2022

Two studies in resource-efficient inference: structural testing of networks, and selective classification

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

TWO STUDIES IN RESOURCE-EFFICIENT INFERENCE: STRUCTURAL TESTING OF NETWORKS, AND SELECTIVE CLASSIFICATION

by

ADITYA GANGRADE

B.Tech., Indian Institute of Technology Bombay, 2014

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2022

Approved by

First Reader

Bobak Nazer, PhD
Associate Professor of Electrical and Computer Engineering

Second Reader

Venkatesh Saligrama, PhD
Professor of Electrical and Computer Engineering
Professor of Systems Engineering
Professor of Computer Science

Third Reader

Daniel Sussman, PhD
Assistant Professor of Mathematics and Statistics

Fourth Reader

Ery Arias-Castro, PhD
Professor of Mathematics
Professor of Data Science
University of California, San Diego

To the memory of my grandmother.

Acknowledgments

It is meant to be a lucky thing to find a doctoral adviser that you get along with intellectually and personally. I am then at least quadratically as lucky to have found two such people in Bobak and Venkatesh, both of whom have contributed immensely, not only to my understanding of research, but also to my understanding of life. I want to thank them both for being so generous with their time and ideas, for always being encouraging, and for exposing me to lines of thought that I never would have gotten to on my own. I especially want to thank Bobak for always lending an ear, and for guiding with a light touch; and Venkatesh for sharing his wisdom, and for being an unending font of interesting questions.

I am fortunate to have had Ery Arias-Castro and Daniel Sussman on my thesis committee, and am thankful to them for their free sharing of ideas, and for the vast amounts of encouragement they gave me.

A number of the faculty in the broad IDS group at BU have enriched my experience here over the years, through their contributions to various reading groups, through conversations, and through advice. I want to especially thank David Castañón, Ashok Cutkosky, Prakash Ishwar, Brian Kulis, Alex Olshevsky, and Yannis Paschalidis. I would also like to thank the administrators of the Systems Engineering department for the leniency shown to me over the years, and especially Elizabeth Flagg for her infinite patience, and for what I like to call 'the great auditing of 2019.'

Since this dissertation represents a culmination of my academic study, I want to thank my teachers over the years that have been crucial to me getting here. Special thanks are due to Ankur Kulkarni at IITB, Vinod Prabhakaran at TIFR, and Rajesh Sundaresan at IISc, who very kindly hosted me in 2014 and 2015, and were my first introduction to concrete research; to B.G. Fernandes, H. Narayanan, Punit Parmananda, and Sibi Raj Pillai at IITB, without whose encouragement and advice

I never would have set down this path; and to Vashali Basu and Vipul Mehra at Cathedral, and Deepak Mala Sawhney at DPS, who made learning so enjoyable.

My time at BU has been greatly enriched by the many talented and generous students I have met here (and beyond), and the myriad conversations, academic and otherwise, that I have had with them. I want to especially thank the people I have collaborated closely with over the years, namely Alp Acar, Tianrui Chen, Anil Kag, Feng Nan, Ali Siahkamari, and Praveen Venkatesh. It has been a great privilege to learn from such a varied and brilliant group of people.

Stepping outside work, I would like to thank my mates, and especially the Windsor street crew of Shibani, Rushina, and Ishita, and Bo and PPP, who have contributed more to my general sanity than I would ever have the grace to properly let on or appreciate them for.

My parents, Smita and Vivek Gangrade, are ultimately responsible for this work by engendering an interest in science so long ago, and giving me the room to explore it in my way. I am very grateful, for this, and for so much more. I further want to thank the extended clan, and to remember my grandmother, Usha Gangrade, who I miss dearly, and to whom this work is dedicated.

TWO STUDIES IN RESOURCE-EFFICIENT INFERENCE: STRUCTURAL TESTING OF NETWORKS, AND SELECTIVE CLASSIFICATION

ADITYA GANGRADE

Boston University, College of Engineering, 2022

Major Professors: Bobak Nazer, PhD
Associate Professor of Electrical and Computer Engineering

Venkatesh Saligrama, PhD
Professor of Electrical and Computer Engineering
Professor of Systems Engineering
Professor of Computer Science

ABSTRACT

Inference systems suffer costs arising from information acquisition, and from communication and computational costs of executing complex models. This dissertation proposes, in two distinct themes, systems-level methods to reduce these costs without affecting the accuracy of inference by using ancillary low-cost methods to cheaply address most queries, while only using resource-heavy methods on 'difficult' instances.

The first theme concerns testing methods in structural inference of networks and graphical models, the proposal being that one first cheaply tests whether the structure underlying a dataset differs from a reference structure, and only estimates the new structure if this difference is large. This study focuses on theoretically establishing separations between the costs of testing and learning to determine when a strategy such as the above has benefits. For two canonical models—the Ising model, and the stochastic block model—fundamental limits are derived on the costs of one- and two-sample goodness-of-fit tests by determining information-theoretic lower bounds, and developing matching tests. A biphasic behaviour in the costs of testing is demonstrated: there is a critical size scale such that detection of differences smaller than this size is nearly as expensive as recovering the structure, while detection of larger differences has vanishing costs relative to recovery.

The second theme concerns using Selective classification (SC), or classification with an option to abstain, to control inference-time costs in the machine learning framework. The proposal is to learn a low-complexity selective classifier that only abstains on hard instances, and to execute more expensive methods upon abstention. Herein, a novel SC formulation with a focus on high-accuracy is developed, and used to obtain both theoretical characterisations, and a scheme for learning selective classifiers based on optimising a collection of class-wise decoupled one-sided risks. This scheme attains strong empirical performance, and admits efficient implementation, leading to an effective SC methodology. Finally, SC is studied in the online learning setting with feedback only provided upon abstention, modelling the practical lack of reliable labels without expensive feature collection, and a Pareto-optimal low-error scheme is described.

Contents

List of Tables

List of Figures

List of Abbreviations

BL	....	Budget Learning
DNN	....	Deep Neural Network
EoF	....	Error-of-Fit
ERM	....	Empirical Risk Minimisation
GoF	....	Goodness-of-Fit
i.i.d.	....	Independent and identically distributed
IoT	....	Internet of Things
ML	....	Machine Learning
SBM	....	Stochastic Block Model
SC	....	Selective Classification
SL	....	Structure Learning
SNR	....	Signal-to-Noise Ratio
TST	....	Two-Sample Testing

Chapter 1

Systems-Level Approaches to Resource-Efficient Inference

The design of practical inference systems must account for a variety of costs associated with making accurate inferences. Such costs arise from a two main sources.

First are the statistical costs associated with the quality and amount of data required for reliable inference to be possible. The nature of this expense can differ depending on the setting - for instance, accurately fitting a parametric model to a source of data usually requires a certain minimum number of samples from the same, or to accurately classify an object might require us to collect enough features about it. These requirements are fundamental, and are often referred to as Information-theoretic requirements in the computer science literature.

The second major set of costs is those associated with running the inference system itself. For instance, a very data efficient but computationally inefficient method for fitting parameters is completely infeasible. Less drastically, but no less importantly, large models such as deep neural networks used in modern classification systems are accurate, but suffer from huge practical computational costs in deployment to the extent that they cannot be implemented on typical ‘edge’ devices such as mobile phones or embedded Internet of Things (IoT) devices. Of course, these costs need not be purely computational - for instance, one common strategy for handling queries at edge devices is to communicate these to a ‘cloud’ server that implements a complex model. The costs associated with such a deployment include the energy costs and latency of communicating with the server, as well as the (amortised over queries) monetary cost of implementing the infrastructure associated with the server.

Of course, these two types of costs are not present in isolation, but interact to inform the design of inference systems. Indeed, one might choose to implement a data-inefficient method simply because it is computationally feasible, while known data-efficient methods are not. Conversely, despite the high costs of communication, the cloud-edge design described previously *is* implemented, since this is usually required to deliver accurate results. Balancing these costs while delivering accurate inference has been a rich vein of research over the past century, and strongly informs the study of methods for inference.[1]

This dissertation is focused on *systems-level* approaches to controlling the costs incurred by a system at the time of inference. The basic model adopted is that an accurate but expensive core method is available as a black-box, which the system can call upon at will. Given this, we will design ancillary methods that quickly process either cheaply available or small amounts of data to decide whether the expensive method needs to be executed in order to get an accurate answer. If the core method is not executed, these ancillary methods must produce an answer of their own, and avoid the expense of running this core method. Otherwise, the core method *is* executed, and the answer is assured to be accurate. The basic goal of these ancillary systems thus is to minimise the usage of the expensive core method, but without any significant loss of accuracy in the process. This directly ameliorates the average cost of inference in (common) settings where most queried instances can be handled easily, but a minority requires the increased precision of the expensive methods. The overall system can be represented as the block diagram in Figure 1·1.

[1] Of course, along with all of these inference-time costs, the statistical and computational costs of finding good methods and models for data are also important - for instance, a machine learner needs to collect data and fit a model, which incurs both statistical and computational costs. This dissertation does not focus on these.

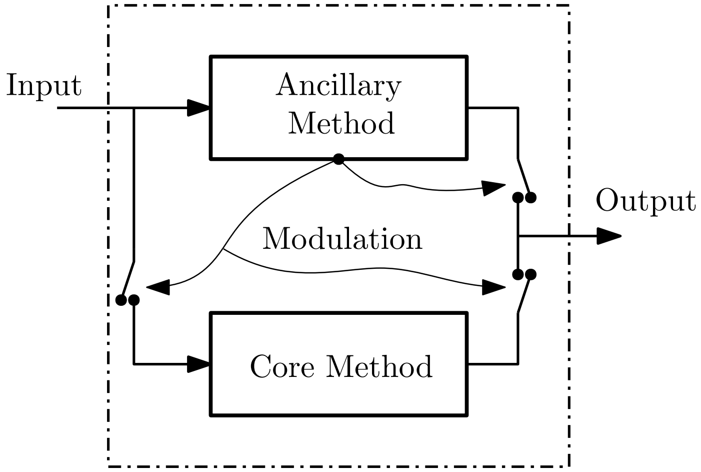


Figure 1·1: A representation of the basic structure of the methods - the ancillary method cheaply processes the input, and decides whether the expensive core method should be executed, and modulates the corresponding access to the input. It should be noted that the core method may draw more or richer input data, or employ greater amounts of resources such as computation or energy, which is not represented in the figure.

The following studies this broad approach in two quite distinct themes, both in setting and the nature of the primitives explored. The first of these is the use of testing methods to modulate structure recovery methods in networks and graphical models, and the second is the use of low-cost selective classifiers to modulate the use of complex methods in the machine learning setup. In my view the primary difference between these settings is that the former studies concrete parametric distributions, while the latter domain is primarily discriminative and relatively unstructured. This distinction strongly colours the type of study in either domain - in the first structured setting, the bulk of the study regards theoretically establishing for what scenarios these testing methods have a significant advantage over always running the expensive methods; in contrast, in the second unstructured setting, we describe novel methodologies along with theoretical results of the effectiveness of such methods, and further characterise the advantage of these methods in terms of approximation theoretic properties.

It should be noted that while each of these primitives have strong relevance to resource-efficient inference, they have significance beyond these considerations as well. Subsequent chapters will embrace this breadth in their contextualisation, but nevertheless the study remains informed by the needs of resource-efficient inference, and this determines the types of questions we ask, and the solution concepts we adopt. The remainder of this chapter gives a more detailed overview of these investigations, along with a short survey of the results obtained, and highlights the relevance to resource-efficient inference.

1.1 Structural Testing of Networks and Graphical Models

Network-structured data is ubiquitous in scientific domains, whether 'hard' or social, where the large scale features of a system derive from the interaction of smaller objects or agents within it. Common examples include the relationships between people in a sociologial context, which naturally form networks that inform the behaviour of individuals [OR02]; protein-protein or gene interactions in molecular biology, whose in situ dynamics inform the overall behaviour, and are typically represented as a network [Cos+10; PF95]; and in neuroscience, where the dynamics of neuran populations is driven by the 'connectome' - the network of connections between individual neurons [Orl+15]. Recent advances in each of these areas has lead to the proliferation of large scale data collection - the rise of social media in sociology, or high throughput gene and protein interaction techniques in molecular biology [Cos+10; PF95], and of calcium fluorescence imaging in neuroscience [YY17] each allows observation of thousands of agents at the same time, with the goal of identifying both fine and crude underlying structures in these domains. For instance, geneticists attempt to reason about groups of genes that together encode some gross behaviour[Cli+07], and neuroscientists try to establish the broad structures of connectomes in various scenarios [Orl+15].

This motivates the study of network structured or network driven probabilistic models, and that of basic inference tasks on the same. Indeed, there has been a flurry of activity in the past decade on the mathematical aspects of the recovery of these underlying structures, with new algorithms, and information theoretic characterisation of the sample costs in various settings - for example [DM17; SW12; Abb18; ABH16; MNS15; CRV15; Bre15; VMLC16; KM17] and many others. However, due to the intrinsic variability of the underlying data, such methods are typically expensive in the sense that a lot of high-quality data is required to reliable recover these latent structures.

This dissertation explores the use of hypothesis testing as a precursor to such estimation methods in order to mollify the high cost of recovery. The following illustrates the premise - suppose a biologist is interested in determining how the interactions between proteins in certain cell systems differ in diseased populations as opposed to healthy ones [IK12]. Typical formulations identify plausible sub-systems that might exhibit a difference in structure, but many of these in fact would not be significantly different - indeed, this constitutes one of the major challenges of discovering effects in such domains. Further, for most of these sub-systems, well characterised information about the baseline stuctures in the commonplace healthy populations is available, simply because this constitutes the bulk of data. With this background, the idea of testing is natural - we can first try to determine if the latent structure of a sub-system is significantly different from the baseline structure or not - if not, then this baseline structure itself can be taken as the estimate, and we need only undergo the intensive data collection process for recovering the structure if it is indeed significantly different. This problem corresponds to 'one-sample' Goodness-of-fit (GoF) testing. This type of problem is commonplace, and applies also to marketers trying to determine if the state of a social network has changed compared to historical data, or to neuroscientists studying how if learning a task manifests as changes in the connectome [Moh+16]. Of course, such tests are also more broadly relevant, in particular to validating recovered models, and to detecting the presence of structural effects in response to changes in external conditions (such as changes in protein-interaction networks in the presence of a drug [IK12]).

The potential for benefit of the above testing-based strategy lies mainly in a potential reduction in statistical costs, in that since testing is trying to recover much simpler information than the structure itself, it should be much cheaper than structure learning. In fact, this is pretty much a requirement for the strategy to be meaningful - if the costs of testing are instead comparable to those of structure recovery, then there is no benefit, and in fact an extra cost to testing. An additional desideratum is that the tests themselves should be computationally simple, which is a second benefit when the procedure exits at the testing phase.

Concretely, we study two families of models, that represent different forms of observations that are encountered in such network settings.

- The *Stochastic Block Model* (SBM), which is a random graph with a latent clustering structure, captures settings where the underlying network of interactions is directly observable, and the latent structure represents groups of agents such that the level of interactions between agents depends on the groups they belong to. Such scenarios are relevant to, e.g. , gene interaction networks, where the goal is typically to reason about functional groups of genes that tend to interact more with than outside the group [Abb18].

- The *Ising Model*, which is a pairwise binary graphical model, captures settings where behaviour of the individual agents can be directly observed, but the links between them cannot. The latent structure here is the underlying Markov network (see Ch.3) that governs the interactions of these agents. This is relevant to, e.g. , neuron studies via calcium fluorescence, where the activity of individual neurons can be observed, but the connections between neurons cannot, but the overall behaviour is driven by these connections.

The following describes the above problem more concretely, and gives a brief survey of the results.

Generic Problem Description We will work with a parametric family of $\mathcal{X}$-valued distributions $\{P_\theta^\lambda\}$, where $\theta \in \Theta$ denotes a set of parameters, and $\lambda \in \mathbb{R}$ denotes a notion of statistical quality - for our purposes, this can be a signal-to-noise ratio (SNR) in the SBM, or the number of independent and identically distributed (i.i.d.) samples we can draw from P_θ in the Ising model. In addition, we will assume a given pseudo-metric ρ on Θ, which encodes the structure we are interested in reasoning about[2]. Finally, we will assume a given distortion parameter s, which controls how precisely we want to recover the underlying parameters.

With this preamble, we can define the basic inference problems of interest as follows

Structure learning (SL; also known as estimation or recovery) over $\{P_\theta^\lambda\}_{\theta\in\Theta}$ with distortion s:

> *Design a procedure* $\hat\theta : \mathcal{X} \to \Theta$ *such that given* $X \sim P_\theta^\lambda$, *for an unknown* θ, *such that* $\rho(\theta, \hat\theta(X)) < s$ *with high probability.*

Goodness-of-fit testing (GoF; also known as identity testing) over $\{P_\theta^\lambda\}_{\theta\in\Theta}$ with distortion s:

[2] A pseudo-metric is just a metric but relaxed so that the ‘distance’ between two distinct points can also be zero. This is important in capturing relevant parts of the latent structure - for instance, if we are working with weighted graphs but are interested in the corresponding unweighted structure, then d can equate graphs with the same edges even if they have distinct weights

Design a procedure $\varphi : \mathcal{X} \times \Theta \to \{0,1\}$ such that given $X \sim P_\theta^\lambda$, for an unknown θ, and a baseline parameter θ_0, φ has the following behaviour with high probability

- *If $\theta = \theta_0$, then $\varphi(X, \theta_0) = 0$.*
- *If $\rho(\theta, \theta_0) \geq s$, then $\varphi(X, \theta_0) = 1$.*

Note that depending on how good the data is (i.e., how large λ is), the above problems may not be solvable - for instance it is impossible to recover anything if we observe no data. The principal objects of study are the statistical complexity of the tasks, which capture the quality needed for the above tasks to be reliably performed. We will symbolically denote these as follows - these definitions are informal, and details are left to the specific chapters.

$$\Lambda_{\mathrm{SL}}(s) := \inf\{\lambda : \text{SL over } \{P_\theta^\lambda\}_{\theta \in \Theta} \text{ is solvable}\},$$
$$\Lambda_{\mathrm{GoF}}(s) := \inf\{\lambda : \text{GoF over } \{P_\theta^\lambda\}_{\theta \in \Theta} \text{ is solvable}\}.$$

It should be noted that one can always solve GoF via SL with a small cost in distortion - indeed, if $\hat{\theta}$ solves SL with distortion ${}^{s}/_{2}$, then we can produce an estimate $\hat{\theta}$ such that if $\theta \neq \theta_0$, then $\rho(\hat{\theta}, \theta_0) \geq {}^{s}/_{2}$, while if $\theta = \theta_0$, then $\rho(\hat{\theta}, \theta_0) < {}^{s}/_{2}$. Importantly, the costs Λ are not very sensitive to constant factors in the situtaion we study. Thus, we can bound

$$\Lambda_{\mathrm{GoF}}(s) \leq \Lambda_{\mathrm{SL}}({}^{s}/_{2}) \approx \Lambda_{\mathrm{SL}}(s).$$

Now, for the test-then-recover strategy of the previous section to be effective, we require that $\Lambda_{\mathrm{GoF}}(s) \ll \Lambda_{\mathrm{SL}}(s)$ - indeed, otherwise we do not gain anything when we avoid executing the expensive recovery method. As it turns out, the question of whether this indeed holds depends both on the family $\mathcal{P}$, and on s. Characterising this dependence, as encapsulated below, forms the main subject of investigation in Chapters 2 and 3.

> *For the SBM and Ising Models, characterise $\Lambda_{\mathrm{GoF}}(s)$, and in particular determine for what values of s this is much smaller than $\Lambda_{\mathrm{SL}}(s)$.*

A Brief Summary of Results The main result of our study is that the statistical complexity of testing exhibits a biphasic behaviour as the distortion s varies. More concretely, we observe that there are critical size scales[3] $\mathfrak{s}(\mathcal{P})$ that vary mainly with the dimensionality of the parameter space such that

- If $s \ll \mathfrak{s}$, then $\Lambda_{\mathrm{GoF}}(s)$ is comparable to $\Lambda_{\mathrm{SL}}(s)$.
- If $s \gg \mathfrak{s}$, then $\Lambda_{\mathrm{GoF}}(s) \ll \Lambda_{\mathrm{SL}}(s)$.

More concretely, for SBMs on p nodes, this $\mathfrak{s}$ is $\sqrt{p}$. For Ising models, the story is more complex (see Ch. 3), but for the subclass of Ising models on p nodes with *tree-structured* networks, again $\mathfrak{s} = \sqrt{p}$. These results sharply characterise when the test-then-learn approach has a strong advantage - namely if the distortion to which one wishes to learn the structures is exceeds $\mathfrak{s}$.

It should be noted that biphasic behaviour in Λ_{GoF} is relatively common in structured settings - for instance, this occurs in the testing of linear models with a fixed design.[4] However, typically in such simple settings, the costs of GoF are uniformly much smaller than those of recovery, which is in sharp contrast with the above observations, where too small a tolerance s makes testing as hard as recovery. A further point of comparison is the testing of completely unstructured distributions (such as arbitrary law on a finite alphabet), which do not suffer phase transitions[5],

[3] To be absolutely clear - we do not show a generic result demonstrating such a phase transition. Instead, we observe similar effects by explicitly controlling Λ_{GoF} and Λ_{SL} for the models we study.

[4] This is mainly folklore. Roughly speaking, under a known Gaussian design G, testing problems can be reduced to the distinguishability problem of differentiating data $Y = Gx + W$ for a s-sparse p-dimensional x and standard noise w from the pure noise model $Y = W$. By simply running sparse recovery, or by considering the energy of the observed signal, it is possible to detect these differences with $\widetilde{\Theta}(\min(s, p/s))$ samples, again encountering a phase transition at $s = \sqrt{p}$. Importantly, this is always smaller than the $\Theta(p)$ samples needed for recovery if $s \ll p$. In fact, very small changes are also easy to detect here! Please write to me if you'd like a concrete proof.

[5] at least for the simple null as above

and testing costs are always much lower than those of recovery [DKW18; Gol17].

The approach taken to showing the above results is two-fold. Firstly, we use information theoretic methods to establish minimax lower bounds and thus derive impossibility results. These all utilise the approach of designing appropriate alternates and analysing the χ^2-divergence between a mixture over these and a generic null, as developed, e.g., by Ingster [IS12].

We complement the above lower bounds by designing simple test statistics that match these bounds for large changes, thus yielding upper bounds on the complexity. Concretely, we show sharp bounds for the SBM and tree-structured Ising models, as well as sharp bounds for the restricted problem of testing for deletions in ferromagnetic Ising models. These test statistics are efficient to compute, in a practical sense, and further are 'global' in the sense that a single statistic that accounts for all possible alternatives is developed, rather than more local scan-based methods like generalised likelihood ratios that involve an M-estimation to find the 'closest' alternate. This non-local property is crucial - the scan based methods instead suffer a huge blow-up in statistical complexity due to the need to ensure a small size of the in the face of combinatorially many close-by alternates that must all be protected against. Indeed, this aspect of needing to test combinatorially many possibilities intrinstically present in these high-dimensional setting is the key statistical challenge in testing these network structures in a data efficient manner, and the global strategies represent an interesting strategy to address this that should be more broadly relevant.

In addition to the above, we also develop lower bounds on approximate recovery and for property testing of Ising model structures, and develop a tight two-sample test for the SBM. We will leave the discussion of these results to the appropriate chapters.

1.2 Selective Classification For Resource-Efficient Inference

Modern machine learning (ML) models are remarkably accurate, and as remarkably complex. For instance, the best models can obtain a staggering 90% accuracy on complex tasks such as ImageNet [Den+09] but even 'efficient' models that achieve this requires deep neural network (DNN) of nearly half a billion parameters [PDXL21]. Out of necessity, then, such models are typically implemented on powerful machines running graphics processing units.

Conversely, the queries to these ML models often arise at so-called 'edge' devices - mobile phones, internet-of-things (IoT) sensor et c. Such devices are severely battery and processing power limited. This imposes severe constraints on the methods implementable in such settings - for instance, the typical CPU-based structure of such devices precludes the use of many convolutional layers in vision tasks due to computational latency [ZWTD19], imposing architectural constraints. In particular, modern high accuracy methods like deep neural networks are seldom implementable in these settings. At the same time, edge devices are required to give fast and accurate decisions. Enabling such mechanisms is an important technical challenge.

There are two main strategies for this - pure edge solutions, and cloud based solutions. The former consists of learning weak models using methods like resouce aware training or distillation that can be implemented on the edge directly (e.g. [Wu+19; KGV17; HVD15]). However, such approaches inevitably results in a large drop in accuracy - for instance, edge-implementable models like MCUNETs only have accuracy of about 50% [Lin+20]. The cloud approach instead provides the edge devices with a communication link to a server that implements one of the aforementioned complex models, thus gaining accuracy, but at the cost of increased latency and energy consumption due to communication. These costs are very significant - for instance, the battery drain of typical edge devices is dominated by the energy cost of communication

[Zhu+19; Hol17; Nor19].

The above sets the stage for trying to efficiently utilise the expensive-but-accurate cloud models. The principal observation underlying this is that many queries to such a system are relatively simple to process. The ideal of operation of such a system should handle such ‘easy’ queries at the edge, and only use the cloud server for more subtle cases that require its extra complexity. Such an operation would retain the overall high accuracy of a solution that always uses the cloud model, but with a reduced resource cost due to all the instances processed at the edge.

Situations like the above are common in inference systems. For example, when deciding if a mammary mass is benign or malignant, a physician may predict based on ultrasound imaging tests, and, in more subtle cases, abstain and refer the patient to a specialist, or recommend specialised imaging such as CT scans; or an automated content moderating system can either verify whether a post satisfies guidelines, or fall back to an expensive but accurate human moderator. There are two main sources of the extra resource costs of the expensive option, as discussed previously - it could be the case that the expensive method is much more complex to execute, as above, or it could be the case that the expensive method uses extra information, such as the features collected by a CT scan in the medical example. Figure 1·2 illustrates these models. The noisy left image describes a situation in which the available features are not enough to separate the two classes, and so a central region where one should abstain emerges. In contrast, the features are enough to separate the data in the right image, but only in a complex way. When the decision boundaries are required to be simple (for instance, straight lines), this boundary cannot be represented, again leading to a central region where one should abstain. Of course, these aspects can arise together.

Selective classification (SC) [Cho57; Cho70] is a classical paradigm of relevance

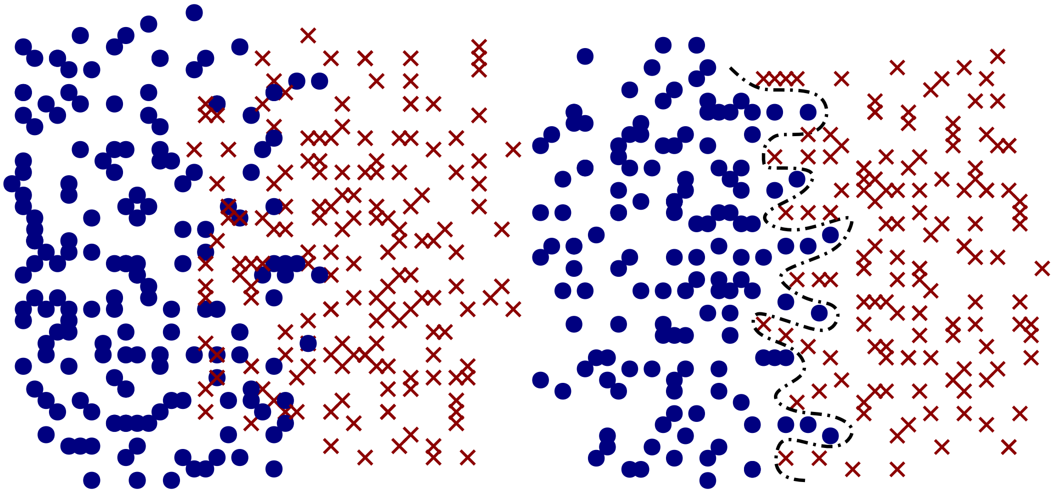


Figure 1·2: Two modes by which the need for SC may emerge. Left represents situations where cheap features are unable to separate classes. Right represents situations where the features are enough, but the decision boundary is complex, and thus cannot be cheaply computed.

to such settings. The setup allows a predictor to abstain (also called reject) from classifying some instances (without incurring a mistake). This abstention models adaptive decisions to invoke more resource-intensive methods on subtle cases, like in the above examples. Our concrete proposal is to learn a low-complexity selective classifier to serve as the 'ancillary method' in the situation of Figure 1·1. If this classifier abstains on an instance, the expensive core method can be executed, and otherwise the selective classifier produces an appropriate answer.

The primary desiderata for such a selective classifier is that *accuracy* is high, while *coverage*—the fraction of points that the selective classification does not abstain upon—is as high as possible. The former means that whenever the classifier does answer, it is very accurate, so that the overall error rate of the system can be controlled, while the coverage is a direct measure of efficiency of the system. This is a subtle problem, mainly because there is no direct supervision for whether a point is easy or not, and this must be inferred with standard class labels. Further, implementing such a solution requires a reasonable methodology to practically train such classifiers.

The most common prior SC formulation, and the resulting methods, use a 'gating-structure,' wherein abstention is explicitly modelled by a binary-valued function γ, and classification is handled by a function π. An instance, x, is predicted as $\pi(x)$ if $\gamma(x) = 1$, and otherwise rejected. Within this formulation, recent work has proposed a number of methods, ranging from alternating minimisation based joint training, to the design of new surrogate losses, and of new model classes to accommodate rejection. Despite this increased complexity, these methods lack power, as shown by the fact that their performance is essentially matched by naïve schemes that rely on abstaining on the basis of post-hoc uncertainty estimates for a trained standard classifier. This represents a significant gap in the practical effectiveness of selective classification.

A Brief Summary of Results **Selective Classification** In Chapter 4, we describe a novel formulation for the SC problem, that comprises of directly learning *disjoint* classification regions $\{\mathcal{S}_k\}_{k\in\mathcal{Y}}$, each of which corresponds to labelling the instance as k respectively. Rejection is *implicitly defined* as the gap, i.e., the set $\mathcal{R} = \mathcal{X} \setminus \bigcup \mathcal{S}_k$. We show that this formulation is equivalent to earlier approaches, thus retaining expressivity.

The principal benefit of our formulation is that it admits a natural relaxation, via dropping the disjointness constraints, into *decoupled* 'one-sided prediction' (OSP) problems. We show that at design error ε, this relaxation has the coverage optimality gap bounded by ε itself, and so the relaxation is statistically efficient in the practically relevant high target accuracy regime. In addition, we provide parametric learnability results for both the OSP and the SC problems.

We pose OSP as a standard constrained learning problem, and due to the decoupling property, they can be approached by standard techniques. We design a method that efficiently adjusts to inter-class heterogeneity by solving a minimax program, controlled by one parameter that limits overall error rates. This yields a powerful SC training method that does not require designing of special losses or model classes, instead allowing use of standard discriminative tools.

To validate these claims, we implement the resulting SC methods on benchmark vision datasets. We empirically find that the OSP-based scheme has a consistent advantage over SOTA methods in the regime of low target error. In particular, we show a clear advantage over the naïve scheme described above, which in our opinion is a significant first milestone in the practice of selective classification.

Budget Learning Next, in Chapter 5 we explicitly study the situation of the first example above, wherein black-box access to a very accurate but complex classifier is available, and we wish to learn *simple* selective classifiers, which we term budget learning. By reinterpreting the the prior SC formulation, we draw a connection between budget learning and the approximation theoretic notion of brackets - in the regime where a simple selective classifier attempts to match the performance of a given high complexity model, the problem is equivalent to learning a good bracketing of this high complexity model.

This interpretation also identifies the key budget learning problem as an approximation theoretic question - *which complex classes have 'good' bracketings by simple classes?* We characterise this for a binary version of Hölder smooth classes, and also provide partial results for generic classes with bounded VC dimension.

Finally, we describe empirical results that concretely construct low-complexity selective classifiers to match high complexity model on benchmark vision tasks. Even with a strong disparity in the cloud and edge models (§5.5), we obtain usages of $20 - 40\%$ at accuracies higher than 98% with respect to the cloud. Further, we outperform existing methods in usage by factors of $1.2 - 1.4$ at these high accuracies.

Online Selective Classification Finally, in Chapter 6, we study selective classification in the online learning setting. This setting is significantly different from the above batch setting, mainly since queries arrive in a streaming way, and the model must decide whether each should be abstained upon or not. Notice that in this streaming setting, for the model to get reliable labels, it actually needs to run the expensive method. This sets up the critical difference that in this online setting, we impose the restriction that feedback is provided to the learner only when it abstains.

This sets up the problem of online selective classification with limited feedback, we study in both the harsh adversarial setting, as well as the more benign stochastic setting. Concretely, an adversary sequentially produces contexts and labels (X_t, Y_t), and the learner uses the X_ts to produce a decision $\widehat{Y}_t$ that may either be one of K classes, or an abstention, which we represent as $\perp$. Feedback in the form of Y_t is provided if and only if $\widehat{Y}_t = \perp$, and the learner incurs a mistake if $\widehat{Y}_t$ was non-abstaining and did not equal Y_t.

Throughout, the emphasis is on ensuring very few mistakes, to account for the need for very accurate decisions. With this motivation, we study regrets achievable when compared to the behaviour of the best-in-hindsight error-free selective classifier from a given class - that is, one that makes no mistakes, while abstaining the fewest number of times. Notice that our situation is non-realisable, and therefore this competitor may abstain in the long-run. The two metrics of importance here are the number of mistakes the learner makes, and its excess abstention over this competitor. An effective learner must control both abstention and mistakes, and it is not enough to make one small, e.g. a learner that makes a lot of mistakes but incurs a very negative excess abstention is no good. This *simultaneous* control of two regrets raises particular challenges.

We construct a simple scheme that, when competing against finite classes, simultaneously guarantees $O(T^{\mu})$ mistakes and $O(T^{1-\mu})$ excess abstentions against adaptive adversaries (for any $\mu \in [0, 1]$), and show that these rates are Pareto-tight [OR94]. We further show that against stochastic adversaries, the same rates can be attained with improved dependence of the regret bounds on the size of the class, and we also describe schemes that enjoy similar improvements against adaptive adversaries, but at the cost of the T-dependence of the regret bounds. The main schemes randomly abstain at a given rate in order to gain information, and otherwise play $\widehat{Y}_t$ consistent with the 'version space' of classifiers that have not been observed to make mistakes. For the adversarial case, the analysis of the scheme relies on a new 'adversarial uniform law of large numbers'(ALLN) to argue that such methods cannot incur too many mistakes. This ALLN uses a self-normalised martingale concentration bound, and further yields an adaptive continuous approximation guarantee for the Bernoulli-sampling sketch in the sense of Ben-Eliezer & Yogev [BY20; Alo+21]. The theoretical exploration is complemented by illustrative experiments that implement our scheme on two benchmark datasets.

1.3 Format

We will separately discuss the two themes in two separate 'parts', which can be read independently, with separate chapters devoted to each of the directions of exploration we described above. These chapters are essentially lightly edited reproductions of the papers wherein this research was originally represented. Nearly all of these publications are in machine learning conferences wherein, due to length constraints, the usual style is to present the main results and intuitions in the main text, and relegate technical details and proof to an appendix. I have come to appreciate this style (after non-trivial teething pains), and accordingly, we shall commit to the same here - so, the chapter provide exposition, and describe and interpret the main results, while the concrete details of proofs are left to the (copious) appendices of this dissertation. Each chapter is roughly self-contained (perhaps with the exception of Chapter 5, which frequently refers to Chapter 4), and may be read in any order desired, although I prefer the order given within the parts. Each chapter also discusses some open problems at the end.

Part I

Structural Testing of Networks and Graphical Models

Chapter 2

Testing Community Structures of Stochastic Block Models

While community detection and recovery for the stochastic block model (SBM) [Abb18] and, more generally, inference of community structures underlying large-scale network data [GN02; New06; For10] has received significant interest across the machine learning, statistics and information theory literature, there has been limited work on the important problem of testing changes in community structures. The general problem of testing changes in networks naturally arises in a number of applications such as discovering statistically significant topological changes in gene regulatory networks [Zha+08] or differences in brain networks between healthy and diseased individuals [Bas+08]. Building upon this perspective, we propose testing of differences in the underlying community structure of a network, which can encompass scenarios such as detecting structural changes over time in social networks [AG05; For10], determining whether a set of genes belong to different communities in disease and normal states [JTZ04], and deciding whether there are changes in functional modules, which represent communities, in protein-protein networks [CY06].

Testing structural changes in networks is statistically challenging due to the fact that we may have relatively few independent samples to evaluate combinatorially-many

This chapter is a lightly edited reproduction of [GVNS19], which was written in collaboration with Bobak Nazer, Praveen Venkatesh, and Venkatesh Saligrama.

potential changes. In this chapter, we propose methods for goodness-of-fit (GoF) testing and two-sample testing (TST) for detecting changes in community memberships under the SBM. The SBM naturally captures the community structures commonly observed in large-scale networks, and serves as a baseline model for more complex networks. Specifically, there are n nodes partitioned into two equal-sized communities, and the network is observed as a random $n \times n$ adjacency matrix, representing the instantaneous pairwise interactions among individuals in the population. Both intra- and inter-community interactions are allowed. Members within the same community interact with uniform probability a/n, while members belonging to different communities with a smaller probability b/n. We restrict attention to the commonly-considered and practically-relevant setting of $a/b = \Theta(1)$.

For our testing problems, we assume that the network samples are aligned on $n \gg 1$ vertices, and that the latent communities are either the same, or they differ in at least some $s \ll n$ nodes. We pose the GoF problem as: *Decide whether or not the observed random graph is an instantiation of a given community structure.* For the TST problem, we ask: *Given two random graphs, decide whether or not their latent community structure is identical.*

Sparse vs. Dense Graphs. We focus on scenarios where the observed random incidence matrices are sparse with average degree-per-node bounded by a constant independent of the network size. Ours is the first work to develop minimax optimal methods for GoF and TST in this context. We are motivated by both practical and theoretical concerns.

Practically, as observed in [Chu10], realistic graphs such as social networks are sparse (friendships do not grow with network size); in temporal settings, at any given time, only a small subset of interactions are observed; and in other cases ascertaining the presence or absence of each edge in the network being observed is an expensive process, and it makes sense to understand the fundamental limits for when testing is even possible.

From a theoretical standpoint, the sparse setting is challenging due to signal-to-noise ratio (SNR) constraints that do not arise in the dense case. Recovery of the latent community with up to s errors is possible iff $\Lambda \gtrsim \log(n/s)$ [CRV15; ZZ16; FC19], where Λ is a SNR parameter that, in the setting $a/b = \Theta(1)$, scales linearly with the mean degree. In particular, for Λ of constant order, recovery with sublinear distortion fails. The question of *whether testing is possible when recovery fails* is mathematically intriguing. Further, this is the *only* theoretically interesting setting. Indeed, if testing for s changes requires a graph dense enough to allow recovery with $\sim s$ errors, then one might as well recover these communities and compare them.

Contributions. We show that optimal tests exhibit a biphasic behaviour:

1. For $s \gg \sqrt{n}$, or 'large changes,' we propose computationally-efficient schema for GoF and TST that succeed with $\Lambda = O(1)$ - far below the SNR threshold for recovery. For GoF, this requirement is even weaker - we only need $\Lambda \gtrsim n/s^2$, which vanishes with n since $s \gg \sqrt{n}$. Further, we match these bounds up to constants with information-theoretic lower bounds.

2. In contrast, we show via an information-theoretic lower bound that for $s \ll \sqrt{n}$, or 'small changes,' both testing problems require $\Lambda = \Omega(\log(n))$ for reliable testing. This means that the naïve strategy of recovering communities and comparing them is tight up to constants in this regime.

We complement the above theoretical study by three experiments: the first implements the above tests on synthetic SBMs, and the second on the political blogs dataset - a popular real world dataset for community detection [AG05]. Both of these experiments show excellent agreement with the theoretical predictions. The third experiment casts a wider net, and instead studies the related problem of testing the underlying community structure of a Gaussian Markov Random Field that has precision matrix $I + \gamma G$ for G drawn from an SBM. This experiment explores the more realistic setting where instead of receiving a graph, we obtain observations at each node of a hidden graph, and wish to reason about the underlying structure. Remarkably, a simple adaptation of our procedure for SBMs shows excellent performance for this problem. This indicates that our observations are not restricted to raw SBMs, but may signal a more general phenomenon that merits exploration.

Related Work. For work on recovery communities we refer to the excellent survey by Abbe [Abb18]. However, we explicitly point out the papers [CRV15; ZZ16; FC19], which provide various schemes and necessary conditions that show that the partial recovery problem with distortion s can be solved with vanishing error probability if and only if $\Lambda \gtrsim \log(n/s)$. We further point out the lower bound of [DAM17], which assert that if $\Lambda < 2$, then asymptotically, the best possible distortion for partial recovery (or weak recovery, as it is referred to in this constant SNR regime) is $n/2 - o(n)$. Note that reporting a uniformly random community achieves distortion of $s = n/2 - O(\sqrt{n})$.

Ours is the first work to study GoF and TST where both hypothesized models are SBMs. Nevertheless, both GoF and TST in the context of network data as well as SBMs have been studied. Below we highlight the key differences in modeling assumptions and the ensuing technical implications, which renders much of the prior work inapplicable to our setting.

With regards to GoF, [AV14; VA15] study the problem of detecting if a graph is an unstructured Erdős-Rényi (ER) graph, or if it has a planted dense subgraph, providing detailed characterizations of the feasiblity regions and statistical phase transitions in this setting. While this work is aligned with ours in the techniques used, the modeled setting and problem there are different (ER vs. planted dense subgraph), and TST is not explored. Particularly, the dense subgraph model and the SBM are qualitatively different, and conclusions from one cannot be transferred to the other directly.

A number of papers, including [Lei16; BS16; BMNN16; GL17] study various techniques and regimes of determining if a graph is a SBM or an unstructured ER graph, and if the former, the number of communities in the model. Of these, [GL17] approach the problem by counting small motifs in the graphs, [BMNN16] propose a simple scan and [Lei16; BS16] propose testing of the number of communities on the basis of the top singular values of the graph.

[Tan+17] study TST of the model parameters in random dot product graphs, and propose the distance between aligned spectral embeddings of the two graphs as a statistic to do so. They use this to test equality against various transformations of the underlying models, and in particular for SBMs, test if the connectivity probabilities $(a/n, b/n)$ are identical or not for two graphs with latent communities that are randomly drawn. [LL18] adapt these tests by considering the same distance, but weighted by the corresponding singular values of one of the graphs, and use this to study two-sample testing of equality of the latent communities in the graphs - as in this chapter.

In contrast to the low-rank structure assumptions in the above work, Ghoshdastidar, von Luxburg, and collaborators study two-sample testing of inhomogeneous ER graphs (i.e., ER graphs where each edge may have a distinct probability of existing) [GGCV20; GGCvL17; GvL18]. Within this setting, they provide a number of statistics based both on estimates of the Frobenius and operator norms of the differences of the expected graph adjacency matrices, as well as those based on motifs such as triangles, and explore the limits of these tests.

A fundamental drawback of these approaches, in our context, is their reliance on singular values, spectral norms and Frobenius norms. Singular embeddings are particularly sensitive to noise, and stable embeddings require significant edge density (particularly when a sublinear number of alterations to the communities are to be tested). Indeed, in this context, we note that, in contrast to our low SNR, sparse setting, [LL18] require both a degree of $n^{1/2-\varepsilon}$ and an SNR of $\log(n)$ corresponding to a high SNR, high edge-density regime, where full community recovery is possible.

Similarly, Frobenius and Spectral norms based tests of [GvL18; GGCV20] are not stable enough to test a sublinear number of changes in a low SNR regime. Functionally, this can be seen by the fact that the square-Frobenius norm of the difference of two graphs is equal to the number of edges that appear in one graph but not the other, and for sparse graphs, *most* edges appear in only one of the two graphs. Similarly, arguments about spectral norms rely on concentration of the same for ER graphs, but the best known concentration radius [LLV17] is far too large to allow testing of small differences in sparse graphs. Indeed, for any of the statistics of [GvL18] to have power in our setting, the results of the paper require that the expected degree diverges with n, and that $\Lambda \gtrsim n/s$, which is exponentially above the SNR required to recover communities up to distortion $s/2$.

2.1 Definitions

The Stochastic Block Model. A vector $x \in \{\pm 1\}^n$ is said to be a *balanced community vector* (or partition) if $\sum x_i = 0$. The *stochastic block model* is defined as a random, simple, undirected graph G on n nodes such that all edges are drawn mutually independently given x, and

$$P(\{i,j\} \in G|x) = \frac{a+b}{2n} + \frac{a-b}{2n} x_i x_j.$$

Note that we treat x as a deterministic but unknown quantity, and thus, $P(\cdot|x)$ is a slight abuse of notation. The parameters (a,b) may vary with n, and we focus on the setting $a, b = O(\log n)$, and $a/b = \Theta(1)$. For technical convenience, we require that

$a + b < n/4$.

The *signal-to-noise ratio* (SNR) of an SBM is the quantity

$$\Lambda := \frac{(a-b)^2}{a+b},$$

which characterises the recovery problem, as described in earlier discussions.

Note that the partitions x and $-x$ induce the same distribution. Accordingly, the *distortion* between partitions x and y is

$$d(x, y) := \min(d_{\mathrm{H}}(x, y), d_{\mathrm{H}}(x, -y)),$$

where d_{H} is the Hamming distance.

Minimax Testing Problems. We formally define two minimax hypothesis testing problems.

Goodness-of-Fit. We are given a balanced partition x_0 and a parameter s. We receive a graph $G \sim P(G|x)$, where x is an unknown balanced partition that is either exactly equal to x_0 or differs in at least s places. Our goal is to solve the hypothesis test:

$$H_0 : d(x, x_0) = 0 \qquad \text{vs.} \qquad H_1 : d(x, x_0) \geq s.$$

We measure the minimax risk of this problem by

$$R_{\mathrm{GoF}}(n, s, a, b) := \inf_{\varphi} \sup_{x_0} \Big\{ P(\mathrm{FA}) + \sup_{x} P(\mathrm{MD}(x)) \Big\} \tag{2.1}$$

where $\varphi(G)$ outputs either 0 or 1,

$$P(\mathrm{FA}) := P(\varphi(G) = 1 \mid x_0),$$
$$P(\mathrm{MD}(x)) := P(\varphi(G) = 0 \mid x),$$

and the second supremum is over all x such that $d(x, x_0) \geq s$.

Two-Sample Testing. We are given a parameter s and two independent graphs $G \sim P(G|x), H \sim P(H|y)$, where x and y are unknown balanced communities satisfying $d(x, y) \in \{0\} \cup [s : n/2]$. The goal is to solve the following (composite null) testing problem:

$$H_0 : d(x, y) = 0 \qquad \text{vs.} \qquad H_1 : d(x, y) \geq s,$$

with the measure of risk

$$R_{\mathrm{TST}}(n, s, a, b) := \inf_{\varphi} \sup_{x,y} P\big(\varphi(G, H) \neq \mathbf{1}\{x = y\} \,|\, x, y\big), \tag{2.2}$$

where $\varphi(G, H)$ outputs either 0 or 1 and the supremum is over balanced x, y such that $d(x, y) \geq s$.

As we vary n and (s, a, b) with n as some functions (s_n, a_n, b_n), the above define a sequence of hypothesis tests. We say that the GoF problem can be solved *reliably* for such a sequence if $R_{\mathrm{GoF}}(n, s_n, a_n, b_n) \to 0$ as $n \nearrow \infty$, and similarly for TST. Below, we will target $O(1/n)$ bounds. For conciseness, we will suppress the dependence of risks on (n, s, a, b), writing just $R_{\mathrm{GoF}}/R_{\mathrm{TST}}$.

On constants: We use C and c, and their modifications, as unspecified constants that may change from line to line. While these can be explicitly bounded, we do not expect them to be tight.

2.2 Community Goodness-of-Fit

We begin by stating our main results regarding the *community goodness-of-fit problem.*

Theorem 2.2.1. *Community goodness-of-fit testing is possible with risk* $R_{\mathrm{GoF}} \le \delta$ *if* $s\Lambda \ge C\log(2/\delta)$ *and* $\Lambda \ge C\frac{n}{s^2}\log(2/\delta)$ *for some constant* $C > 0$.

Conversely, in order to attain $R_{\mathrm{GoF}} \le \delta \le 0.25$, *we must have that* $s\Lambda \ge C'\log(1/\delta)$ *and* $\Lambda \ge C'\log\left(1+\frac{n}{s^2}\right)$ *for some constant* $C' > 0$.

These bounds reveal the following behavior in terms of large and small changes:

- For large changes ($s \ge n^{1/2+c}$ for some $c > 0$), since $n/s^2 \le 1$ and $\log(1+x) \ge x/2$ for $x \le 1$, the second converse bound behaves as $\Lambda \ge Cn/s^2$, matching the sufficient condition up to a constant.
- For small changes ($s \le n^{1/2-c}$ for some $c > 0$), since $n/s^2 \sim n^{2c}$, the second converse bound instead behaves as $\Lambda \gtrsim \log n$. In this regime, community recovery up to $s/2$ errors requires $\Lambda \ge C\log 2n/s = \tilde{C}\log n$. Thus, estimating x from G and comparing it to x_0 is optimal up to constants.
- The above indicate a phase transition in the GoF testing problem at $\sigma := \log_n(s) = 1/2$. Consider the thermodynamic limit of $n \nearrow \infty$. For $\sigma < 1/2$, the problem is 'hard' in that the SNR Λ is required to diverge to ∞, while for $\sigma > 1/2$, the SNR can tend to zero.

Proof Sketch for the Achievability. Let us begin with an intuitive development of the test. Since we start with a partition x_0 in hand to test, it is natural to look at the edges across and within the cut defined by x_0. We thus define the number of edgess *across* and *within* this cut:

$$\begin{aligned} N_a^{x_0}(G) &:= |\{(i,j) \in G : x_{0,i} \neq x_{0,j}\}| = \frac{1}{4}x_0^\top(D(G) - G)x_0 \\ N_w^{x_0}(G) &:= |\{(i,j) \in G : x_{0,i} = x_{0,j}\}| = \frac{1}{4}x_0^\top(D(G) + G)x_0 \end{aligned} \tag{2.3}$$

where the expressions treat G as an adjacency matrix and $D(G) = \mathrm{diag}(\mathrm{degree}(i))$ collects the degress of each node in a diagnoal matrix. Note that $D(G) - G$ is the

Laplacian of the graph, which commonly features in spectral clustering methods. In the null case, the above statistics are respectively $\mathrm{Bin}(n^2/4, b/n)$ and $\mathrm{Bin}(n^2/4, a/n)$ random variables, while in the alternate case some $s/2 \cdot (n-s)/2$ of each behave like edges of the opposite polarity (i.e. as b/n instead of a/n and vice versa), leading to a excess/deficit of edges of this type. Note that while the 'average signal strength', i.e., the amount by which edges are over- or underrepresented is the same in both cases ($\sim s|a-b|$), the group with the larger null parameter suffers greater fluctuations. Thus, we base our test only on edges of smaller bias. This reduces the SNR by at most a factor of 4.

We now define the test. C_1 below is the constant implicit in Lemma A.1.1 in Appendix A.1.1.

- If $a > b$, we use the test $N_a^{x_0}(G) \overset{H_1}{\underset{H_0}{\gtrless}} \dfrac{bn}{4} + C_1 \max\left(\sqrt{nb\log(1/\delta)}, \log(1/\delta)\right).$
- If $b > a$, we use the test $N_w^{x_0}(G) \overset{H_1}{\underset{H_0}{\gtrless}} \dfrac{an}{4} - \dfrac{a}{2} + C_1 \max\left(\sqrt{na\log(1/\delta)}, \log(1/\delta)\right).$

The risks of these tests can be controlled by separating the null and alternate ranges using Bernstein's inequality. Indeed, the threshold above is just the the expectation plus the concentration radius of the statistic under the null distribution. Let us briefly develop the statistic's behaviour in the alternate - considering only the case $a > b$, we find that under the alternate, $\binom{n-s}{2} + \binom{s}{2}$ of the edges in $N_a^{x_0}$ continue to behave like $\mathrm{Bern}(b/n)$ bits, while the remaining $s(n-s)/2$ edges behave as $\mathrm{Bern}(a/n)$ bit. Thus, the expectation of $N_a^{x_0}$ is increased by an amount greater than $s(n-s)\frac{a-b}{2n} \geq s(a-b)/4$. Next, Bernstein's inequality controls the fluctuations at scale $\sqrt{\max(nb, s(a-b))\log(2/\delta)}$. The conclusion is straightforward to draw from here, and the proof is carried out in Appendix A.1.1.

Proof Sketch for the Converse. The proof is relegated to Appendix A.1.2, and we discuss the strategy here. The converse proof follows Le Cam's method, which lower bounds the minimax risk by the Bayes risk for conveniently chosen priors - which can be expressed using the TV distance.

To show $\Lambda \gtrsim \log(1 + n/s^2)$, we pick the null x_0 to be any balanced community, and choose the uniform prior on communities that are exactly s-far from x_0 (in fact, we only use a subset of these in order to facilitate easier computations). This is an obvious choice for this setting - we are interested in balanced communities that are at least s far, and choosing a large number of them allows for a greater 'confusion' in the testing problem due to a richer alternate hypothesis. The bound follows by invoking inequalities between TV and χ^2 divergences and a lengthy calculation due to the combinatorial objects involved.

To show $s\Lambda \gtrsim -\log(\delta)$, we again pick the null to be any balanced community, and pick the alternate to be an s-far singleton. It then proceeds to control d_{TV} by the Hellinger divergence.

2.3 Two-Sample Testing

We again begin with the main results on *community two-sample testing problem.*

Theorem 2.3.1. *Assume, for some $\gamma > 0$, $s \geq n^{\frac{1}{2}+\gamma}$. There exist constants C, C' such that if $C' \leq a, b \leq (n/2)^{1/3}$, then two-sample testing of s changes with $R_{\mathrm{TST}} \leq 4/n$ is possible if the SNR satisfies $\Lambda \geq C$.*
Conversely, for $n \geq 200$, there exist constants c, c' such that if $s < (\frac{1}{2} - c')n$, then two-sample testing of s changes cannot be carried out with $R_{\mathrm{TST}} \leq 1/4$ unless $\Lambda \geq c$.

Large Changes. The above theorem makes an achievability claim for the setting of large changes. Notice that in this regime the stated upper and lower bounds match up to constants. Specifically, if $n^{\frac{1}{2}+\gamma} < s < (\frac{1}{2} - c')n$, two-sample testing can be solved iff $\Lambda \gtrsim 1$. Further, the condition $a, b \gtrsim 1$ is also tight, as it follows from $a/b = \Theta(1)$, and the necessary condition $\Lambda \gtrsim 1$, since $\Lambda \leq a + b$.

This leaves the condition $\max(a, b) \leq (n/2)^{1/3}$, which we suspect is an artifact of the proof technique and conjecture that, even for our proposed test, it can be removed. In any case, observe that this condition is irrelevant in the setting $a, b = O(\log n)$ considered in this chapter. Further, if a/b is bounded away from 1, then TST is directly possible when $a, b = \Omega(\log n)$ by recovering the communities and comparing them, demonstrating that this condition is not present in general.

Small Changes. We claim that for small changes - $s < n^{\frac{1}{2}-\gamma}$ for some $\gamma > 0$ - the naïve scheme of recovering the communities and comparing them is minimax. To see this, note that that GoF testing is reducible to TST - given a TST scheme of a known risk, one may construct a GoF tester of that risk by feeding the TST algorithm the observed graph and a graph drawn from $P(\cdot|x_0)$. Thus, the lower bounds of Theorem 2.2.1 apply to TST, and for $a/b = \Theta(1)$, we find that it is necessary that $s\Lambda = \omega(1)$ and that $\Lambda \gtrsim \log(1 + n/s^2)$ to attain vanishing R_{TST}. For small s, the latter lower bound is $\Omega(\log n)$, the claim follows since recovery with up to s errors is possible if $\Lambda \gtrsim \log n$.

Efficiency. Finally, we point out that the above bounds can be attained with computationally efficient tests. Further, for large changes, the test can be made agnostic to knowledge of (a, b). Instead, it only requires one to be able to estimate $n(a + b)$ to within an additive error of $\widetilde{O}(\sqrt{n(a + b)})$, which can be done by simply counting the number of edges in the graphs.

Proof Sketch of the Achievability. We describe the proposed test, and sketch its risk analysis below, completing the same in Appendix A.2.1. Recall the definition of N_w^z, N_a^z from (2.3) in §2.2, and let

$$T^{\hat{x}}(G) := N_w^{\hat{x}}(G) - N_a^{\hat{x}}(G). \tag{2.4}$$

We show that the routine 'TwoSampleTester' below attains a risk smaller than $4/n$. In words, the test computes a partition $\hat{x}$ for the graph G by using about half the edges in the graph. This is represented in the 'PartialRecovery' step below, for which any such method may be used - concretely, that of [CRV15]. Next, we compute the statistic $T^{\hat{x}}$ above for both the remaining part of the first graph, and for the second graph. Notice that unlike the GoF statistic, which was only N_a, $T^{\hat{x}}$ takes the difference of N_a and N_w. This is necessary because the partition $\hat{x}$ derived from partial recovery cannot be very well correlated with the true partition x. This means the reduced fluctuations from only considering one part does not apply, and we instead use the whole cut.

Algorithm 1 TwoSampleTester(G, H, δ)

```
$G_1 \leftarrow$ subsampling of edges of $G$ at rate $1/2$ uniformly at random.
$\widetilde{G} \leftarrow G - G_1$.
$\widehat{x} \leftarrow$ PartialRecovery$(G_1)$.
Compute $T^{\widehat{x}}(\widetilde{G}), T^{\widehat{x}}(H)$.
$T \leftarrow 2T^{\widehat{x}}(\widetilde{G}) - T^{\widehat{x}}(H)$.
Return $T \underset{H_0}{\overset{H_1}{\gtrless}} \sqrt{Cn(a+b)\log(6n)}$.
```

Since the edges within communities, and across communities in the graph are (separately) exchangable, the errors made in $\hat{x}$ distribute uniformly over the two communities[1]. This allows us to explicitly control the behaviour of T as defined in the test *provided $\hat{x}$ is non-trivially correlatd with x* - i.e., given that it makes $< (1/2 - c)n$ errors for some $c > 0$. The condition $\Lambda \gtrsim 1$ in the theorem arises from this.

A complication in this strategy is that the remaining graph $\widetilde{G}$ in the scheme is not independent of the recovered community $\hat{x}$. This is handled in the analysis by introducing an independent copy of G, called G', and arguing that $T^{\hat{x}}(\widetilde{G}) \approx 1/2 T^{\hat{x}}(G')$. This step is the origin of the nuisance condition $\max(a, b) \lesssim n^{1/3}$ in the theorem.[2]

Proof Sketch of the Converse. The argument uses Le Cam's method, but with the twist that the null model is chosen to be a two-step procedure - one that draws a balanced community uniformly at random, and then generates a graph according to it, while the alternate models are drawn uniformly from the balanced communities that are at least s-far from the chosen null. This allows a comparison to the unstructured Erdős-Rényi graph on n vertices with mean degree $(a+b)/2$. Bounds can then be drawn in from the study of the so-called distinguishability problem, and we invoke results from [WX18] to show that total variation distance between the null and alternate distributions is small when Λ is a small enough constant, allowing us to conclude using Neyman-Pearson. See Appendix A.2.2 for a detailed argument.

[1] For a proof: since $x, -x$ induce the same law, and since the communities are balanced, for every realization of G such that $\hat{x}$ makes e_+, e_- errors in the community $+, -$ respectively, there is a realization of equal probability where it makes e_-, e_+ errors. Further, within community exchangability implies that errors distribute uniformly.

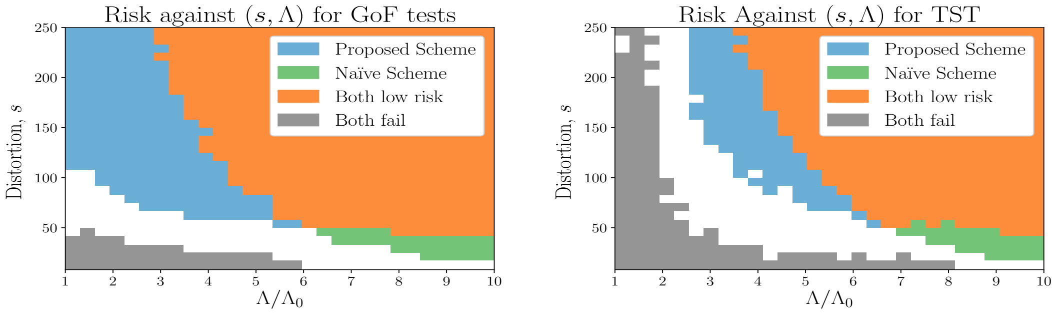
[2] This is something of a cliché by now, but I'd like to know that someone has read this whole bloody thing. So, I'll send fifty bucks to the first four people to email me claiming to have done so - attach a screenshot of Algorithm 1 for proof.

2.4 Experiments

We perform three different sets of numerical experiments. We first run our tests on SBMs with 1000 nodes. Next, we demonstrate that our tests perform similarly for a real dataset, specifically the Political Blogs dataset [AG05]. Finally, we examine SBM-supported Gaussian Markov Random Fields (GMRFs) as an example of a "node observation" model, where the SBM-generated edges form the precision matrix for the Gaussian vector consisting of the random variables assigned to each node. In particular, we need to determine if the underlying community of the graph has changed without explicitly observing (or recovering) the edges of the graph. For the sake of brevity, precise details of the experiments are moved to Appendix A.3.

2.4.1 SBM Experiments

We perform experiments implementing our GoF and TST strategies as well as the naïve scheme of reconstructing communities and comparing. Recovery is performed by regularised spectral clustering, for which a detailed description is given in Appendix A.3.1. The graphs are drawn on $n = 1000$ nodes for a range of (s, Λ) pairs and the high and low risk regimes are plotted in Figure 2·1. First, note that for 'large changes,' $s \geq \sqrt{n \log(10)} \approx 50$, our GoF and TST tests can succeed for lower SNR values. In contrast, for 'small changes,' $s < \sqrt{n} \approx 30$, the naïve test is more powerful in the high SNR regime. Additionally, both tests fail for TST unless the SNR is larger than a constant, as predicted by our lower bound in Theorem 2.3.1.


Risk against (s, Λ) for GoF tests

Risk Against (s, Λ) for TST

Figure 2·1: Risks of the proposed tests from sections 2.2 and 2.3 for GoF and TST respectively, and the performance of the naïve scheme, on synthetic SBMs with $n = 1000, a/b = 3$. Both schemes attain high risk ($> 1 - \delta$) in the grey region, intermediate risk in the white, and the colours indicate which of the schema attain low risk ($< \delta$), where $\delta = 0.01$ for GoF and $\delta = 0.1$ for TST.

2.4.2 Political Blogs Dataset [AG05]

The political blogs dataset [AG05] is canonical in the study of community detection, and consists of $n = 1222$ nodes. Here, we vary the effective SNR by randomly subsampling the edges of the graphs at rate ρ. See Appendix A.3.2 for further details. In this dataset, the ground truth partition x_{True} is available, which in turn yields accurate estimates of the connectivity probabilities (a, b). For this graph $a/b \approx 10$. Further, spectral clustering alone incurs ≈ 50 errors in this graph, which is larger than $\sqrt{1222} \approx 35$. As a consequence, the behaviour in the 'small changes' regime where the test relies on recovery - is not well illustrated in the following.

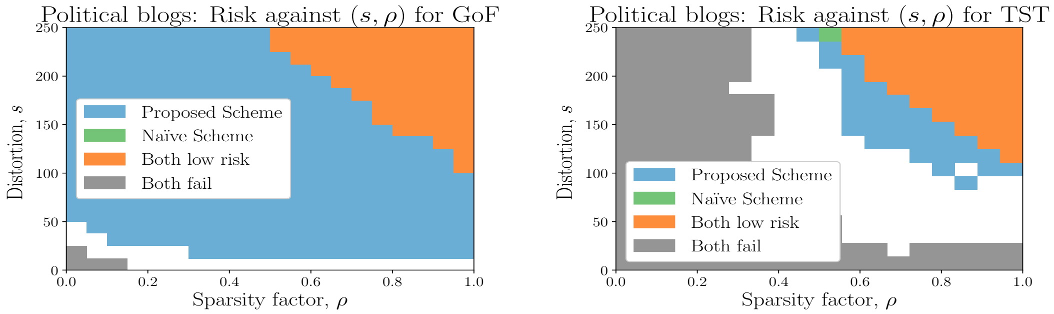


Figure 2·2: Risks of the tests applied to the Political Blogs graphs - colour scheme is retained from Fig. 2·1. The X-axis plots the sparsification factor, which serves as a proxy for SNR. Features similar to Fig. 2·1 can be seen. The GoF plot improves since a/b is bigger, while the TST plot suffers since the political blogs graph is not completely described as a 2-community SBM [Lei16].

Goodness-of-Fit. We determine the size of the test by running the GoF procedures against x_{True}. To determine power, we construct a partition y by relabelling a random set of nodes of size s, and running the GoF procedures against y *with the same graph.*

Two-Sample Testing. We compare the political blogs graph G against two other graphs drawn from SBMs. Size is detemined by drawing G' according to an SBM of community x_{True} and running the TST procedure, and power is determined by drawing a y as above, generating H according to an SBM of community y, and running the TST procedure. Note that this experiment is thus semi-synthetic.

2.4.3 Gaussian Markov Random Fields (GMRFs)

Frequently instead of simply receiving a graph, one receives i.i.d. samples from a graph-structured distribution, and it is of interest to be able to cluster nodes with respect to the latent graph. For example, in large-scale calcium imaging, it is possible to simultaneously record the activity pattern of thousands of neurons, but not their underlying synaptic connectivity [Pne+16]. Here, we explore the behavior of our tests for GMRFs where the underlying graph structure is randomly drawn from an SBM and and we only observe the nodes.

A heuristic reason for why our methods might succeed in such a situation arises from the local tree-like property of sparse random graphs (see, e.g. [DM10]). For graphs with mean degree $d \ll n$, typical nodes do not lie in cycles shorter than $\sim \frac{\log n}{2 \log d}$. In MRFs, this tree-like property induces correlation decay: the correlation between two nodes decays geometrically up to graph-distance $\sim \frac{\log n}{2 \log d}$. Thus, the covariance matrix closely approximates $\sigma_1 G + \sum_{i=2}^{k} (\sigma_1 G)^i + \sigma_0 \mathbf{1}\mathbf{1}^\top$ for some $\sigma_0 \ll \sigma_1$, small k, and G, the adjacency matrix of the graph. Since the local structure of the graph is so expressed, both clustering and testing applied directly to the covariance matrix should be viable.

We report experimentation on the GMRF (see, e.g. [WJ08, Ch. 3]), which comprises random vectors $\zeta \sim \mathcal{N}(0, \Theta^{-1})$, where the non-zero entries of the precision matrix Θ encode the conditional dependence structure of ζ. Using standard parametrisations [WWR10], we set $\Theta = I + \gamma G$, where $G \sim P(G|x)$ is an adjacency matrix from an SBM with latent parameter x, and γ is a scalar. Below, we fix the SBM parameters a, b and the level γ, and explore risks against s and sample size t.

Following the above heuristic, we naïvely adapt community recovery and testing to this setting, by replacing all instances of the graph adjacency matrix in previous settings with the sample covariance matrix. Figure 2·3 presents our simulations of the risk of this test when $n = 1000$, and $(a, b) \approx (12.3 \log n, 1.23 \log n)$, at $\Lambda \approx 9 \log(n)$ (for details see Appx. A.3.3). This large SNR is chosen so that community recovery would be easy if the graph was recovered;[3] this emphasizes the role of the sample size, t. Importantly, in this implementation, the threshold for rejecting the null has been fit using data (unlike in the previous sections). This is since we lack a rigorous theoretical understanding of this problem, and have not analytically derived expressions for

[3] Note, however, we expect graph recovery to be impossible at these sample sizes. Lower bounds from [WWR10] indicate this would require > 3300 samples theoretically.

the thresholds. As a result, these plots should be treated as speculative research intended to underscore the presence of interesting testing effects in this scenario, and to encourage future work along these lines.

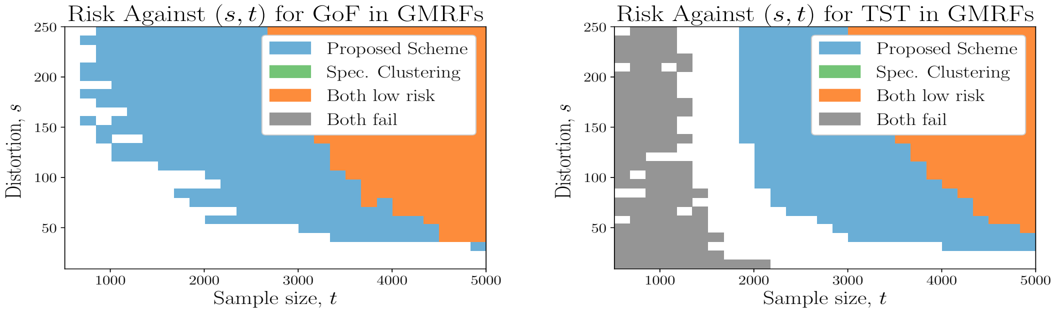


Figure 2·3: Risks for adaptation of our tests to GMRFs - colour scheme is retained from Fig. 2·1. The plots show structural similarity to Fig. 2·1, but with two differences - In GoF, we don't find a high risk region at the sample sizes considered, and the proposed scheme always outperforms the Naïve scheme based on spectral clustering.

2.5 Directions For Future Work

Our analysis of the SBM establishes the basic fact that testing of communities is possible far below the recovery threshold at a comparable granularity if and only if the changes are large enough. Below, we propose a number of refinements of the approach in the chapter, to broaden the characterisation to greater parameter regimes with a richer analysis, and to improve upon the algorithms developed therein. Before proceeding, we acknowledge that the problem of determining exact constants in the various SNR bounds has also been left open in the above.

More Communities Most realistic networks are considerably better described as block models with $k > 2$ communities as opposed to with only two communities. This is even the case in networks with two broad categories - indeed, as noted in [Lei16], this is also the case for the popular political blogs dataset [AG05] even though this consists of ground labelling of blogs as Democratic or Republican.

The case of the SBM with many communities is non-trivially different from that of the two community SBM. The most important difference is due to computational effects - for $k \geq 4$, it is known that the information theoretic threshold for recovery and the efficient threshold, as described by the achievability region of the message passing schemes (or the Kesten-Stigum threshold) separates [Abb18].

With this in mind, it is natural to study the GoF and TST problems for k-community symmetric SBMs as a basic case. There are two main challenges. First, our lower bounds do not extend to this regime, and it is unclear if the analysis along the same lines can be run in a simple manner. Second, while our achievability schemes admit natural extensions to this regime, these extensions are efficient, and so it is unclear whether they would be powerful up to the information theoretic limit of testing, or if they would suffer a similar computational separation.

Refined Analyses in various parameter regimes We discuss two natural situations. The first regards strongly imbalanced community structures. The theory of §2 extends naturally to 2-SBMs where both communities have size linear in n. Does the same occur for imbalanced settings, where one of the communities is very small (of size $o(n)$)? One challenge with this setting is that the recovery of imbalanced communities is poorly understood - to our knowledge, optimal recovery thresholds are unknwon when the smaller community has size $o(n/\log n)$.

Second, the theory of the above chaper concentrates on the regime $a/b = \Theta(1)$. It is unclear how the testing problems behave in the easier setting where $\rho := \frac{\max(a,b)}{\min(a,b)}$ diverges with n. Preliminary work, discussed in A.1.3, shows that for large change GoF, the testing threshold for SBMs is controlled by $\mu := \frac{(a-b)^2}{\min(a,b)} = (1+\rho)\Lambda \gg \Lambda$ - for large changes, reliable GoF is possible if and only if $\mu \gtrsim {}^{n}/_{s^2}$. For TST, the lower bound of Theorem 2.3.1 continues to hold even when ρ diverges, and thus the TST story for large changes remains the same. However, it is unknown if this is also the case for small changes, since the lower bound for this setting utilises the lower bound of Theorem 2.2.1. We propose to study this effect for small changes in an both to determine the correct notion of SNR for testing, and to investigate if this is different from the SNR for recovery.

Improved Schemes The TST scheme described in Algorithm 1 uses a partial recovery step to seed what is effectively a GoF statistic. This is both aesthetically dissonant, and statistically profligate - the former because the study of testing should be possible independently of any knowledge of partial recovery thresholds, and the latter because passing through such a recovery step inevitably requires some subsampling of the graph, thus degrading the SNR available for testing. This motivates the question of if a TST scheme that does not pass through partial recovery can be constructed.

A promising direction for this is afforded by the statistic $\max_R \langle \mathbf{1}_R, G - H\rangle$, where $\mathbf{1}_R$ for a set $R \subset [1:n] \times [1:n]$ is the matrix with entries 1 for coordinates in R, the inner product is $\langle A, B\rangle = \mathrm{Tr}(AB)$, and the maximisation is carried over combinatorial rectangles in $[1:n] \times [1:n]$ of dimension ${}^{n-s}\!/_2 \times {}^{s}\!/_2$. The intuition behind this lies in the fact that for G and H with differing communities, $\mathbb{E}[G-H]$ develops a block structure with two (combinatorial) rectangles of the above dimensions taking values ${}^{+(a-b)}\!/_n$ and ${}^{-(a-b)}\!/_n$. The statistic thus investigates the existence of such a rectangle with a large weight. The main challenge here is controlling the size of this test, as the setting requires control of about $n^{O(s)}$ rectangles in the null. A natural tool to exploit here is that of the generic chaining, since the structure of the objective is precisely that of a Bernoulli chaos. A more promising initial direction may be to study such schemes for the spiked Wigner model, which is a Gaussian analogue of the SBM that typically admits somewhat easier analysis.

Testing in Richer Models of Random Networks The SBM has a number of well known shortcomings when it comes to modeling real world networks. The primary of these shortcomings is that the SBM does not capture heterogeneity in the degrees of different agents, and that it does not capture homophily effects that tend to increase the number of small motifs such as triangles in real world social networks compared to the level obtained by the Erdős-Rényi based structure of the SBM.

A number of models have been proposed to rectify these structural properties, and have been analysed in the recent literature. These include degree-corrected SBMs [KN11; GMZZ18], which include node dependent latent variables modulating the degree of each node, or more geometrically oriented graph models, such as the Geometric Block Model [GPMS18; GMPS19] or labelled Euclidean Random Graphs [ABS17; SB17], each of which captures structural features such as over-representation of triangles by planting an underlying geometric structure in the model generation. The study of testing in these more realistic models would lead to a more representative theory for real world community testing applications.

In a similar vein to the above, in the real world community boundaries are often fuzzy, and a node may belong partially to more than one community. Models capturing such effects have been proposed and analysed, although considerably less is known about these settings than about the SBM [NP15; ABFX08; Abb18, §3.1]. Further extensions may be pursued by considering settings where edge labels beyond existence are given, or where edges are censored. We refer the reader to [Abb18, §2.4]. Testing of such structures is a wide open and interesting question.

Generic Testing in the Graphical Channel Setup Consider the following view of the SBM: n binary message bits $x_1, \dots, x_n \in \{0, 1\}$ are passed through an asymmetric memoryless binary input binary output channel

$$p_n(g = 1|z = 0) = a/n$$
$$p_n(g = 1|z = 1) = b/n$$

via the rate $2/n$ code $z_{ij} = x_i \oplus x_j$. This clearly demarcates the SBM as an information theoretic problem, but with three twists - asymptotically in the blocklength, the rate at which the message is passed is vanishing, the channel code is fixed, and further, the statistics of the channel depend on the blocklength.

Graphical channels, proposed by Abbe & Montanari [AM13], generalise this view to consider an arbitrary code that is encapsulated by a hypergraph - the message bits are associated with the nodes of a hypergraph, and each edge e in the same has an attached variable z_e. The z_e form an encoding of the message, which is passed through a DMC, and the underlying problem is that of inference of the message from this output. This setup is a considerable generalisation, encapsulating not only does the SBM and its many variants, but also hypergraphical block models, random constraint satisfaction problems with planted solutions, and various noise models with planted signal. Generic results on the the graphical channel are limited, although we refer the reader to the excellent presentation of the exact recovery problems over the same in [Kim18] along with the aforementioned paper. The critical features of the graphical channel relative to standard communication problems are precisely those observed in the SBM - the code pushed through the channel is fixed (i.e., this is a problem of communication without coding), the rate of communication is vanishing, and that the channel's capacity may decay with the blocklength.

Testing in the graphical channel setup is an unexplored question, and, in our opinion, the broad question of characterising the distortion to which GoF and TST can be performed across generic graphical channels is very interesting. The analogous problem with coding is that of identification via channels [AD89], and thus, just as recovery in graphical channels is a 'communication without coding' problem, testing is an identification without coding problem. This was considered for the special case of the graphical channel with only self loops (i.e., when the message is directly transmitted) by JaJa [JaJ85] in the 80's, although not much has developed since[4]. We note that the observations of [JaJ85] align with ours - they find that for any non-zero rate channel, identification is possible. More precisely translated, the results correspond to saying that for channels that do not depend on the blocklength, one can do GoF testing of the message at distortion level $O(\sqrt{n})$.

[4] more accurately, attention shifted to identification *with* coding, as explored in [AD89], and, to our knowledge, the connection with GoF testing has not been pursued.

Chapter 3

Testing Network Structures of Ising Models

3.1 Introduction

The Ising model is a canonical pairwise graphical model, which captures settings where the behaviour of many agents is governed in a pairwise manner by an underlying network that typifies which agents interact. This makes it a natural object of study when considering situations such as inference of a connectome via calcium fluorescence studies of neuron populations. This setting is much harder than the SBM testing and recovery scenarios described in the previous section. This is because of two reasons - first, instead of directly observing the edges of the underlying graph, we are only observing node level information. Secondly, instead of reasoning about a low dimensional structure like a planted partition, the inference task is focused on the entire graph itself.

As always, a baseline approach for testing is to estimate the network, and then compare the differences. However, such observations exhibit significant variability, and the amount of data available may be too small for this approach to yield meaningful results. On the other hand, *reliably recovering network changes should be easier than*

This chapter is a lightly edited reproduction of the paper [GNS20]. Weaker versions of the results discussed below appeared previously in [GNS17; GNS18]. All of these papers were written in collaboration with Bobak Nazer and Venkatesh Saligrama. Additionally, I would like to thank Bodhi Vani and Anil Kag for discussions that helped with the simulations described in §B.3.3, on which Figure 3·3 is based.

full reconstruction. While prior works have proposed inference algorithms to explore this possibility [ZCL14; XCC15; FB16; BVB16; BZN18; Zha+19; CLMX19], we do not have a good mathematical understanding of when this is indeed easier.

To shed light on this question, we propose to derive information-theoretic limits for two structural inference problems over degree-bounded Ising models. The first is goodness-of-fit testing (GoF). Let $G(P)$ be the network structure (see §3.2) of an Ising model P. GoF is posed as follows.

> GoF : *Given an Ising model P and i.i.d. samples from another Ising model Q, determine if $P = Q$ or if $G(P)$ and $G(Q)$ differ in at least s edges.*

The second is a related estimation problem, termed error-of-fit (EoF), that demands localising differences in $G(P)$ and $G(Q)$ (if distinct).

> EoF: *Given an Ising model P and i.i.d. samples from another Ising model Q that is either equal to P, or has a network structure that differs from that of P in s edges or more, determine the edges where $G(P)$ and $G(Q)$ differ.*

Notice that the above problems are restricted to models that are either identical, or significantly different. 'Tolerant' versions (separating small changes from large) are not the focus for us (although we discuss this setting for a special case in §3.4). The main question of interest is: *For what classes of Ising models is the sample complexity of the above inference problems significantly smaller than that of recovering the underlying graph directly?*

Contribution. We prove the following surprising fact: up to relatively large values of s, the sample complexities of GoF and EoF are *not* appreciably separated from that of structure learning (SL). Our bound is surprising in light of the fact that prior works [Liu+14; Liu+17; FB16; KLK19; CLMX19] propose algorithms for GoF and

EoF, and claim recovery of *sparse* changes is possible with sample complexity much smaller than SL. Concretely, for models with p nodes, degrees bounded by d, and non-zero edge weights satisfying $\alpha \le |\theta_{ij}| \le \beta$ (see §3.2), the sample complexity of SL is bounded as $O(e^{2\beta d}\alpha^{-2}\log p)$. We show that if $s \ll \sqrt{p}$, then the sample complexity of GoF is at least $e^{2\beta d - O(\log(d))}\alpha^{-2}\log p$, and that if $s \ll p$, then the sample complexity of EoF has the same lower bound. We further show that the same effect occurs in the restricted setting of detecting edge deletions in forest-structured Ising models, and, to some extent, in detecting edge deletions in high-temperature ferromagnets. In the case of forests, we tightly characterise this behaviour of GoF, showing that for $s \ll \sqrt{p}$, GoF has sample complexity comparable to SL of forests, while for $s \gg \sqrt{p}$, it is vanishingly small relative to SL. For high-temperature ferromagnets, we show that detecting changes is easier than SL if $s \gg \sqrt{pd}$, while this does not occur if $s \ll \sqrt{pd}$. These are the first structural testing results for edge edits in natural classes of Ising models that show a clear separation from SL in sample complexity.

Technical Novelty. The lower bounds are shown by constructing explicit and flexible obstructions, utilising Le Cam's method and χ^2-based Fano bounds. The combinatorial challenges arising in directly showing obstructions on large graphs are avoided by constructing obstructions with well-controlled χ^2-divergence on small graphs, and then *lifting* these to p nodes via tensorisation in a process that efficiently deals with combinatorial terms. The main challenge is obtaining precise control on the χ^2-divergence between graphs based on cliques, which is attained by an elementary but careful analysis that exploits the symmetries inherent in Ising models on cliques. The most striking instance of this is the 'Emmentaler clique' (Fig. 3·2), which is constructed by removing $\Theta(d^2)$ edges from a d-clique in a structured way. Despite this large edit, we show that it is exponentially hard (in low temperatures) to distinguish this clique with large holes from a full clique.

3.1.1 Related Work

Statistical Divergence Based Testing. Related to our problem, but different from our setup, GoF of Ising models has been studied under various statistical metrics such as the symmetrised KL divergence [DDK19] and total variation [Bez+19]. More refined results and extensions have appeared in [GLP18; DDK17; CDKS17; ABDK18]. These are tests that certify whether or not a particular statistical distance between two distribution is larger than some threshold. In contrast, our focus is on *structural* testing and estimation, namely, whether or not the change in the network is a result of edge-deletions or edge-additions. As such, statistically-based GoF tests do not have a direct bearing on structural testing. Divergences can be large in structurally irrelevant ways, e.g., if a few isolated nodes in a large graph become strongly interacting, a large KL divergence is induced, but this is not a significant change in the network on the whole (Also see §B.5.1). In light of applications which demand structure testing as a means to interpret phenomena, and this misalignment of goals, testing in the parameter space is compelling, and testing the network is the simplest instance of this.

Sparse-Recovery-Based Structural Testing Methods. More directly related to our work, are those that are based on direct change estimation (*DCE*) [FB16; Liu+14; Liu+17; LFS17; KLK19], which attempt to directly characterize the difference of parameters $\delta^* = \theta_P - \theta_Q$ by leveraging sparsity of δ^*. These works leverage the 'KL Importance Estimation Procedure' (KLIEP), the key insight of which is that the log-likelihood ratios can be written in a form that is suggestive of expressions from sparse-pattern recovery methods, to define the empirical loss function

$$\mathcal{L}(\delta) = -\langle \delta, \hat{\mathbb{E}}_Q[XX^T] \rangle + \log \hat{\mathbb{E}}_P[\exp\left(X^T \delta X\right)],$$

where $\hat{\mathbb{E}}$ denotes an empirical mean, and δ is sparse. The second term, which is the only non-linear term, is reminiscent of normalization factors in graphical models. In this context, it is useful to recall the key ideas from high-dimensional sparse estimation theory (see [NRWY12]), which has served as a powerful generic tool. At a high-level, these results show that for a loss function $\mathcal{L}(\delta)$ paired with a decomposable regulariser (such as an ℓ_1 norm on δ), if the loss function satisfies restricted strong convexity, namely, strong convexity only in a suitable descent error set, as characterised by the regulariser and the optimal value δ^*, minimising the penalised empirical loss leads to a non-trivial estimation error bound. Leveraging these concepts of high-dimensional estimation, and exploiting sparsity, the sparse DCE works show that testing can be done in $O(\text{poly}(s) \log p)$ samples (for any P, Q!), which is further much smaller than the number needed for SL, a result which contradicts bounds we derive in this chapter. The situation warrants further discussion.

From a technical perspective, the sample complexity gains of these methods arise from assuming law-dependent quantities to be constants. For example, [Liu+14; Liu+17] require that for $\|u\| \le \|\delta^*\|, \nabla^2 \mathcal{L}(\delta^* + u) \preccurlyeq \lambda_1 I$, and that for S the support of δ^*, the submatrix $(\nabla^2 \mathcal{L}(\delta^*))_{S,S} \succcurlyeq \lambda_2 I$, where λ_1, λ_2 are constants independent of P, Q. [FB16] removes the second condition, and shows that $\mathcal{L}$ has the λ_2-RSC property, where λ_2 is claimed to be independent of P, Q. In each case, sample costs increase with λ_1 and λ_2^{-1}. However, the assertion that λ_1, λ_2 are independent of (P, Q) cannot hold in general – the only non-linear part in $\mathcal{L}$ is $\log \hat{\mathbb{E}}_P[\exp\left(X^T \delta X\right)]$, which clearly depends on P! This dependence also occurs if P is known. Thus, the 'constants' λ_1, λ_2 are affected by the properties of P. More generically, the efficacy of sparse recovery techniques is questionable in this scenario. Since the data is essentially distinct across samples, and internally dependent, and since the sparse changes, δ^*, and the underlying distributions interact, it is unclear if meaningful notions of design matrix that allow testing with sub-recovery sample costs can be developed.

Nevertheless, it is an interesting question to understand what additional assumptions on P, Q or topological restrictions are useful in terms of benefiting from sparsity. Our results suggest that these conditions are stronger than typical incoherence conditions such as high temperatures, and further that the topological restrictions demand more than just ‘simplicity’ of the graphs.

Other Methods.[CLMX19] propose a method, whereby the parameters θ_P and θ_Q are only crudely estimated, and then tests using the biggest (normalised) deviations in the estimates as a statistic. The claims made in this paper are more modest, and do not show sample complexity below n_{SL}. We point out, however, that d-dependent terms are treated as constants in this as well.

Much of the structural testing work studies Gaussian GMs instead of Ising (see the recent survey [Sho20]). We do not discuss these, but encourage the same careful examination of their assumptions.

Structural Testing Extensions. A number of structural testing problems other than GoF have been pursued. For instance, [BN18] tests if the model is mean field or supported on a structured graph (sparse, etc.), [BN19] tests mean-field models against those on an expander, [CNL18] tests independence against presence of structure in high temperatures, [NL19] tests combinatorial properties of the underlying graph such as whether it has cycles, or the largest clique it contains (also see §B.5.2).

Structure Learning The structure learning literature is by now quite expansive, with many recent efficient algorithms with close-to-optimal sample complexity [KM17; HKM17; LVMC18; WSD19], and exploration of refined settings such as learning under corruptions [GKK19]. Detailed discussion of this literature would take up too much space, but we highlight [SW12] as the original paper to establish information-theoretic bounds for the same, and [BK20] for a neat analysis of the Chow-Liu algorithm, which are the only SL papers directly used in the following. Additionally, our methods offer improvements to the minimax lower bounds of [SC16] by improving the exponents in their $\exp(\Omega(\beta d))$ bounds.

3.2 Problem Definitions and Notation

The zero external field Ising Model specifies a law on a p-dimensional random vector $X = (X_1, \dots, X_p) \in \{\pm 1\}$, parametrised by a symmetric matrix θ with 0 diagonal, of the form

$$P_\theta(X = x) = \frac{\exp\left(\sum_{i<j} \theta_{ij} x_i x_j\right)}{Z(\theta)},$$

where $Z(\theta)$ is called the partition function. Notice that given X_j for all $j \in \partial i := \{j : \theta_{ij} \neq 0\}$, X_i is conditionally independent of $X_{[1:p]-\{i\}-\partial i}$. Thus, the θ determine the local interactions of the model. With this intuition, one defines a simple, undirected graph $G(P_\theta) = ([1:p], E(P_\theta))$ with $E(P_\theta) = \{(i,j) : \theta_{ij} \neq 0\}$. This graph is called the *Markov network structure* of the Ising model, and θ can serves as a weighted adjacency matrix of $G(P_\theta)$. We often describe models by an unweighted graph, keeping weights implicit until required.

The model above can display very rich behaviour as θ changes, and this strongly affects all inference problems on Ising models. With this in mind, we make two explicit parametrisations to help us track how θ affects the sample complexity of various inference problems. The first of these is degree control - we assume that the degree of every node is $G(P), G(Q)$ is at most d. The second is weight control - we assume that if $\theta_{ij} \neq 0$, then $\alpha \leq |\theta_{ij}| \leq \beta$.

These are natural conditions: small weights are naturally difficult to detect, while large weights mask the nearby small-weight edges; degree control further sets up a local sparsity that tempers network effects in the models. The class of laws so obtained is denoted $\mathcal{I}_d(\alpha, \beta)$. We will usually work with a subclass $\mathcal{I} \subset \mathcal{I}_d$ which has *unique network structures* (i.e., for $P, Q \in \mathcal{I}, G(P) \neq G(Q)$). Note that we do not restrict α, β, d to have a particular behaviour - these are instead used as parametrisation to study how weights and degree affects sample complexity. In particular, they may vary with p and each other. We do demand that $d \leq p^{1-c}$ for some constant $c > 0$, and that p is large ($\gg 1$).

We let $\mathcal{G}$ be the set of all graphs on p nodes, and $\mathcal{G}_d \subset \mathcal{G}$ be those with degree at most d. The symmetric difference of two graphs G, H is denoted $G \triangle H$, which is a graph with edge set consisting of those edges that appear in exactly one of G and H.

Lastly, we say that two Ising models are *s-separated* if their networks differ in at least s edges. The 'anti-ball' $A_s(P) := \{Q \in \mathcal{I} : |G(Q) \triangle G(P)| \geq s\}$ is the set of $Q \in \mathcal{I}$ s-separated from P.

3.2.1 Problem Definitions

Below we define three structural inference problems: goodness-of-fit testing, error-of-fit identification, and approximate structure learning.

Goodness-of-Fit Testing Given P and the dataset $X^n \sim Q^{\otimes n}$ where $Q \in \{P\} \cup A_s(P)$, we wish to distinguish between the case where the model is unchanged, $Q = P$, and the case where the network structure of the model differs in at least s edges, $Q \in A_s(P)$. A goodness-of-fit test is a map $\Psi^{\mathrm{GoF}} : \mathcal{I} \times \mathcal{X}^n \to \{0, 1\}$. The n-sample risk is defined as

$$\begin{aligned} &R^{\mathrm{GoF}}(n, s, \mathcal{I}) \\ &:= \inf_{\Psi^{\mathrm{GoF}}} \sup_{P \in \mathcal{I}} \left\{ P^{\otimes n}(\Psi^{\mathrm{GoF}}(P, X^n) = 1) + \sup_{Q \in A_s(P)} Q^{\otimes n}(\Psi^{\mathrm{GoF}}(P, X^n) = 0) \right\}. \end{aligned}$$

Error-of-Fit Recovery Given P and the dataset $X^n \sim Q^{\otimes n}$ where $Q \in \{P\} \cup A_s(P)$ we wish to identify where the structures of P and Q differ, if they do. The error-of-fit learner is a graph-valued map $\Psi^{\mathrm{EoF}} : \mathcal{I} \times \mathcal{X}^n \to \mathcal{G}$. The n-sample risk is defined as

$$\begin{aligned} &R^{\mathrm{EoF}}(n, s, \mathcal{I}) \\ &:= \inf_{\Psi^{\mathrm{EoF}}} \sup_{P \in \mathcal{I}} \sup_{Q \in \{P\} \cup A_s(P)} Q^{\otimes n}\left(\left|\Psi^{\mathrm{EoF}}(P, X^n) \triangle \left(G(P) \triangle G(Q)\right)\right| \geq (s-1)/2\right). \end{aligned}$$

In words, Ψ^{EoF} attempts to recover $G(P) \triangle G(Q)$, and the risk penalises answers that get more than $(s-1)/2$ of the edges of this difference wrong. This problem is very similar to the following.

s-Approximate Structure Learning Given the dataset $X^n \sim Q^{\otimes n}$ we wish to determine the network structure of Q, with at most s errors in the recovered structure. A structure learner is a graph-valued map $\Psi^{\mathrm{SL}} : \mathcal{X}^n \to \mathcal{G}$, and the risk of structure learning is

$$R^{\mathrm{SL}}(n, s, \mathcal{I}) := \inf_{\Psi^{\mathrm{SL}}} \sup_{Q \in \mathcal{I}} Q^{\otimes n}(|\Psi^{\mathrm{SL}}(X^n) \triangle G(P)| \geq s).$$

The sample complexity of the above problems is defined as the smallest n necessary for the corresponding risk to be bounded above by $1/4$, i.e.

$$n_{\mathrm{GoF}}(s, \mathcal{I}) := \inf\{n : R^{\mathrm{GoF}}(n, s, \mathcal{I}) \leq 1/4\},$$

and similarly n_{EoF} and n_{SL} but with the risk lower bound of $1/8$.[1]

The above problems are listed in increasing order of difficulty, in that methods for SL yield methods for EoF, which in turn solve GoF. This is captured by the following statement, proved in §B.1.1.

Proposition 3.2.1. $n_{\mathrm{SL}}((s-1)/2, \mathcal{I}) \geq n_{\mathrm{EoF}}(s, \mathcal{I}) \geq n_{\mathrm{GoF}}(s, \mathcal{I})$.

Our main point of comparison with the literature on SL is the following result, which (mildly) extends [SW12, Thm 3a)] due to Santhanam & Wainwright. We leave

[1] $1/4$ is convenient for bounds for GoF, but any risk smaller than 1 is of interest, and can be boosted to arbitrary accuracy by repeating trials and majority. For EoF, SL we use $1/8$ for ease of showing Prop. 3.2.1.

the proof of this to Appx. B.1.2.

Theorem 3.2.2. *If $\mathcal{I} \subset \mathcal{I}_d(\alpha, \beta)$ has unique network structures, then for $s \le pd/2, \exists C \le 64$ such that*

$$n_{\mathrm{SL}}(s, \mathcal{I}) \le C \frac{de^{2\beta d}}{\sinh^2(\alpha/4)} \left(1 + \log \frac{p^2}{2s} + O(1/s)\right).$$

3.3 Lower Bounds for GoF and EoF over $\mathcal{I}_d(\alpha, \beta)$

This section states our results, and discusses our proof strategy, but proofs for all statements are left to §B.2. The bound are generally stated in a weaker form to ease presentation, but the complete results are described in §B.2. We begin by stating lower bounds for the case of $s = O(p)$. Throughout $500 > K > 1$ is a constant independent of all parameters.

Theorem 3.3.1. *If $20 \le d \le s \le {}^p/_K$, then there exists a $C > 0$ independent of (s, p, d, α, β) such that*

$$n_{\mathrm{GoF}}(s, \mathcal{I}) \ge C \max \left\{ \frac{e^{2\beta}}{\tanh^2 \alpha}, \frac{e^{2\beta(d-3)}}{d^2 \min(1, \alpha^2 d^4)} \right\} \log \left(1 + C \frac{p}{s^2}\right)$$
$$n_{\mathrm{EoF}}(s, \mathcal{I}) \ge C \max \left\{ \frac{e^{2\beta}}{\tanh^2 \alpha}, \frac{e^{2\beta(d-3)}}{d^2 \min(1, \alpha^2 d^4)} \right\} \log \left(C \frac{p}{s}\right)$$

This statement is enough to make our generic point - for small s (i.e., if $s \le p^{1/2-c}$ in GoF and if $s \le p^{1-c}$ in EoF), the above bounds are uniformly within a $O(\mathrm{poly}(d))$ factor of the the upper bound on n_{SL} in Theorem 3.2.2. Notice also that the max-terms are uniformly $\tilde{\Omega}(d^2)$ in the above - if $\beta d \ge 2 \log d$, then the second term in the max is $\Omega(d^2)$, while if smaller, the first term is $\Omega(({}^d/_{\log d})^2)$ because $\alpha \le \beta$. Thus, over $\mathcal{I}_d$, the best possible sample complexity of GoF and EoF scales as $\tilde{\Omega}(d^2 \log p)$, and in particular cannot be generally d-independent.

Of course, graphs in $\mathcal{G}_d$ have upto $\sim pd$ edges, and so many more changes can be made. Towards this, we provide the following bound for GoF. A similar result for EoF is discussed in §B.2.

Theorem 3.3.2. *If for some $\zeta > 0, s \leq pd^{1-\zeta}/K$, and $d \geq 10$, then there exists a constant $C > 0$ independent of (s, p, d, α, β) such that*

1. *If $\alpha d^{1-\zeta} \leq 1/32$ then $n_{\mathrm{GoF}} \geq C\frac{1}{d^{2-2\zeta}\alpha^2}\log\left(1 + C\frac{pd^{3-3\zeta}}{s^2}\right)$.*
2. *If $\beta d \geq 4\log(d-4)$ then $n_{\mathrm{GoF}} \geq C\frac{e^{2\beta d(1-d^{-\zeta})}}{d^2\min(1, \alpha^2 d^4)}\log\left(1 + C\frac{pd^{2-3\zeta}}{s^2}\right)$.*

Thm. 3.3.2 leaves a (small) gap, since as $\zeta \to 0$, $\alpha d^{1-\zeta} \leq 1$ and $\beta d \geq 4\log(d)$ do not completely cover all possibilities. Barring this gap, we again notice that for $s \ll \sqrt{pd^{1-\zeta}}$, n_{GoF} is separated from n_{SL} by at most a poly(d) factor. The first part of the above statement is derived using results of [CNL18]. For the limiting case of $\zeta = 0$, i.e. when s is linear in pd, we recover similar bounds, but with the distinction that the $2\beta d$ in the exponent is replaced by a βd. See §B.2.

Finally, since often the interest in DCE lies in *very sparse* changes, we present the following -

Theorem 3.3.3. *If $s \leq d$, then there exists a $C > 0$ independent of (s, p, d, α, β) such that*

$$n_{\mathrm{GoF}}(s, \mathcal{I}) \geq C\max\left\{\frac{e^{2\beta}}{\tanh^2\alpha}, \frac{e^{2\beta(d-1-2\sqrt{s})}}{d^6\sinh^2(\alpha\sqrt{s})}\right\}\log\left(1 + C\left(\frac{p}{s^2}\wedge\frac{p}{d}\right)\right)$$
$$n_{\mathrm{EoF}}(s, \mathcal{I}) \geq C\max\left\{\frac{e^{2\beta}}{\tanh^2\alpha}, \frac{e^{2\beta(d-1-2\sqrt{s})}}{d^6\sinh^2(\alpha\sqrt{s})}\right\}\log\left(C\frac{p}{d}\right)$$

Structure of the Bounds Each of the bounds above can be viewed as of the form $(\mathrm{SNR})^{-1}\log(1 + f(p, s, d))$, where we call the premultiplying terms SNR since they naturally capture how much signal about the network structure of a law relative to its fluctuations is present in the samples. This SNR term in Thms. 3.3.1 and 3.3.3 is developed as a max of two terms. The first of these is effective in the high temperature regime (where βd is small), while the second takes over in the low temperature regime of large βd. Similarly, the first and second parts of Thm. 3.3.2 are high and low temperature settings, respectively, and have different SNR terms. The SNR in all of the above is within a $\mathrm{poly}(d)$ factor of the corresponding term in the upper bound for n_{SL}.

The term $f(p,d,s)$ thus captures the hardness of testing/error localisation. For EoF, as long as s is small, this term takes the form p^c for some c. Thus, generically, localising sparse changes is nearly as hard as approximate recovery. This is to be expected from the form of the EoF problem itself. More interestingly, for GoF, these take the form pd^c/s^2. When $s \ll \sqrt{pd^c}$, this continues to look polynomial in p, and thus GoF is as hard as recovery. On the other hand, for s much larger than this, f becomes $o(1)$ as p grows, and so $\log(1+f) \approx f$ itself and the resulting bounds look like $(\mathrm{SNR})^{-1}pd^c/s^2$. In the setting of low temperatures with non-trivially large degree, these can still be super-polynomial in p, but relative to n they are essentially vanishing.

Notice that in high temperatures ($\beta d \le 1$), the bounds of Thms. 3.3.1 and 3.3.3 are only $O(d)$ away from n_{SL} for small s, fortifying our claim that GoF and EoF are not separated from SL in this setting.

Counterpoint to Sparse DCE efforts The above bounds, especially Thm. 3.3.3, show that for small s GoF and EoF are as hard as recovery of $G(Q)$ itself. A possible critique of these bounds when considering DCE is that the DCE schemes demand that the changes are smaller than s, while our formulations only require the changes to have size at least s. To counter this, we point out that the constructions for Thms. 3.3.1, 3.3.2, and 3.3.3 make at most $2s$ changes when computing bounds for any s (in fact, smaller edits lead to stronger bounds). Thus, the above results catergorically contradict the claim that a generic $O(\mathrm{poly}(s)\log p)$ bound that is d independent and much smaller than n_{SL} can hold for DCE methods on $\mathcal{I}_d$. Since α, β, d are only parameters, and are not restricted in any way, this shows that the assumptions made for DCE cannot be reduced to some conditions on only α, β, d, and further topological conditions must be implicit. In particular, these are stronger than typical incoherence conditions such as Dobrushin/high-temperature ($\beta d < 1$;e.g.,[DDK17; GLP18]).

3.3.1 Proof Technique

The above bounds are shown via Le Cam's method with control on the χ^2-divergence of a mixture of alternatives for GoF, and via a Fano-type inequality for the χ^2-divergence, due to Guntuboyina [Gun11] for EoF. These methods allow us to argue the bounds above by explicit construction of distributions that are hard to distinguish. We briefly describe the technique used for GoF below.

Definition *A s-change ensemble in $\mathcal{I}$ is a distribution P and a set of distributions $\mathcal{Q}$, denoted $(P, \mathcal{Q})$, such that $P \in \mathcal{I}, Q \subseteq \mathcal{I}$, and for every $Q \in \mathcal{Q}$, it holds that $|G(P) \triangle G(Q)| \geq s$.*

Each of the testing bounds we show will involve a mixture of n-fold distributions over a class of distributions. For succinctness, we define the following symbol for a set of distibutions $\mathcal{Q}$

$$\langle \mathcal{Q}^{\otimes n} \rangle := \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} Q^{\otimes n}.$$

Le Cam's method (see e.g. [Yu97; IS12]) shows that if $(P, \mathcal{Q})$ is a s-change ensemble in $\mathcal{I}$, then

$$R^{\mathrm{GoF}}(n, s, \mathcal{I}) \geq 1 - \sqrt{\frac{1}{2} \log(1 + \chi^2(\langle \mathcal{Q}^{\otimes n} \rangle \| P^{\otimes n})}.$$

Therefore, if we find a change ensemble and an n such that $1 + \chi^2(\langle \mathcal{Q}^{\otimes n} \rangle \, \| \mathcal{P}^{\otimes n}) \leq 3$, then we would have established that $n_{\mathrm{GoF}}(s, \mathcal{I}) \geq n$. So, our task is set up as constructing appropriate change ensembles for which the χ^2-divergence is controllable.

Directly constructing such ensembles is difficult, essentially due to the combinatorial athletics involved in controlling the divergence. We instead proceed by constructing a pair of separated distributions (P_0, Q_0) on a small number of nodes, and then 'lifting' the resulting bounds to the p nodes via tensorisation - P is contructed by collecting disconnected copies of P_0, while $\mathcal{Q}$ is constructed by changing some of the P_0 copies to Q_0. The process is summarised as follows.

Lemma 3.3.4. *(Lifting) Let P_0 and Q_0 be Ising models with degree $\leq d$ on $\nu \leq p/2$ nodes such that $|G(P_0)\triangle G(Q_0)| = \sigma$, and $\chi^2(Q_0^{\otimes n}\|P_0^{\otimes n}) \leq a_n$. Let $m := \lfloor p/\nu \rfloor$. For $t < m/16e$, there exists a $t\sigma$-change ensemble $(P, \mathcal{Q})$ over p nodes such that $|\mathcal{Q}| = \binom{m}{t}$ and*

$$1 + \chi^2(\langle \mathcal{Q}^{\otimes n}\rangle \| P^{\otimes n}) \leq \exp\left(\frac{t^2}{m} a_n\right).$$

A similar argument is used for the EoF bounds, along with a similar lifting trick, discussed in §B.2. Due to the tensorisation of the χ^2-divergence, we obtain results of the form $a_n \leq (1+\kappa)^n - 1$, where κ depends on (P_0, Q_0) but not n. Plugging this into the above with $t = \lceil s/\sigma \rceil$ yields

$$n_{\mathrm{GoF}}(s, \mathcal{I}) \geq \frac{1}{\log(1+\kappa)} \log\left(1 + \frac{p\sigma^2}{8\nu s^2}\right).$$

Notice that this κ is an SNR term, while $\log(1 + p\sigma^2/8\nu s^2)$ captures combinatorial effects.

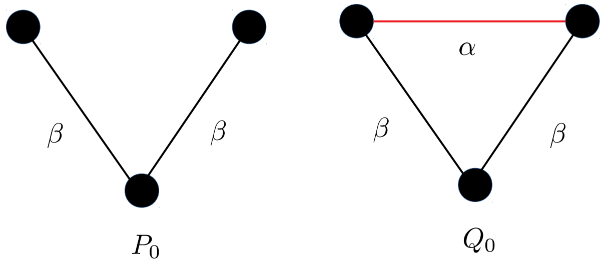


Figure 3·1: Graphs used to construct high-temperature obstructions. Labels indicate edge-weight, and the red edge is added in Q_0.

The procedure thus calls for strong χ^2 bounds for various choices of small graphs, or 'widgets'. We use two varieties of these - the first, 'star-type' widgets, are variations on a star graph. These allow direct calculations in general, and provide bounds that extend to the high-temperature regime. The second variety is the 'clique-type' widgets, that are variations on a clique, and provide low-temperature obstructions. Classical Curie-Weiss analysis shows that cliques tend to 'freeze' - for Ising models on a k-clique with uniform weight λ, the probability mass concentrates on the set $\{(1)^{\otimes k}, (-1)^{\otimes k}\}$ w.p. roughly $1 - e^{-\Theta(\lambda k)}$. The clique-type obstructions implicitly argue that this effect is very robust.

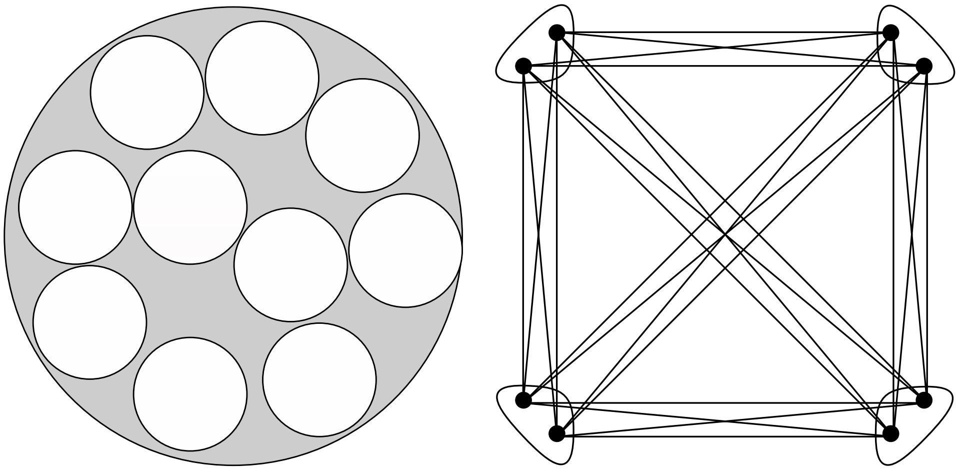


Figure 3·2: Two views of Emmentaler cliques. Left: the base clique is the large grey circle, uncoloured circles represent the groups with no edges within (this is $d, \ell \gg 1, {}^{d+1}/_{\ell+1} = 10$); Right: Emmentaler as the graph $K_{\ell+1,\ell+1,\dots,\ell+1}$ $(d = 7, \ell = 1)$.

The particular graphs used to argue the high temperature bounds in Theorems 3.3.1 and 3.3.3 are a 'V' versus a triangle as seen in Fig. 3·1, while in Theorem 3.3.2 the empty graph is compared to a $d^{1-\zeta}$-clique. The low temperature obstructions of Theorem 3.3.1 and3.3.2 compare a full $d+1$-clique as P_0 to an 'Emmentaler' clique (Fig. 3·2). These are constructed by dividing the $d+1$ nodes into groups of size $\ell + 1$, and removing the $\ell + 1$-subclique within each group. The graph can thus be seen either as a clique with many large 'holes' - corresponding to the deleted subcliques - which inspires the name, or as the complete ${}^{d+1}/_{\ell+1}$-partite graph on $d+1$ nodes. Notice that in the Emmentaler clique we have deleted $\approx {}^{d\ell}/_{2}$ edges. We will show in §B.4 that this is still hard to distinguish from the full clique for $\ell \sim d/10$ - a deletion of $\Omega(d^2)$ edges!

On Tightness Prima facie the above bounds suggest that one may find sample efficient schemes in, say, GoF for $s \gg \sqrt{pd}$. However, it is our opinion that these bounds are actually loose. Particularly, while the SNR terms are relatively tight, the behaviour of $f(p, d, s)$ is not. To justify this opinion, consider the setting of forest-structured graphs. By the same techniques, we show a similar bound with $f = p/s^2$ for GoF in forests in §3.4.1 - this is the best possible by the methods employed. For $s \gg \sqrt{p}$, the resulting overall lower bound is the trivial $n \geq 1$ unless $\alpha \leq (p/s^2)^{1/2}$. On the other hand, [DDK19, Thm. 14] can be adapted to show a lower bound for forests of $\Omega(\alpha^{-2} \wedge \alpha^{-4}/p)$ for the particular case of $s = p/2$, which is non-trivial for all $\alpha \lesssim p^{-1/4}$. Our results trivialise for $\alpha \gtrsim p^{-1/2}$ for this case, demonstrating looseness.

The reason for this gap lies in the lifting trick used to show these bounds. The tensorisation step involved in this constricts the set of 'alternates' one can consider, thus diminishing f. More concretely - there are about ${}^{p^2-pd}\!/_2$ potential ways to add an edge (and $O(pd)$ to delete an edge), while the lifting process as implemented here restricts these to at most $O(pd)$. It is important to recognize this lossiness, particularly since *most* lower bounds, for both testing and recovery, proceed via a similar trick, e.g. [SW12; TSRD14; SC16; GNS17; NL19; CNL18]. [DDK19, Thm. 14] is the only exception we know of. We conjecture that for GoF in $\mathcal{I}_d$, f should behave like p^2/s^2, while for EoF, it should behave like p^2/s. Note that for GoF, since s can be as big as pd, this indicates that one should look for sample-efficient achievability schema in the setting of $s > pd^c$.

However, for simpler settings this technique *can* recover tight bounds. For instance, §3.4.1 presents a matching upper bound for testing of edge-*deletion* in a forest. Notice that in this case there are only $O(p)$ possible ways to edit. This raises the further question of if the same effect extends to $\mathcal{I}_d$, i.e., can deletion of edges in $\mathcal{I}_d$ be tested with $O(1 \vee e^{2\beta d}\alpha^{-2}(pd/s^2))$ samples when $s \gg \sqrt{pd}$? §3.4.2 offers initial results in this direction in the high temperature regime.

3.4 Testing Edge Deletions

Continuing on the theme that concluded our discussion of the tightness of our lower bounds, we study the testing of edge deletions in two classes of Ising models - forests, and high-temperature ferromagnets - with the aim demonstrating natural settings in which the sample complexity of GoF testing of Ising models is provably separated from that of the corresponding recovery problem.

In the deletion setting, we consider the same problems as in §3.2, but with the additional constraint that if $Q \neq P$, then $G(Q) \subset G(P)$, that is, the network structures of alternates can be obtained by dropping some edges in that of the null. For a class of Ising models $\mathcal{J}$, we thus define

$$R^{\mathrm{GoF,del}}(n, s, \mathcal{J}) = \inf_{\Psi} \sup_{P \in \mathcal{J}} P^{\otimes n}(\Psi(P, X^n) = 1) + \sup_{\substack{Q \in A_s(P) \cap \mathcal{J} \\ G(Q) \subset G(P)}} Q^{\otimes n}(\Psi(P, X^n) = 1),$$

and, analogously define $R_{\mathrm{EoF,del}}(n, s, \mathcal{J})$, and the corresponding sample complexities $n_{\mathrm{GoF,del}}(s, \mathcal{J})$ and $n_{\mathrm{EoF,del}}(s, \mathcal{J})$.

We will look at testing deletions for two choices of $\mathcal{J}$ which both have uniform edge weights

- **Forest-Structured Models** ($\mathcal{F}(\alpha)$) are Ising models with uniform weight α such that their network structure is a forest (i.e., has no cycles).
- **High-Temperature Ferromagnets** ($\mathcal{H}_d^\eta(\alpha)$) are models with max degree at most d, uniform *positive* edge weights α, and further such that there is an $\eta < 1$ such that $\alpha d \leq \eta$.

We note that while our motivation for the study of the above is technical, both of these subclasses of models have been utilised in practice, and indeed are the subclasses of $\mathcal{I}_d$ that are best understood.

3.4.1 Testing Deletions in Forests

Forest-structured Ising models are known to be tractable, and have thus long served as the first setting to explore when trying to establish achievability statements. We show a tight characterisation of the sample complexity of testing deletions in forests for large changes, and also demonstrate the separation from the corresponding EoF (and thus also SL) problem. In addition, we also show that for the restricted subclass of trees, essentially the same characterisation follows for *arbitrary* changes (i.e., not just deletions), and that the methods support some amount of tolerance directly. We begin with the main result for testing deletions in forests (all proofs are in §B.3.1).

It is worth noting that degrees are not assumed to be explicitly bounded in this section - i.e. the results hold even if the max degree is $p-1$ (a star graph).

Theorem 3.4.1. *There exists a constant C independent of (s, p, α) such that the sample complexity of* GoF *testing of forest-structured Ising models against deletions is bounded as*

$$n_{\mathrm{GoF,del}}(s, \mathcal{F}(\alpha)) \le C \max\left\{1, \frac{1}{\sinh^2(\alpha)} \frac{p}{s^2}\right\}.$$

Conversely, for $s \le p/32e$, there exists a constant C' independent of (s, p, α), such that

$$n_{\mathrm{GoF,del}}(s, \mathcal{F}(\alpha)) \ge \max\left\{1, \frac{1}{C'} \frac{1}{\sinh^2 \alpha} \log\left(1 + \frac{p}{C' s^2}\right)\right\},$$
$$n_{\mathrm{EoF,del}}(s, \mathcal{F}(\alpha)) \ge \frac{1}{C' \sinh^2 \alpha} \log\left(\frac{p}{C' s}\right).$$

The upper bound is constructed using the global statistic $\mathscr{T}_P = \sum_{(i,j)\in G(P)} X_i X_j$, averaged across the samples. Again, the behaviour of the lower bound shifts as s crosses $\sqrt{p}$ - for larger s, it scales as $1 \vee \sinh^{-2}(\alpha) p/s^2$, while for much smaller s it is $1 \vee \sinh^{-2}(\alpha) \log p$. Further, for large changes, the lower bound is matched, up to constants, by the achievability statement above. For the smaller case, the same holds in the restricted setting of $\alpha < 1$, since exact recovery in $\mathcal{F}(\alpha)$ only needs

$\tanh^{-2}(\alpha)\log p$ samples (Chow-Liu algorithm, as analysed in [BK20]).[2] Finally, the EoF lower bound (which is also tight for $\alpha < 1$, show that the sample complexity of GoF is separated from error of fit (and thus SL) for large changes.

Fig. 3·3 illustrates Thm. 3.4.1 via a simulation for testing deletions in a binary tree (for $p = 127, \alpha = 0.1$), showing excellent agreement. In particular, observe the sharp drop in samples needed at $s = 21 \approx 2\sqrt{p}$ versus at $s < \sqrt{p} \approx 11$. We note that SL-based testing fails for all $s \leq 60$ for this setting even with 1500 samples (Fig. B·1 in §B.3.3), which is far beyond the scale of Fig. 3·3. See §B.3.3 for details.

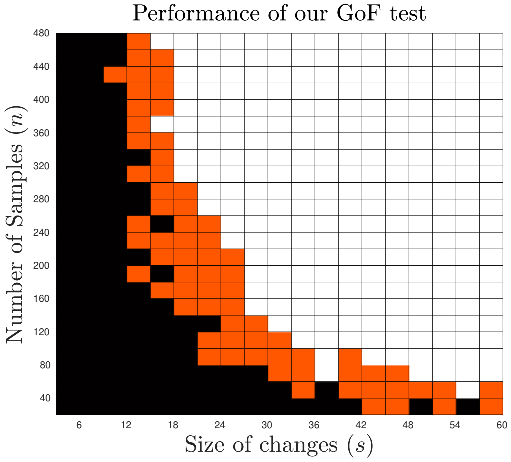


Figure 3·3: Testing deletions in binary trees for $p = 127, \alpha = 0.1$. Entries are coloured black if risk is > 0.35, white if < 0.15, and orange otherwise.

Testing arbitrary changes in trees The statistic $\mathscr{T}$ is good at detecting deletions in edges, but is insensitive to edge additions, which prevents it from being effective in general for forests. However, if the forest-models P and Q are restricted to have the same *number of edges*, then $\mathscr{T}$ should retain power, since any change of s edges must delete $s/2$ edges. This, of course, naturally occurs for trees! Let $\mathcal{T}(\alpha) \subset \mathcal{F}(\alpha)$ denote tree-structured Ising models.

[2] While the $\alpha < 1$ regime is certainly more relevant in practice, it is an open question whether for larger α, and for small s, the correct SNR behaviour is $\sinh^{-2}$ or $\tanh^{-2}$ in testing.

Theorem 3.4.2. *There exists a C independent of (p, s, α) s.t.*

$$n_{\mathrm{GoF}}(s, \mathcal{T}(\alpha)) \le C \max\left(1, \frac{1}{(1-\tanh(\alpha))^2 \sinh^2(\alpha)} \frac{p}{s^2}\right).$$

Conversely, there exists a c independent of (p, s, α) such that

$$n_{\mathrm{GoF}}(s, \mathcal{T}(\alpha)) \ge c \frac{1}{\tanh^2(\alpha)} \log\left(1 + \frac{cp}{s^2}\right).$$

Tolerant Testing The achievability results of Thm.s 3.4.1,3.4.2 can be made 'tolerant' without much effort (see §B.3.1.3). 'Tolerance' here refers to updating the task to separate models that are εs-close to P from those that are s-far from it.

Concretely, let $\mathcal{J}$ be a class of Ising models, s, p, n as before, and let $\varepsilon \in (0, 1)$ be a tolerance parameter. We set up the following risks of tolerant testing of s changes at tolerance ε, and of tolerant testing of deletion at the same levels, as

$$R^{\mathrm{GoF}}_{\mathrm{tol}}(n, s, \varepsilon, \mathcal{J}) = \inf_{\Psi} \sup_{P \in \mathcal{J}} \left\{ \sup_{\widetilde{P} \in A_{\varepsilon s}(P)^c \cap \mathcal{J}} \widetilde{P}^{\otimes n}(\Psi = 1) + \sup_{Q \in A_s(P) \cap \mathcal{J}} Q^{\otimes n}(\Psi = 0) \right\},$$

$$R^{\mathrm{GoF,del}}_{\mathrm{tol}}(n, s, \varepsilon, \mathcal{J}) = \inf_{\Psi} \sup_{P \in \mathcal{J}} \left\{ \sup_{\substack{\widetilde{P} \in A_{\varepsilon s}(P)^c \cap \mathcal{J} \\ G(\widetilde{P}) \subset G(P)}} \widetilde{P}^{\otimes n}(\Psi = 1) + \sup_{\substack{Q \in A_s(P) \cap \mathcal{J} \\ G(Q) \subset G(P)}} Q^{\otimes n}(\Psi = 0) \right\}.$$

Analogously to §3.2, the sample complexities $n^{\mathrm{tol}}_{\mathrm{GoF}}(s, \varepsilon, \mathcal{J})$ and $n^{\mathrm{tol}}_{\mathrm{GoF,del}}(s, \varepsilon, \mathcal{J})$ are the smallest n required to drive the above risks below $1/4$. Our claim in the above may be summarised as follows.

Theorem 3.4.3. *There exists a constant C independent of $(s, p, \alpha, \varepsilon)$ such that*

$$n^{\mathrm{tol}}_{\mathrm{GoF,del}}(s, \varepsilon, \mathcal{F}(\alpha)) \le C \max\left\{1, \frac{1}{\sinh^2(\alpha)} \frac{p}{(1-\varepsilon)^2 s^2}, \frac{1}{(1-\varepsilon)^2 s}\right\}.$$

Further, if $\varepsilon < {}^{1-\tanh(\alpha)}/_{2}$, *then*

$$n^{\mathrm{tol}}_{\mathrm{GoF}}(s,\varepsilon,\mathcal{T}(\alpha)) \le C\max\left\{1, \frac{1}{\sinh^2(\alpha)}\frac{p}{(1-2\varepsilon-\tanh(\alpha))^2 s^2}, \frac{1}{(1-2\varepsilon-\tanh(\alpha))^2 s}\right\}.$$

The key point for showing the above is that the mean of the statistic $\mathscr{T}$ doesn't move too much under small changes - for $\tau = \tanh(\alpha)$, changing εs edges reduces the mean of $\mathscr{T}_P$ by at most $\varepsilon s\tau$ in both cases, while changing $\ge s$ edges reduces it by at least $s\tau$ for forest deletion, and ${}^{s\tau(1-\tau)}/_{2}$ for arbitrary changes in trees. Comparing this upper bound in the drop in the mean against the lower bound when $\ge s$ changes are made (along with the common noise scale of the problem) directly gives the above blowups in the costs of tolerant testing. This should be contrasted with statistical distance based formulations of testing, for which tolerant testing is a subtle question, and, at least in unstructured settings, requires using different divergences to define closeness and farness in order to show gains beyond learning [DKW18].

3.4.2 Testing Deletions in High-Temperature Ferromagnets

Testing deletions in ferromagnets is amenable due to two technical properties of the statistic $\mathscr{T}_P = \sum_{(i,j)\in G(P)} X_i X_j$. The first of these is that due to the ferromagneticity, deleting an edge can only reduce the correlations between the values that the variables take. Coupling this fact with a structural result that is derived using [SW12, Lemma 6] yields that if $G(Q) \subset G(P)$ and $|G(P)\triangle G(Q)| \ge s$, then $\mathbb{E}_P[\mathscr{T}_P] - \mathbb{E}_Q[\mathscr{T}_P] \gtrsim s\alpha$. The second technical property is that bilinear functions of the variables, such as $\mathscr{T}_P$, exhibit concentration in high-temperature Ising models. In particular, using the Hoeffding-type concentration of [AKPS19, Ex. 2.5], $\mathscr{T}_P$ concentrates at the scale $O(\sqrt{pd})$ around its mean for all high-temperature ferromagnets. With means separated, and variances controlled, we can offer the following upper bound on the sample complexity, while the converse is derived using techniques of previous sections. See §B.3.2 for proofs.

Theorem 3.4.4. *There exists a constant C_η depending only on η and not otherwise on (s, p, d, α) such that*

$$n_{\mathrm{GoF,del}}(s\mathcal{H}_d^\eta(\alpha)) \le C_\eta \left(\frac{pd}{\alpha^2 s^2} \vee 1 \right).$$

Conversely, there exists a $c < 1$ independent of (s, p, d, α) such that if $\eta \le 1/16, s \le cpd$ then

$$\begin{aligned} n_{\mathrm{GoF,del}}(s, \mathcal{H}_d^\eta(\alpha)) &\ge \frac{c}{\alpha^2 d^2} \log \left(1 + \frac{cpd^3}{s^2} \right), \\ n_{\mathrm{EoF,del}}(s, \mathcal{H}_d^\eta(\alpha)) &\ge \frac{c}{\alpha^2 d^2} \log \left(1 + \frac{cpd}{s} \right). \end{aligned}$$

Unlike in Thm. 3.4.1, the lower bounds above are not very clean, and so our characterisation of the sample complexity is not tight. Nevertheless, we once again observe a clear separation between sample complexities of GoF and of EoF and a fortiori that of SL. Concretely, our achievability upper bound and the EoF lower bound show that for $s > \sqrt{pd^3}$, the sample complexity of testing deletions is far below that of structure learning in this class. Further, our testing lower bound tightly characterises the sample complexity for $s \ge \sqrt{pd^3}$.

As an aside, note that unlike in the forest setting, it is not clear if $\mathscr{T}$ is generically sensitive to edge deletions, since network effects due to cycles in a graph can bump up correlation even for deleted edges. However, we strongly suspect that a similar effect does hold in this setting, raising another open question - can testing of changes in the subclass of $\mathcal{H}_d^\eta$ with a fixed number of edges be performed with $O(\alpha^{-2}pd/s^2)$ samples for large s? A similar open question arises for tolerant testing, which requires us to show that small changes do not alter the mean of $\mathscr{T}$ too much.

3.5 Discussion

The chapter was concerned with the structural goodness-of-fit testing problem for Ising models. We first argued that this is instrinsically motivated, and we distinguished this formulation from GoF testing under statistical measures that has been pursued in the recent literature. The main problem we studied was that of the sample complexity of GoF testing, with a refined question asking when this was significantly separated from that of structure learning. Alternatively, we can view this question as asking when testing via structure learning is suboptimal in sample costs. In addition, we considered the EoF estimation problem, which serves as a proxy for approximate structure recovery, and also aligns with the focus of the sparse DCE literature. We showed that quite generically, if the number of edge edits being tested are small, then the GoF testing and EoF testing problems are not separated from structure learning in sample complexity. This concretely rebuts the approach taken by the sparse DCE methods, and instead suggests that algorithmic work on structural testing should concentrate on large changes. In addition, we identified inefficiencies in our lower bound technique, namely that the number of changes the constructions allow is too small, which reduces the effectiveness of the lower bounds below the level we believe them to hold (in that the bounds trivialise for too small an s, in our opinion). In order to demonstrate that this is the only source of looseness, we demonstrated upper bounds for GoF testing in the deletion setting. This was helped by the fact that the deletion problem is much simpler than full testing, because the relevant test statistic is pretty obvious for this case, while it is unclear what statistic is appropriate to construct general tests. Along the way we controlled the sample costs of generic testing in tree structured models, and showed that the same tests easily admit some level of tolerance around the null model.

A number of questions are left open, and we point out a few here. From the perspective of lower bounds, the chief is to remove the inefficiencies in our lower bound technique. As a beginning towards this, it may be worth exploring if the methods used to show [DDK19, Thm. 14] can be extended to deal with $s < p$ changes. In addition, we note that while the SNR terms in the lower bounds are relatively tight, there are still extraneous factors that need to be addressed. Coming around to upper bounds, the main open problem is that of constructing tests for degree bounded Ising models in the setting $s = pd^c$ for some $c > 0$. Further, we ask if our bounds on testing deletions in high-temperature ferromagnets can be extended to generic ferromagnets (which would require replacing the concentration argument), or to generic changes in high-temperature ferromagnets (which would require development of new statistics that are sensitive to edge additions et c.). In addition, can the deletion result be extended to testing under the constraint the the null and alternate models have the same number of edges (analogously to how the forest deletion results extend to changes in trees), and can the deletion result be made tolerant?

3.5.1 Open Questions Regarding Testing and Recovery of Community Structure with Node Level Observations

Both of the models discussed in Chapters 2 and 3 capture different critical features of real world network oriented tasks, but leave out others.

The SBM captures the fact that often instead of some fine grained information about each pair of interactions, practitioners are interested in underlying low rank structure, such as communities or collections of fundamental groups. However, the SBM assumes that individual connections are observable, which is a strong assumption - frequently, only node level observations are available, and practitioners generate networks from these via crude correlation based thresholding heuristics (this is especially the case in neuroscientific contexts). The Ising model captures the latter feature of typical data, in that it studies node level observations, but suffers from the fact that the behaviour of the model, and particularly the inference problems commonly studied are concerned with the whole graph.

It thus behoves us to attempt to unite these models to produce a node level graphical model that encodes low rank structure such as communities. Two natural models have been proposed in the recent literature[3] In the rest of this section, I'll discuss these two models, and some open directions of inquiry into the same that may be pursued.

The Ising Block Model

The *Ising Block Model* (IBM), also called the 'Block Spin Ising Model' was proposed by Berthet et al. [BRS19]. The model is mathematically convenient, but unrealistic. Due to its convenience, however, this is a natural proving ground for algorithms. The Ising Block Model is a law on $\{-1,1\}^p$-valued random variables parametrised by three quantities - an partition of the nodes $z \in \{-1,+1\}^p$, and intra- and inter-community edge weights α, β. The law is that Curie-Weiss model, but with edge weights taking a block structure (recall that $\mathbf{1} = (1, \dots, 1)^\top$).

$$P(X = x; z, \alpha, \beta) \propto \exp\left(\frac{x^T J x}{p}\right)$$
$$J = \frac{\beta+\alpha}{2}\mathbf{1}\mathbf{1}^\top + \frac{\beta-\alpha}{2} z z^\top .$$

Analysis techniques and intuitions The principle fact underlying the convenience of this model is that due to the symmetry of the potential, the covariance of the model develops a block structure, i.e., there exist two constants Δ, Ω, depending on p, α, β such that

$$\mathbb{E}[XX^T] = \frac{\Delta+\Omega}{2}\mathbf{1}\mathbf{1}^T + \frac{\Delta-\Omega}{2} z z^T + (1-\Delta)I.$$

[3] For consistency of notation, I'll refer to the intra- and inter- block connection probabilities of a SBM as c_{in}/n and c_{out}/n respectively. Throughout, edge strengths are encoded as factors of β, α with the implicit understanding that $\beta > \alpha$.

In addition to the above, notice that sample covariance is a sufficient statistic for this model - indeed, the law is simply that of an exponential family of the form $\exp\left(\langle XX^T, J\rangle\right)$. Thus, the fundamental analytic characterisation of this model is as a block covariance model, but with the non-trivial twist that the concentration of this matrix is driven by the block Curie-Weiss structure underlying it.

The above observations render natural the approach of reasoning about the structure of the IBM by studying projections of the sample covariance matrix. Indeed, the results of [BRS16] follow from such considerations, with the determination of exact recovery thresholds via semi-definite programming, and some partial recovery results via spectral methods.

Notice, however, the subtlety inherent in the above that Δ and Ω are themselves non-trivial to ascertain as functions of the parameters p, α, β. This imposes probabilistic challenges in determining sample complexities in terms of the natural parameters. Such issues have drawn probabilistic attention, and various refined characterisations extending the analysis of [BRS16] are now available [LS18; KLSS20].

Proposed investigation We begin by pointing out that a natural generalisation of our GoF scheme to consider a sample covariance instead a graph adjacency matrix is easily analysable, and shows a similar advantage for the IBM as it does in the SBM, in that large changes can be tested with very few samples. However, the same cannot be said for our scheme for two sample testing Algorithm 1, because this proceeds critically via a partial recovery step, the sample costs of which are not as well established in the IBM as they are in the SBM. This presents two natural directions of extension. The first is to develop TST schemes for the IBM that do not pass through a partial recovery step. Indeed, this is a broader question that is already of interest in the SBM (see the discussion in the following section). For this reason, perhaps the better place to start when pursuing this thrust is to develop improved TST schemes in the SBM itself.

The second line of investigation is that of the partial recovery problem in the IBM itself. [BRS16] show that exact recovery in the IBM is possible via an SDP based scheme, and show achievability bounds for the partial recovery problem using spectral methods. In the analogous SBM, it is known that while SDP based schemes extend to partial recovery at minimax rates [FC19], the same is not true for naïve applications of spectral methods, which typically require some mild regularisation of the obtained graph in order to succeed (for example, see [CRV15]). This raises two concrete questions - the first is if the SDP based analysis of [FC19] for the partial recovery problem can be extended to the IBM. The second is to determine what the appropriate analogue of graph regularisation is for covariance matrices, and to adapt regularised spectral schemes in the SBM that carry this out to the IBM setting.

The Stochastic Ising Block Model

The *Stochastic Ising Block Model* (SIBM) is perhaps the more natural integration of the SBM and the Ising model. The idea is to draw a single graph G from a two community SBM with connectivities $(c_{\text{in}}/p, c_{\text{out}}/p)$ and then to place an Ising Model on this graph, from which observations are drawn. Formally, the model is parametrised by a community vector $z \in \{\pm 1\}^p$, and a edge weight parameter β, along with the connectivity parameters, and takes the following form:

1. A graph G is drawn according to an SBM with planted community z and connection probabilities c_{in}/p and c_{out}/p.

2. n samples $X^{(1)}, \dots, X^{(n)}$ are drawn independently from the Ising model on G with weight β, i.e. the law

$$P(X = x; G, \beta) \propto \exp\left(\beta \sum_{\{i,j\} \in G} X_i X_j\right).$$

Notice that the samples of the Ising model are drawn from the same graph - this is an important feature, since in real world inference tasks we expect the underlying network of relationships between agents to be constant at the time scale of experimentation. The analogous setting for continuously valued data is the Stochastic Gaussian Block Model (SGBM), in which samples are drawn from a Gaussian with precision matrix $\Theta = \Sigma^{-1} = I - \gamma G$ for a small enough γ. The recent work of Ye [Ye21] studies exact recovery in a variant of the SIBM in which negative biases are introduced in between variables that are not connected according to the underlying graph in order to facilitate recovery.

While our concrete intuition for this question is limited, the main heuristic reasoning behind why testing of the above models was presented in §2.4.3, which carried out experiments on the SGBM as defined above - roughly, this posits that due to the local tree-like structure of large sparse graphs, the empirical covariance matrix approximates the local graph structure well, at least up to some (graph) distance. This suggests a natural two-step line of attack - firstly, to demonstrate that this heuristic actually bears out, and secondly to argue that statistics such as the one designed in §2.4.3 are actually effective given the extent to which the graph structure is expressed in such data.

Part II

Selective Classification for Resource-Efficient Inference

Chapter 4

Selective Classification in the Batch Setting

4.1 Introduction

Selective Classification is a classical problem that goes back to the work of Chow [Cho57; Cho70]. The setup allows a learner to classify a query into a class, or to abstain from doing so (we also call this 'rejecting' the query). This abstention models real-world decisions to gather further data/features, or engage experts, all of which may be costly. Such considerations commonly arise in diverse settings, including healthcare[1], security, web search, and the internet of things ([Xu+14; Zhu+19]), all of which require very low error rates (lower even than the Bayes risk of standard classification). The challenge of SC is to attain such low errors while keeping coverage (i.e., the probability of not rejecting a point) high. This is a difficult problem because any choice of what points to reject is intimately coupled with the classifiers chosen for the remaining points.

The most common SC method is via 'gating,' in which rejection is explicitly modelled by a binary-valued function γ, and classification is handled by a function π. An instance, x, is predicted as $\pi(x)$ if $\gamma(x) = 1$, and otherwise rejected. Within this

This chapter is a lightly edited version of the paper [GKS21], which was written in collaboration with Anil Kag and Venkatesh Saligrama.

[1] For example, when deciding if a mammary mass is benign or malignant, a physician may predict based on ultrasound imaging tests, and, in more subtle cases, abstain and refer the patient to a specialist, or recommend specialised imaging such as CT scans.

formulation, recent work has proposed a number of methods, ranging from alternating minimisation based joint training, to the design of new surrogate losses, and of new model classes to accommodate rejection. Despite this increased complexity, these methods lack power, as shown by the fact that they do not significantly outperform naïve schemes that rely on abstaining on the basis of post-hoc uncertainty estimates for a trained standard classifier. This represents a significant gap in the practical effectiveness of selective classification.

Our Contributions. We describe a new formulation for the SC problem, that comprises of directly learning *disjoint* classification regions $\{\mathcal{S}_k\}_{k\in\mathcal{Y}}$, each of which corresponds to labelling the instance as k respectively. Rejection is *implicitly defined* as the gap, i.e., the set $\mathcal{R} = \mathcal{X} \setminus \bigcup \mathcal{S}_k$. We show that this formulation is equivalent to earlier approaches, thus retaining expressivity.

The principal benefit of our formulation is that it admits a natural relaxation, via dropping the disjointness constraints, into *decoupled* 'one-sided prediction' (OSP) problems. We show that at design error ε, this relaxation has the coverage optimality gap bounded by ε itself, and so the relaxation is statistically efficient in the practically relevant high target accuracy regime.

We pose OSP as a standard constrained learning problem, and due to the decoupling property, they can be approached by standard techniques. We design a method that efficiently adjusts to inter-class heterogeneity by solving a minimax program, controlled by one parameter that limits overall error rates. This yields a powerful SC training method that does not require designing of special losses or model classes, instead allowing use of standard discriminative tools.

To validate these claims, we implement the resulting SC methods on benchmark vision datasets - CIFAR-10, SVHN, and Cats & Dogs. We empirically find that the OSP-based scheme has a consistent advantage over SOTA methods in the regime of low target error. In particular, we show a clear advantage over the naïve scheme described above, which in our opinion is a significant first milestone in the practice of selective classification.

4.1.1 Related Work

State of the Art (SOTA) methods: The SOTA, in terms of performance, for SC is encapsulated by three methods. The Naïve method, i.e., rejecting when the output of a soft classifier is non-informative (e.g. classifier margin is too small), and this is surprisingly effective when implemented for modern model classes such as DNNs ([GE17]). The only other methods that can (marginally) beat this are due to Liu et al., who design a loss function for DNNs [Liu+19], and to Geifman & El-Yaniv, who design a new architecture for DNNs that incorporates gating [GE19].

The methods of [Liu+19; GE19] are both based on the **Gating formulation**, mentioned earlier. This formulation was popularised by Cortes et al. [CDM16], although similar proposals appeared previously [EW10; WE11]. A number of papers have since extended this approach, e.g. designing training algorithms via alternating minimisation [NS17a; NS17b], designing loss functions [Liu+19; NCHS19; RTA18], and model classes, such as an architecturally augmented deep neural network (DNN) [GE19]. In contrast, our work develops an alternate formulation that directly solves SC without use of specialised losses or model classes.

The naïve method has its roots in the **Direct SC** formulation, which is based on learning a function $f : \mathcal{X} \to \{1, \dots, K, ?\}$ (where ? denotes rejection), and is pursued by Wegkamp and coauthors [HW06; BW08; Weg07; WY11; YW10]. The main disadvantage of this formulation is that the methods emerging from it consider very restricted forms of rejection decisions, e.g. $\{|\varphi - {}^1\!/_2| < \delta\}$, where φ is a softmax output of a binary classifier.

Rather than including an explicit gate, our formulation and method for learning abstaining classifiers uses an **implicit abstention criterion**, by modelling regions of high confidence directly. Such an approach was theoretically considered by Kalai et al. [KKM12] for the binary setting, although an implementable methodology was not developed from the same. This paper also suggests a decoupled approach to learning. Independently and concurrently of our work, Charoenphakdee et al. [CCZS21] also propose an implicitly gated method by observing that in the situation where abstention has a fixed cost, the Bayes optimal classifier can be derived using a cost-sensitive objective. They develop this into a methodology for learning selective classifiers that bears significant commonalities to ours in the structure of the losses constructed and approach taken, although their exploration is focused on the situation with a fixed cost for abstention (the so-called Chow loss). Together these papers suggest that the approach we design can be motivated in multiple ways.

An alternate **Confidence Set formulation** (which also features an implicit abstention criterion) has been pursued in the statistics literature [Lei14; DH19] (for the binary case), and involves learning sets $\{\mathcal{C}_k\}_{k\in[1:K]}$ such that $\bigcup \mathcal{C}_k = \mathcal{X}$, and each $\mathcal{C}_k$ covers class k in the sense $\mathbb{P}(\mathcal{C}_k|Y = k)$ is large.[2]Points which lie in two or more of the $\mathcal{C}_k$s are rejected, and otherwise points are labelled according to which $\mathcal{C}_k$ they lie in. While this has subsequently been extended to the multiclass setting [SLW19; DH17; CDH19], these papers study 'least ambiguous set-valued classification', which is a different problem from selective classification and does not express it well (see the appendix of [GKS21]). A limitation of existing work in this framework is their reliance on estimating the regression function $\eta(x) := \mathbb{P}(Y = k|X = x)$ to ensure efficiency. Proposals typically go via using non-parametric estimates of η, which are then filtered. On a practical level, this reliance on estimation reduces statistical efficiency, and on

[2] More accurately, this precise formulation has not appeared for the multiclass setting, and only appears for the binary problem in [Lei14; DH19]. Here we are expressing the natural multiclass extension of this, that turns out to be equivalent to selective classification (§4.2.3). The existing literature instead pursues the multiclass extension to LASV classification, as mentioned above.

a principled level, this violates Vapnik's maxim of avoiding solving a more general problem as an intermediate step to solving a given problem ([Vap00, §1.9]).

While our formulation is most closely related to the confidence set formulation, and is equivalent to a change of variables of this (§4.2.3), it is directly motivated. Furthermore, our framework naturally leads to relaxations to OSP that let us study discriminative methods on high-dimensional datasets and large model classes, which are unexplored in these works.

In passing, we mention the *uncertainty estimation* (UE), and *budget learning* (BL) problems. UE involves estimating model uncertainty at any point [GG16; LPB17], which can plug into both naïve classifiers, and the other methods. As such, UE is a vast generalisation of SC. BL is a restricted form of SC that aims at reaching the accuracy of a complex model using simple functions, and is relevant for efficient inference constraints.

We highlight a recent *decoupling-based* method for BL that involves the first and last authors [AGS20]. The present work can be seen as considerable extension of this paper to full SC. While the broad strategies of decoupling schemes are similar, significant differences arise since much of the structure developed in the prior work does not generalise to SC, and development of new forms is necessary. Additionally, our experiments study large multiclass models going beyond best achievable standard accuracy, while the previous work only studies small binary models getting to standard accuracy achievable by larger models.

4.2 Formulation and Methods

Notation. Probabilities are denoted as $\mathbb{P}$, random variables are capitalised letters, while their realisations are lowercase (X and x). Sets are denoted as calligraphic letters, and classes of sets as formal script ($\mathcal{S} \in \mathscr{S}$). Parameters are denoted as greek letters. For a set $\mathcal{S} \subset \mathcal{X}$, $\mathbb{P}(\mathcal{S})$ is shorthand for $\mathbb{P}(X \in \mathcal{S})$.

We adopt the supervised learning setup - data is distributed according to an unknown joint law $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$, and we observe n i.i.d. points $(X_i, Y_i) \sim \mathbb{P}$. For K classes, we set $\mathcal{Y} = [1:K]$, where K is a constant independent of $|\mathcal{X}|$. We use $\mathscr{S}$ to denote the class of sets from which we learn classifiers.

4.2.1 Formulation of SC

We set up the SC problem (Fig. 4·1(top) illustrates binary case) as that of directly recovering disjoint classification regions, $\{\mathcal{S}_k\}_{k\in[1:K]}$ from a class of sets $\mathscr{S}$, under the constraint that the error rate is smaller than a given level ε, which we call the target error. Each such K-tuple of sets induces two events of interest - the rejection event, and the error event.

$$\begin{aligned}\mathcal{R}_{\{\mathcal{S}_k\}} &:= \left\{X \in \left(\bigcup \mathcal{S}_k\right)^c\right\}\\ \mathcal{E}_{\{\mathcal{S}_k\}} &:= \bigcup \{X \in \mathcal{S}_k, Y \neq k\}.\end{aligned}$$

We will usually suppress the dependence of $\mathcal{R}, \mathcal{E}$ on $\{\mathcal{S}_k\}$. Notice further that $\mathcal{E}$ decomposes naturally into events that depend only on one of the $\mathcal{S}_k$s. We will call these 'one-sided' error events

$$\mathcal{E}^k_{\mathcal{S}_k} = \{X \in \mathcal{S}_k, Y \neq k\}.$$

With the above notation, we pose the problem as a maximisation program. The value of this is said to be the *coverage at target error level* ε, denoted $C(\varepsilon;\mathscr{S})$.

$$\begin{aligned} C(\varepsilon;\mathscr{S}) = \max_{\{\mathcal{S}_k\}_{k\in[1:K]}\in\mathscr{S}} & \sum_{k=1}^{K}\mathbb{P}(\mathcal{S}_k) \\ \text{s.t.} \quad & \mathbb{P}(\mathcal{E}_{\{\mathcal{S}_k\}}) \le \varepsilon, \\ & \mathbb{P}(\bigcup_{k,k'\neq k}\mathcal{S}_k\cap\mathcal{S}_{k'}) = 0, \end{aligned} \tag{SC}$$

where the final constraint is expressing the fact that the $\mathcal{S}_k$s must be pairwise disjoint. Note that if ε equals the Bayes risk of standard classification with $\mathscr{S}$, then (SC) recovers the standard solution and coverage 1.

Example. Consider the case of $K = 2$ where $\mathbb{P}_X$ is uniform on $[0,1]$, $\mathbb{P}(Y = 1|X = x) = x$, and $\mathscr{S}$ consists of single threshold sets $\{x > t\}, \{x \le t\}$ for $t \in [0,1]$. The Bayes risk of standard classification is $1/4$. For any $\varepsilon < 1/4$, the coverage at level ε is $C(\varepsilon;\mathscr{S}) = 2\sqrt{\varepsilon}$, which is attained by $\mathcal{S}_1 = \{x > 1-\sqrt{\varepsilon}\}, \mathcal{S}_2 = \{x \le \sqrt{\varepsilon}\}$.

4.2.1.1 Design choices

We outline alternate ways to set up the SC problem that we don't pursue in this dissertation.

Form of constraints. In (SC), we maximise coverage, while controlling error, which is *error-constrained SC*. Alternately one can pursue the equivalent *coverage constrained SC* problem - minimising $\mathbb{P}(\mathcal{E})$ subject to $\mathbb{P}(\mathcal{R}) \le \varrho$.

As illustrated in the starting example, our interest in SC is driven by the desire to attain very small error rates. We thus find the error constrained form of SC more natural, and since we needed to select one of the two for the sake of brevity, we adopt it in the rest of the paper.[3] We note that our method is also effective for

[3] This is not to imply that the coverage constrained form cannot be more appropriate for some settings. Which one to use in practice is ultimately a problem specific choice.

coverage-constrained SC, as shown empirically in §4.4.

Error criterion. In (SC), we constrain the raw error $\mathbb{P}(\mathcal{E})$. This has the benefit of being both natural, since it directly controls the standard error metric, and further, simple. Alternate forms of the error metric have been studied in the literature - e.g. conditioning on acceptance ($P(\mathcal{E}|\mathcal{R}^c)$) [GE19]; and separately constrained class conditionals ($P(\mathcal{E}|Y = k) \le \varepsilon_k$) [Lei14]. Most of the development below can be adapted to these settings with minimal changes, and we restrict attention to $\mathbb{P}(\mathcal{E})$ for concreteness.

4.2.2 Relaxation and One-sided Prediction

(SC) couples the $\mathcal{S}_k$s via the $\mathbb{P}$-a.s. disjointness constraint. We now develop a decoupling relaxation.

To begin, note that we may decouple the error constraint by introducing variables that trades off the one-sided error rates as below. This program is equivalent to (SC) in the sense that they have the same optimal value, and the same $\{\mathcal{S}_k\}$ achieve this value.

$$\begin{aligned} \max_{\{\mathcal{S}_k\}\in\mathscr{S},\{\alpha_k\}\in[0,1]} & \sum_{k=1}^{K} \mathbb{P}(\mathcal{S}_k) \\ \text{s.t.} \quad & \forall k : \mathbb{P}(\mathcal{E}^k_{\mathcal{S}_k}) \le \alpha_k \varepsilon, \quad \sum \alpha_k \le 1, \\ & \mathbb{P}\big(\bigcup_{k,k'\neq k} \mathcal{S}_k \cap \mathcal{S}_{k'}\big) = 0. \end{aligned} \tag{SC-expanded}$$

Our proposed relaxation is to simply drop the final constraint. The resulting program may be decoupled, via a search over the variables α_k into K *one-sided prediction* (OSP) problems:

$$L_k(\varepsilon_k; \mathscr{S}) = \max_{\mathcal{S}_k\in\mathscr{S}} \mathbb{P}(\mathcal{S}_k) \text{ s.t. } \mathbb{P}(\mathcal{E}^k_{\mathcal{S}_k}) \le \varepsilon_k \tag{OSP-k}$$

Notice that the above OSPk problem demands finding the *largest* set $\mathcal{S}_k$ that has a low false alarm probability for the null hypothesis $Y \neq k$. Structurally this is the opposite to the more common Anomaly Detection problem, which demands finding the smallest set with a low missed detection probability.

We note that while we decouple the SC problem completely above, the main benefit is the removal of the intersection constraint, which is the principal difficulty in SC. The sum error constraint is benign, and for reasons of efficiency we will reintroduce it in §4.3.

Continuing, observe that the sets recovered from the above problems may overlap, which introduces an ambiguous region. This overlap region is necessarily of small mass (Prop. 4.2.1), and so may be dealt with in any convenient way. Theoretically we break ambiguities in the favour of the smallest label. These sets need not belong to $\mathscr{S}$ anymore, and so this is an (weakly) improper classification scheme.

Overall this gives the following infinite sample scheme:

- For each feasible $\alpha \in [0,1]^K$, solve for $\{L_k(\alpha_k \varepsilon)\}$ for each $k \in [1:K]$. Let $\{\mathcal{T}_k^\alpha\}$ be the recovered sets.
- Let $\mathcal{S}_k^\alpha = \mathcal{T}_k^\alpha \setminus \left(\bigcup_{k'<k} \mathcal{T}_{k'}^\alpha\right)$.
- Return the $\{\mathcal{S}_k^\alpha\}$ that maximises $\sum_k \mathbb{P}(\mathcal{S}_k^\alpha)$ over α.

At small target error levels, which is our intended regime of study, the resulting sets are guaranteed to not be too lossy, as in the following statement. The above is shown (in §C.1.1) by arguing that the mass of the overlap between the OSP solutions (the $\mathcal{T}_k$) is at most 2ε. Empirically this is even lower, see Table 4.4.

Proposition 4.2.1. *If $\{\mathcal{S}_k\}$ are the sets recovered by the procedure above, then these are feasible for* (SC)*. Further, their optimality gap is at most 2ε, i.e.*

$$\sum_{k\in[1:K]} \mathbb{P}(\mathcal{S}_k) \geq C(\varepsilon; \mathscr{S}) - 2\varepsilon.$$

4.2.3 Equivalence of SC formulations

We show that the prior gating and confidence frameworks are equivalent to ours, based on transforming feasible solutions of one framework into an other.

Gating: Denote the acceptance set of gating as $\Gamma = \{\gamma = 1\}$, and let the predictions be $\Pi_k = \{\pi = k\}$. Taking $\mathcal{S}_k = \Pi_k \cap \Gamma$ yields disjoint sets that can serve for SC under our formulation that have the same decision regions for each class, and the same rejection region, since $(\bigcup \mathcal{S}_k)^c = \Gamma^c$. Conversely, for disjoint decision sets $\mathcal{S}_k$, the gate $\Gamma = \bigcup \mathcal{S}_k$, and the predictor $\Pi_k = \mathcal{S}_k$ form the corresponding gating solution.

Confidence set: Take confidence sets $\{\mathcal{C}_k\}$ which cover $\mathcal{X}$, and have the rejection set $\mathcal{B} = \bigcup_{k \neq k'} \mathcal{C}_k \cap \mathcal{C}_{k'}$. Then we produce the disjoint sets $\mathcal{S}_k = \mathcal{C}_k \setminus (\bigcup_{k' \neq k} \mathcal{C}_{k'})$, which retain the same decision regions. These also have the same rejection region because we may express $\mathcal{S}_k = \mathcal{C}_k \cap \mathcal{B}^c$, and thus $\bigcap \mathcal{S}_k^c = (\bigcap_k \mathcal{C}_k^c) \cup \mathcal{B}$, and $\bigcap \mathcal{C}_k^c = \varnothing$ since the $\mathcal{C}_k$ cover the space. Conversely, for disjoint $\{\mathcal{S}_k\}$, the sets $\mathcal{C}_k = (\bigcup_{k' \neq k} \mathcal{S}_{k'})^c = \mathcal{S}_k \cup \mathcal{R}$ cover the space, and have the rejection region $\mathcal{R}$ since $\mathcal{C}_k \cap \mathcal{C}_{k'} = \mathcal{R}$ for any pair $k \neq k'$.

Figure 4·1 illustrates these equivalences. Notice that due to the simplicity of the reductions, these equivalences are fine-grained in that the joint complexity of the family of sets used is preserved in going from one to the other. Given these equivalences, we again distinguish our approach from the existing ones.

First, the structure of solutions is markedly different. The gating formulation takes both the rejection and the classification decisions explicitly via the two different sets. The confidence set formulation takes neither explicitly, and instead produces a 'list decoding' type solution. In contrast, we make the classification decisions explicit, and produce the rejection decision implicitly.

Consequently, the salient differences lies in the method. The gating based methods have concentrated on the design of surrogate losses and models, while for the confidence set, methods either go through estimating the regression function, or via a reduction

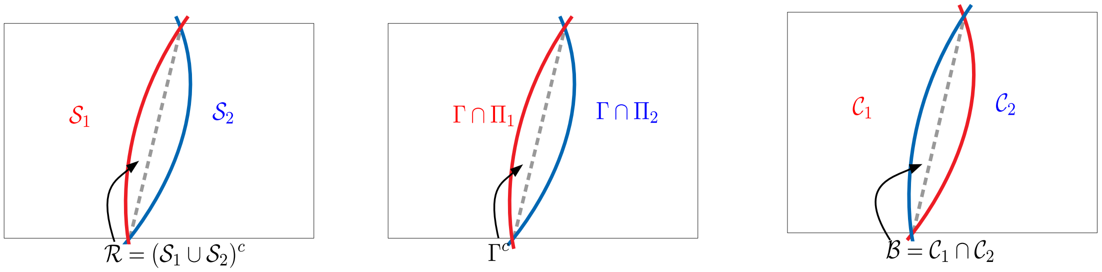


Figure 4·1: An illustration of the equivalence between the three formulations for binary classification. *Left:* our formulation; $\mathcal{S}_i$ denotes disjoint sets; *Middle:* gating with Γ representing gated set; *Right:* $\mathcal{C}_i$ represents confidence sets, and their intersection representing the rejected set. In each case, the coloured curves represent the boundary of the set labelled with the corresponding colour, and the dashed line is the Bayes boundary.

to anomaly detection type problems [SLW19; DH17]In strong contrast, we develop a new relaxation that allows decoupled learning via 'one-sided prediction' problems. These OSP problems are almost opposite to anomaly detection - instead of finding small sets for each class that do not leave too much of its mass missing, we instead learn large sets that do not admit too much of the complementary class' mass.

4.2.4 Finite Sample Properties of OSP

Thus far we have spoken of the full information setting. This section gives basic generalisation analyses for an empirical risk minimisation (ERM) based finite sample approach. Since the one-sided problems are entirely symmetric, we concentrate only on OSP-1, that is (OSP-k) with $k = 1$, below. Note that the SC problem can directly be analysed in a similar way, as discussed in the following section.

We show asymptotic feasibility of solutions, that is, we show that we can, with high probability, recover a set $\mathcal{S}$ for OSP such that $\mathbb{P}(\mathcal{S}) \geq L_1(\varepsilon) - o(1)$ and $\mathbb{P}(\mathcal{E}^1_{\mathcal{S}}) \leq \varepsilon + o(1)$, where the o are as the sample size diverges. This is in contrast to exact feasibility, i.e., insisting on $\mathcal{S}'$ such that $\mathbb{P}(\mathcal{E}^1_{\mathcal{S}'}) \leq \varepsilon$ with high probability. Exactly satisfying constraints via ERM whilst maintaining that the objective is also approaching the optimum is a subtle problem, and has been shown to be impossible in certain cases [RT11]. On the other hand, plug-in methods along with an 'identifiability' condition which imposes that the law of $\eta(X)$ is not varying too fast at any point can be employed to give exact constraint satisfaction along with a small excess risk - the technique was developed by Tong [Ton13], and has been used previously in SC contexts [e.g., SGJ19]. However, since the applicability of plug-in methods to large datasets in high dimensions is limited, we do not pursue this avenue here.

One-Sided Learnability

Definition *We say that a class $\mathscr{S}$ is one-sided learnable if for every $\varepsilon \geq 0$ and $(\delta, \sigma, \nu) \in (0,1)^3$, there exists a finite $m(\delta, \sigma, \nu)$ and an algorithm $\mathfrak{A} : (\mathcal{X} \times [1:K])^m \to \mathscr{S}$ such that for any law $\mathbb{P}$, given m i.i.d. samples from $\mathbb{P}$, $\mathfrak{A}$ produces a set $\mathcal{S}_1 \in \mathscr{S}$ such that with probability at least $1-\delta$ over the data,*

$$\mathbb{P}(\mathcal{S}_1) \geq L_1(\varepsilon; \mathscr{S}) - \sigma, \qquad \textit{and} \qquad \mathbb{P}(\mathcal{E}^1_{\mathcal{S}_1}) \leq \varepsilon + \nu.$$

The characterisation we offer is

Proposition 4.2.2. *A class $\mathscr{S}$ is one-sided learnable iff it has finite VC dimension. In particular, given n samples, we can obtain a set $\mathcal{S}_1$ that, with probability at least $1-\delta$, satisfies*

$$\begin{aligned}\mathbb{P}(\mathcal{S}_1) &\geq L_1(\varepsilon; \mathscr{S}) - \sqrt{C_K \frac{(\mathrm{VC}(\mathscr{S}) \log n + \log(C_K/\delta))}{n}}\\ \mathbb{P}(\mathcal{E}^1_{\mathcal{S}_1}) &\leq \qquad \varepsilon \qquad + \sqrt{C_K \frac{(\mathrm{VC}(\mathscr{S}) \log n + \log(C_K/\delta))}{n}},\end{aligned}$$

where C_K is a constant that depends only on the number of classes K. Equivalently, the label complexity of OSP is bounded as

$$m(\delta, \sigma, \nu) = \widetilde{O}(\mathrm{poly}(K) \max(\nu^{-2}, \sigma^{-2}) \mathrm{VC}(\mathscr{S}))$$

The proof of the necessity of finite VC dimension is via a reduction to standard learning, while the upper bounds on rates above follow from uniform convergence due finite VC dimension. See §C.1.2. The scheme attaining these is a direct ERM that replaces all $\mathbb{P}$s in (OSP-k) by empirical distributions.

On the whole, applying the above result for each of the K OSP problems tells us that if we can solve the empirical OSP problems for the indicator losses and constraints, then we can recover a SC scheme that, with high probability, incurs error of at most $\varepsilon + O(1/\sqrt{n})$ and has coverage of at least $C(\varepsilon; \mathscr{S}) - 2\varepsilon - O(1/\sqrt{n})$.

4.2.5 Finite Sample Analysis for SC

In parallel to the OSP problems, one can directly give finite sample analyses for the SC problem. We begin by defining the solution concept here.

Definition *We say that a class $\mathscr{S}$ is learnable with abstention if for every $(\delta, \zeta, \sigma, \nu) \in (0,1)^4$, there exists a finite $m(\delta, \sigma, \nu, \zeta)$ and an algorithm $\mathfrak{A} : (\mathcal{X} \times \{+,-\})^m \to \mathscr{S}^K$ such that for any law $\mathbb{P}$, and $\varepsilon > 0$, given n i.i.d. samples drawn from $\mathbb{P}$, the algorithm produces sets $\{\mathcal{S}_k\}$ from $\mathscr{S}$ such that with probability at least $1-\delta$ over the data,*

$$\sum_k \mathbb{P}(\mathcal{S}_k) \geq C(\varepsilon; \mathscr{S}) - \sigma$$

$$\mathbb{P}(\mathcal{E}_{\{\mathcal{S}_k\}}) \leq \varepsilon + \nu$$

$$\mathbb{P}(\bigcup_{k,k' \neq k} \mathcal{S}_k \cap \mathcal{S}_{k'}) \leq \zeta.$$

Notice that the recovered sets need not be disjoint, which may be amended by by eliminating the overlap from one of the sets as in §4.2.2. The resulting (improper) sets attain coverage of at least $C - \zeta - \nu$ with high probability.

The main point characterisation here is similar,

Proposition 4.2.3. *A class $\mathscr{S}$ is learnable with abstention if and only if it has finite VC dimension, and further,*

$$m(\delta, \sigma, \nu, \zeta) = \widetilde{O}(\mathrm{poly}(K) \max(\sigma^{-2}, \nu^{-2}, \zeta^{-1}) \mathrm{VC}(\mathscr{S})).$$

The proof of the necessity of finite VC dimension follows from observing that if the data is realisable, i.e., corresponds to $Y = 2\mathbb{1}\{X \in \mathcal{S}\} - 1$ for some $\mathcal{S} \in \mathscr{S}$ then at least one of the recovered sets is a good classifier, at which point standard lower bounds for realisable PAC learning apply. The sufficiency follows from utilising the finite VC property to uniformly bound errors incurred by empirical means. The proof is presented in §C.1.2.

Comment on Semi-Supervised Settings One simple but important observation in this context is that the objective of the SC problem does not depend on the label distribution, and so

4.3 Method

In this section, we derive an efficient scheme, first by replacing indicator losses with two differentiable surrogate variants, and then propose OSP relaxations. A summary of the method expressed as pseudo-code is included in Appx. C.2. Throughout, $\mathscr{S}$ is set to be level sets of the soft output of a deep neural network (DNN), i.e., $\mathscr{S} = \{f(\cdot;\theta) > t\}$, where $f(\cdot;\theta) : \mathcal{X} \to [0,1]$ is a DNN parametrised by θ. The bulk of the exposition concerns learning θs. In this and the following section, $\{(x_i, y_i)\}_{i=1}^n$ refers to a training dataset with n labelled data points.

Relaxed losses. To solve the OSP problem, we follow the standard approach of replacing indicator losses by differentiable ones. This sets up the relaxed problem

$$\min_{\theta_k} \frac{\sum_i \ell(f(x_i;\theta_k))}{n} \text{ s.t.} \frac{\sum_{i:y_i \neq k} \ell'(f(x_i;\theta_k))}{n_{\neq k}} \leq \varphi_k$$

where θ_k parametrises the DNN, φ_k denote relaxed values of the constraints, and ℓ, ℓ' are surrogate losses that are small for large values of their argument, and $n_{\neq k} = |\{i : y_i \neq k\}|$. In the experiments we use $\ell(z) = -\log(z)$ and $\ell'(z) = -\log(1-z)$, essentially giving a weighted cross entropy loss. We refer to the objective of the above problem as $\widetilde{L}_k(\theta_k)$, and the constraint as $\widetilde{C}_k(\theta_k)$.

A more stable loss. Practically, the loss $\widetilde{L}_k$ suffers from instability due to the fact that the first term sums over all instances. This can seen clearly when $\ell = -\log$, for which the objective includes the sum $\sum_{i:y_i \neq k} -\log(f(x_i;\theta_k))$. Since for negative examples we expect $f(x;\theta_k)$ to be small, this sum is very sensitive to perturbations in these values, which reduces the quality of the solutions. To ameliorate this, we formulate the following 'restricted' loss, where the objective instead sums over only the positively labelled samples

$$\min_{\theta_k} \frac{\sum_{i:y_i=k} \ell(f(x_i;\theta_k))}{n_k} \text{ s.t. } \widetilde{C}_k(\theta_k) \le \varphi_k. \tag{4.1}$$

Notice that the constraint $\widetilde{C}_k$ is the same as before. We refer to the restricted objective above as $\widetilde{L}_k^{\text{res.}}(\theta_k)$. This loss underlies all further methods, and §4.4.

Note that the above program remains sound w.r.t. the OSP task, since it is a surrogate for the following

$$\max_{\mathcal{S}_k \in \mathscr{S}} \mathbb{P}(X \in \mathcal{S}_k, Y = k) \text{ s.t. } \mathbb{P}(X \in \mathcal{S}_k, Y \neq k) \le \varepsilon.$$

Comparing (OSP-k) and the above, the constraints are the same, and the objectives differ by $\mathbb{P}(\mathcal{S}_k) - \mathbb{P}(\mathcal{S}_k, Y = k) = \mathbb{P}(\mathcal{S}_k, Y \neq k)$, which, due to the constraint, is at most ε. Thus, the programs are equivalent up to a small gap (that is, optimal solutions for the above attain a value for (OSP-k) that is ε-close to the optimal value for it). For the same reason, we can use the solutions of the above one-sided problem in the scheme of §4.2.2 to yield solutions feasible for (SC) that satisfy an analogue of Prop. 4.2.1 with an optimality gap of 3ε instead of 2ε.

Joint Optimisation and normalisation. A naïve approach with the above relaxations in hand is to optimise the k OSP problems separately. However, this leads to an exponential in K rise in complexity in the model selection process, since different values of $(\varphi_1, \dots, \varphi_K)$ need to be selected - if Φ such values are searched over for each φ_k, then this amounts to a prohibitive grid search over Φ^K values. In addition, due to class-wise heterogeneity, the values of φ_ks need not be calibrated across programs, and thus simple solutions like pinning all the φ_ks to the same value are not viable. A final issue is that a naïve implementation of this setup results in training K separate DNNs, which leads to a K-fold increase in model complexity.

We make two modifications to handle this situation. First, we normalise function outputs by adopting the following architecture: we consider DNNs with K output nodes, each representing one of the f_k. The backbone layers of the network are shared across all OSP problems. Further, we take

$$f(x) = (f_1(x), \dots, f_K(x)) = \mathrm{softmax}(\langle w_k, \xi_\theta(x)\rangle),$$

where ξ_θ denotes the backbone's output, and recall that

$$(\mathrm{softmax}(v))_k = {}^{\exp(v_k)}\!/\!_{\sum \exp(v_k)}.$$

This normalisation and restricted model handles both the class-wise heterogeneity, and the blowup in model complexity.

For the sake of succinctness, we define $\mathbf{w} = (w_1, w_2, \dots, w_K)$, $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_K)$. Next, in order to ameliorate the search, we propose jointly optimising the various OSP problems, by enforcing a joint constraint on the sum of the various constraint values via a single value φ. This mimics the structure of (SC), where the constraint limits the sum of the one-sided errors. The relaxation thus amounts to dropping the disjointness constraint, and softening the indicators in (SC). The resulting problem is

$$\begin{aligned} &\min_{\theta, \mathbf{w}, \boldsymbol{\varphi}} \sum_k \widetilde{L}_k^{\text{res.}}(\theta, \mathbf{w}) \qquad (4.2) \\ &\text{s.t.} \quad \forall k : \widetilde{C}_k(\theta, \mathbf{w}) \le \varphi_k, \quad \sum \varphi_k \le \varphi, \end{aligned}$$

where recall $\widetilde{L}^{\text{res.}}, \widetilde{C}_k$ from above, which are functions of $(\theta, \mathbf{w})$ since the backbone θ is shared, and since all f_k depend on all w_ks due to the softmax normalisation.

Finally, we propose optimising (4.2) via stochastic gradient ascent-descent. We note that one tunable parameter - μ - remains in the problem, corresponding to the sum constraint on the φ_ks, while λ_ks are multipliers for the $\widetilde{C}_k$ constraints. We again denote $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$. The resulting Lagrangian is

$$\begin{aligned} &\widetilde{M}^{\text{res.}}(\theta, \mathbf{w}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \mu) \\ =&\sum_k \widetilde{L}_k^{\text{res.}}(\theta, \mathbf{w}) + \lambda_k(\widetilde{C}_k(\theta, \mathbf{w}) - \varphi_k) + \mu\varphi_k, \end{aligned} \tag{4.3}$$

and we solve the problem

$$\min_{(\theta,\mathbf{w},\boldsymbol{\varphi})} \max_{\boldsymbol{\lambda}:\forall k, \lambda_k \geq 0} \widetilde{M}^{\text{res.}}(\theta, \mathbf{w}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \mu), \tag{4.4}$$

treating μ as the single tunable parameter.

We note that the Lagrangian above bears strong resemblance to a one-versus-all (OVA) multiclass classification objective. The principal difference arises from the fact that the losses are weighted by the λ_k terms, and the optimisation trades these off, which are typically not seen in one-versus-all approaches (of course, we also use the resulting functions very differently).

Thresholding and resulting SC solution. The outputs of the classifiers learned with any given μ yield soft signals for the various OSP problems. To harden these into a decision, we threshold the outputs of the soft classifier at a common level $t \in [0, 1]$. This crucially relies on the earlier normalisation of the soft scores to make them comparable. Finally, to deal with ambiguous regions, we use the soft signals f_k, and assign the label to the one with the largest score. Overall, this leads to the SC solution

$$\mathcal{S}_k(\theta, \mathbf{w}, t) =\{x : f_k(x;\theta,\mathbf{w}) \geq t\} \cap \{x : k = \arg\max_{k'} f_{k'}(x;\theta,\mathbf{w})\}. \tag{4.5}$$

Model Selection. The above setup has two scalar hyperparameters - μ from (4.4), and threshold t at which hard decisions are produced in (4.5), and each choice of these yields a different solution. Our final model is one that performs the best on the validation dataset among all hyperparameter tuples (μ, t). Concretely, let $\widehat{\mathbb{P}}_V$ denote the empirical law on a validation dataset. Denote the solutions from (4.4) with a choice of μ as $(\theta(\mu), \mathbf{w}(\mu))$. Let $\mathbf{M}, \mathbf{T}$ respectively be discrete sets of μ's and t's. The procedure is

- For each $(\mu, t) \in \mathbf{M} \times \mathbf{T}$, and each k, compute $\mathcal{S}_k(\mu, t) = \mathcal{S}_k(\theta(\mu), \mathbf{w}(\mu), t)$ as defined in (4.5).
- For each $(\mu, t) \in \mathbf{M} \times \mathbf{T}$, evaluate $\widehat{C}_V(\mu, t) = \sum_k \widehat{\mathbb{P}}_V(\mathcal{S}_k(\mu, t))$ and $\widehat{E}_V(\mu, t) = \sum_k \widehat{\mathbb{P}}_V(\mathcal{E}^k_{\mathcal{S}_k})$.
- Let $(\mu^*, t^*) = \arg\max_{\mathbf{M}\times\mathbf{T}} \widehat{C}_V(\mu, t)$ subject to $\widehat{E}_V(\mu, t) \leq \varepsilon$.
- Return $(\theta(\mu^*), \{w_k(\mu^*)\}, t^*)$.

4.4 Experiments

4.4.1 Experimental Setup and Baselines

Datasets and Model Class We evaluate all methods on three benchmark vision tasks: CIFAR-10 [Kri09], SVHN-10 [Net+11] (10 classes), and Cats & Dogs[4] (binary). All models implemented below are DNNs with the RESNET-32 architecture [HZRS16], which is a standard model class in vision tasks. 20% of the training data is reserved for validation in each dataset. All models are implemented in the tensorflow framework. The samples sizes and the best standard classification performance is presented in

[4] https://www.kaggle.com/c/dogs-vs-cats

Table 4.1.

Dataset	Num. of Samples			Std. Error
	Train.	Test	Val.	
CIFAR-10	45K	10K	5K	9.58%
SVHN-10	65.9K	26K	7.3K	3.86%
Cats & Dogs	18K	5K	2K	5.72%

Table 4.1: Dataset sizes and standard classification error

Baselines: We benchmark against three state of the art methods. The 'selective net' and 'deep gamblers' methods also require hyperparameter and threshold tuning as in our setup, and we do this in a brute force way on validation data, as in ours.

Softmax Response Thresholding (SR) involves training a neural network for standard classification, and then thresholding its soft output to decide to reject. More formally, the decision is to reject if $\{\text{softmax}(f_1, \dots, f_K) < t\}$, where f is the soft output, and t is tuned on validation data. This simple scheme is known to have near-SOTA performance [GE17; GE19].

Selective Net (SN) is a DNN meta-architecture for SC [GE19]. The network provides three soft outputs - (f, γ, π), where f is an auxiliary classifier used to aid featurisation during training, and γ, π is a gate-predictor pair. Selective net prescribes a loss function that trades off coverage and error via a multiplier c, and by fine-tuning a threshold on γ to reject. We use the publicly available code[5] to implement this, and a comprehensive sweep over the coverage and threshold hyper-parameters. We use 40 valued grid for the parameter c (with 10 equally spaced values in the range $[0.0, 0.65)$ and remaining 30 values in the range $[0.65, 1.0]$). For the gating threshold γ, we use 100 thresholds equally spaced in the range $[0, 1]$, the same as for our scheme.

Deep Gamblers (DG) is a method based on a novel loss function for SC within the gating framework [Liu+19]. The NNs have $K + 1$ outputs - $f_1, \dots, f_K, f_?$. The

[5] https://github.com/geifmany/selectivenet

cross-entropy loss is modified to $\sum \log\left((f_{y_i}(x_i) + \mathscr{o}^{-1} f_{?}(x_i)\right)$, where $\mathscr{o} \in [1, K)$ is a hyperparameter that trades-off coverage and accuracy. Hard decisions are obtained by tuning the threshold of $f_{?}$ on a validation set. We adapt the public torch code[6] for this method to the Tensorflow framework. We used 40 values of $\mathscr{o}$ spaced equally in the range $[1, 2)$[7], and 100 values of thresholds in $[0, 1]$.

4.4.2 Training One-Sided Classifiers

Loss Function We use the loss function $\widetilde{M}^{\text{res.}}$ developed in §4.3. In particular for $\widetilde{L}_k^{\text{res.}}$, we use $\ell(z) = -\log(z)$, and for $\widetilde{C}_k, \ell'(z) = -\log(1 - z)$.

Training of Backbones As previously discussed, our models share a common backbone and have a separate output node for each OSC problem. We intialise this backbone with a base network trained using the cross-entropy loss (i.e. a 'warm start'). Note that this typically yields a strong featurisation for the data, and exploiting this structure requires us to not move too far away from the same. At the same time, due to the changed objective, it is necessary to at least adapt the final layer significantly. We attain this via a two-timescale procedure: the loss is set to the OSP Lagrangian, and the backbone is trained at a *slower rate* than the last layer. Concretely, the last layer is updated at every epoch, while the backbone is updated every 20 epochs. This stabilises the backbone, while still adapting it to the particular OSP problem that the network is now trying to solve.

Hyper-parameters. All of the methods were trained using the train split and the model selection was performed on the validation set. The results are reported on the separate test data (which is standard for all three of the models considered). The

[6] https://github.com/Z-T-WANG/NIPS2019DeepGamblers/

[7] We initially made a mistake and scanned $\mathscr{o}$ in $[1, 2)$ instead of $[1, 10)$. We then redid the experiment. with 40 values in $[1, 10)$, and found that performance deteriorated. This is because the optimal $\mathscr{o}$ for these datasets lies in $[1, 2)$, and the wider grid leads to a less refined search in this domain. Thus, values from the original experiment are reported. See Tables C.2, C.3 in §C.3 for the values with a scan over $[1, 10)$.

minimax program on the Lagrangian was optimised using a two-timescale stochastic gradient descent-ascent, following the recent literature on nonconvex-concave minimax problems [LJJ19]. In particular, we used Adam optimizer for training with initial learning rates of $(10^{-3}, 10^{-5})$ for the min and the max problems respectively for CIFAR-10 and SVHN-10, and of $(10^{-3}, 10^{-4})$ for Cats & Dogs.[8] These initial rates were reduced by a factor of 10 after 50 epochs, and training was run for 200 epochs. The batch size was set to 128.

We searched over 30 values of μ for each of our experiments - 10 values equally spaced in $[0.01, 1]$, and remaining 20 equally spaced in $[1, 16]$. We further used 100 values of thresholds equally spaced in $[0, 1]$.

4.4.3 Results

The key takeaway of our empirical results is the significant increase in performance of our SC scheme when compared to the baselines. We also include some observations about the structure of the solutions obtained.

4.4.3.1 Performance

Tables cited below all appear after the exposition.

Performance at Low Target Error is presented in Table 4.2, which reports coverage at three (small) targeted values of error - $1/2, 1$, and 2 percent - that are in line with the low target error regime that is the main focus of our design. Notice that these target error values are far below the best error obtained for standard classification (Table 4.1). We observe that the performance of our SC methods is significantly higher

[8] These rates were selected as follows: the standard classifier was trained with the rate 10^{-4}, which is a typical value in vision tasks. We then picked one value of μ, and trained models using rates in $(10^{-k}, 10^{-j})$ for $(j, k) \in [2:6] \times [2:6]$, tuned thresholds for models at 0.5% target accuracy using validation data, and chose the pair that yielded the best validation coverage. Performance tended to be similar as long as $j \neq k$, and curiously, we found it slightly better to use a smaller rate for the max problem, which goes against the suggestions of Lin et al. [LJJ19].

than the SOTA methods, especially in the case of CIFAR-10 and Cats & Dogs, where we gain over 4% in coverage at the 0.5% design error. The effect is weaker in SVHN, which we suspect is due to saturation of performance in this simpler dataset.

Performance at High Target Coverage is presented in Table 4.3. This refers to the coverage constrained SC formulation discussed in §4.2.1.1. For these experiments, we use the same μ values (to avoid retraining), but choose thresholds such that the coverage of the resulting model exceeds the stated target, and the models with the lowest error at this threshold are chosen. We observe that at target coverage 100%, the SR solution outperfoms all others. This is expected, since 100% coverage corresponds to standard classification, and the SR objective is tuned to this, while the others are not. Surprisingly, for coverage below 100%, our OSP-based relaxations deliver stronger performance than the benchmarks. Note that this is not due to the low target error performance, because (besides SVHN), the errors attained at these coverage are signficantly above the low target errors investigated in Table 4.2. This shows that our formulation is also effective in the high-coverage regime.

Coverage-Error Curves for the CIFAR-10 dataset are shown in Fig. 4·2. These curves plot the best coverage obtained by training at a given target error level using each of the methods discussed.[9] We find that the coverage obtained by our method uniformly outperform DG and SN, and also outperform SR for the bulk of target errors, except those very close to the best standard error attainable. This illustrates that our scheme is effective across target error levels. We find this rather surprising since we designed our method with explicit focus on the low target error regime. Tables 4.2 and 4.3 can be seen as detailed looks at the left (error < 2) and the upper (coverage > 90) ends of these curves.

Observations regarding baselines. Tables 4.2,4.3, and Figure 4·2 all show that

[9] In particular, we train models at target errors $\varepsilon_i = (i/2)\%$ for $i \in [1:20]$. We then obtain the achieved test error rates $\widehat{\varepsilon}_i$ and coverages c_i for these models. The curves linearly interpolate between $(\widehat{\varepsilon}_i, c_i)$ and $(\widehat{\varepsilon}_{i+1}, c_{i+1})$.

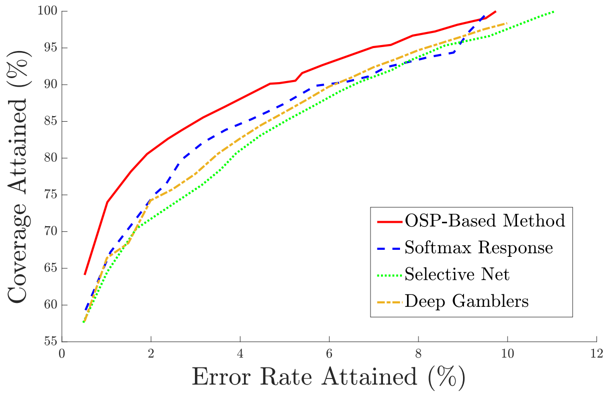


Figure 4·2: Coverage vs Error Curves for the CIFAR-10 dataset. Higher values of coverage are better. Notice the curious behaviour of SR in that the curve's slope sharply changes close to the best standard error rate.

across regimes, DG and SN perform similarly to SR, and are frequently beaten by it. This observation is essentially consistent with the results presented in previous work [GE19; Liu+19], and supports our earlier claims that the prior SOTA methods for selective classification do not meaningfully improve on naïve methods. To alleviate concerns about implementation, we emphasise that we performed a comprehensive hyperparameter search for both SN and DG, and the only change is to use RESNETs instead of VGG.

4.4.3.2 Structure of the Solutions

Overlap of OSP solutions is small. Table 4.4 shows the probability mass of the ambiguous regions for our raw OSP solutions (i.e., the raw sets $\{x : f_k > t\}$ without the max-assignment $\mathcal{S}_k = \{x : f_k > t\} \cap \{x : k = \arg\max f_k\}$) for the models of Table 4.2. We find that this overlap is very small - much smaller than the 2ε bound in Prop. 4.2.1. Empirically, these sets are essentially disjoint, and so the training process is close to tight for the SC problem. We believe that this effect is mainly due to the simple tuning enabled by the softmax normalisation of OSP problem outputs described in §4.3.

Consistency of rejection regions We say that a sequence of models trained at error levels ε_i have consistent rejection regions if for every $\varepsilon_i < \varepsilon_j$, if $\mathcal{R}_i, \mathcal{R}_j$ are the rejection regions for models trained at these errors, then $\mathbb{P}(\mathcal{R}_i \cap \mathcal{R}_j^c)$ is very small. This means that points that are rejected when designing at a higher error level continue to be rejected for stricter error control. Such consistency may be useful for building cascades of models, or for using the error level at which a point is rejected as a measure of uncertainty.

We found that the models obtained by our procedure are remarkably consistent in the high-accuracy regime. Concretely, for $\varepsilon_i = (i/2)\%$ for $\in [1:5]$, for both CIFAR-10 and Cats & Dogs test sets, the models were entirely consistent, i.e. $\mathcal{R}_j \subset \mathcal{R}_i$ for $j > i$[10], while for SVHN, the only violation was that $|\mathcal{R}_{2.5\%} \cap \mathcal{R}^c_{2\%}| = 2$. Since the test dataset for SVHN has size > 7000, this is a tiny empirical probability of inconsistency of $< 0.03\%$.

Dataset	Target	OSP-based		SR		SN		DG	
	Error	Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error
CIFAR-10	2%	**80.6**	1.91	75.1	2.09	73.0	2.31	74.2	1.98
	1%	**74.0**	1.02	67.2	1.09	64.5	1.02	66.4	1.01
	0.5%	**64.1**	0.51	59.3	0.53	57.6	0.48	57.8	0.51
SVHN-10	2%	**95.8**	1.99	95.7	2.06	93.5	2.03	94.8	1.99
	1%	**90.1**	1.03	88.4	0.99	86.5	1.04	89.5	1.01
	0.5%	**82.4**	0.51	77.3	0.51	79.2	0.51	81.6	0.49
Cats & Dogs	2%	**90.5**	1.98	88.2	2.03	84.3	1.94	87.4	1.94
	1%	**85.4**	0.98	80.2	0.97	78.0	0.98	81.7	0.98
	0.5%	**78.7**	0.49	73.2	0.49	70.5	0.46	74.5	0.48

Table 4.2: Performance at Low Target Error. The OSP-based scheme is our proposal. SR, SN, DG correspond to softmax-response, selective net, deep gamblers. Errors are rounded to two decimals, and coverage to one.

[10] Due to time constraints we only checked for higher values of i in the CIFAR-10 case, in which the trend continued until $i = 20$, that is, until full coverage. A curious observation in this case was that in all 20 models for the CIFAR-10 dataset, the same value of μ was best, and the models differed only in the thresholds (this did not occur for SVHN and Cats v/s Dogs). While this obviously implies consistency of the rejection regions, it is unexpected, and suggests that there may be room to improve in our training methodology.

Dataset	Target	OSP-based		SR		SN		DG	
	Cov.	Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error
CIFAR-10	100%	100	9.74	99.99	**9.58**	100	11.07	100	10.81
	95%	95.1	**6.98**	95.2	8.74	94.7	8.34	95.1	8.21
	90%	90.0	**4.67**	90.5	6.52	89.6	6.45	90.1	6.14
SVHN-10	100%	100	4.27	99.97	**3.86**	100	4.27	100	4.03
	95%	95.1	**1.83**	95.1	1.86	95.1	2.53	95.0	2.05
	90%	90.1	**1.01**	90.0	1.04	90.1	1.31	90.0	1.06
Cats & Dogs	100%	100	5.93	100	**5.72**	100	7.36	100	6.16
	95%	95.1	**2.97**	95.0	3.46	95.2	5.1	95.1	4.28
	90%	90.0	**1.74**	90.0	2.28	90.2	3.3	90.0	2.50

Table 4.3: Performance at High Target Coverage. Same notation as Table 4.2.

Dataset	Target Error	Overlap
CIFAR-10	2%	0.09%
	1%	0.01%
	0.5%	0.00%
SVHN-10	2%	0.05%
	1%	0.01%
	0.5%	0.00%
Cats & Dogs	2%	0.07%
	1%	0.01%
	0.5%	0.00%

Table 4.4: Size of overlap between OSP sets in Table 4.2

4.5 Discussion

We have proposed a new formulation for selective classification, which leads to a novel one-sided prediction method. The formulation is naturally motivated, and is equivalent to other formulations appearing in the literature. It is also amenable to standard statistical analyses. The resulting method is flexible, efficiently trainable via standard techniques, and further outperforms state of the art methods across target error regimes. Further, it is the first method to non-trivially outperform naïve post-hoc solutions, and thus, in our opinion, represents a significant step in the practical approaches to selective classification.

Chapter 5

Budget Learning - Selective Classification In the Low Error Limit

5.1 Introduction

This chapter specialises the study of Chapter 4 to the concrete domain where the task is to match the predictions of a given high complexity model that is available as a black box, specifically in the very low error regime. Since the OSP relaxation of 4 is tight when $\varepsilon = 0$, this allows us to study the selective classfication problem in this limit exactly, and offers new characterisations of the problem by developing a connective with brackets, a popular tool in empirical process theory.

In the following, we will explicitly adopt the edge-cloud metaphor discussed in Chapter 1. Rather than expressing the SC problem as learning sets, we will describe it in terms of learning functions (although this is entirely equivalent), and work in the regime of directly approximating the cloud model, which implicitly gives us a noiseless, but complex, feedback, much as the setting of the right half of Figure 1·2, reproduced below. To keep the chapter self-contained, both introductory exposition and basic learnability results are given, although these are essentially a repetition of the discussion of the previous chapter. The main extensions beyond it

This chapter is an edited version of the paper [AGS20], that was written with Alp Acar and Venkatesh Saligrama. Thanks are also due to Pengkai Zhu for discussion regarding the implementation of the empirical experiments.

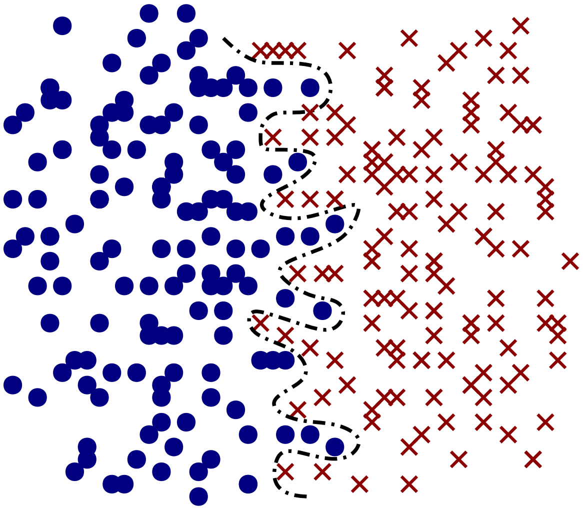


Figure 5·1: The setting of Budget Learning - data labelled by a complex classifier g is available to the learner. This data is ostensibly separable, but only by a complex boundary, while the learner must implement a sufficiently simple model, as encoded by a given hypothesis class $\mathcal{H}$.

are the approximation theoretic definitions of budget learnability and corresponding investigation into budget learnability of simple classes of model. Another important deviation is that rather than coverage, the exposition is in terms of *usage*, which is simply one minus the coverage.

5.1.1 Context for Budget Learning

Edge devices in mobile and IoT applications are battery and processing power limited. This imposes severe constraints on the methods implementable in such settings - for instance, the typical CPU-based structure of such devices precludes the use of many convolutional layers in vision tasks due to computational latency [ZWTD19], imposing architectural constraints. In particular, modern high accuracy methods like deep neural networks are seldom implementable in these settings. At the same time, edge devices are required to give fast and accurate decisions. Enabling such mechanisms is an important technical challenge.

Typically, practitioners either learn weak models that can be implemented on the edge (e.g. [Wu+19; KGV17; HVD15]), which suffer more errors, or they learn a complex model, which is implemented in a cloud[1]. The latter solution is also not ideal - cloud access must be purchased, the prediction pipeline suffers from communication latency, and, since communication consumes the majority of the battery power of such a device [Zhu+19], such solutions limit the device's operational lifetime (see also industry articles, e.g. [Nor19; Hol17]). A third option, largely unexplored in practice, is a hybrid of these strategies - we may learn mechanisms to filter out 'easy' instances, which may be classified at the edge, and send 'difficult' instances to the cloud. The reduction in cloud usage provides direct benefits in, e.g., battery life, yet accuracy may be retained. Similar concerns apply in many contexts, e.g. in medicine, security, and web-search [Xu+14; NS17a].

The key challenge in these applications is to maintain a high accuracy while keeping the usage of the complex classifier, i.e. the budget, low. To keep accuracy high, we enforce that on the locally predicted instances, the prediction *nearly always agrees* with the cloud. This is thus a problem of 'bottom-up' budget learning (BL).

The natural approach to BL is via the 'gating formulation': one learns a gating function γ, and a local predictor π, such that if $\gamma = 1$ then π is queried, and otherwise the cloud is queried. Unfortunately, this setup is computationally difficult, since the overall classifier involves the product $\pi \cdot \gamma$, and optimising over the induced non-covexity is hard. Previous efforts try to meet this head on, but either yield inefficient methods, or require difficult to justify relaxations.

[1] or, more realistically, purchase access to a cloud-based model owned by a company that has sufficient data and computational power (e.g. [ML 19; Cor19]).

Our Contributions

Our main contribution is a novel formulation of the BL problem, via the notion of brackets, that sidesteps this issue. For functions $h^- \le h^+$, the bracket $[h^-, h^+]$ $:= \{f : h^- \le f \le h^+\}$. Brackets provide accurate pointwise control on a binary function - for $f \in [h^-, h^+]$, if $h^+(x) = h^-(x)$, then $f(x)$ takes the same value. We propose to learn a bracketing of the cloud, predicting locally when this condition holds.

The key advantage of this method arises from the surprising property that we may learn optimal brackets via two *decoupled* learning problems - separately approximating the function from above and from below. These one-sided problems are tractable under convex surrogates, with minimal statistical compromises. Further, this comes at negligible loss of expressivity compared to gating - the existence of good gates and predictors implies the existence of equally good brackets.

Since expressivity is retained, bracketings lead naturally to definitions of learnability that are theoretically analysable. We define a PAC-style approach to one-sided prediction, and provide a VC-theoretic characterisation of the same. We also identify the key budget learning problem as an approximation theoretic question - *which complex classes have 'good' bracketings by simple classes?* We characterise this for a binary version of Hölder smooth classes, and also provide partial results for generic classes with bounded VC dimension.

Finally, to validate the formulation, we implement the bracketing framework on a binary versions of MNIST and CIFAR classification tasks. With a strong disparity in the cloud and edge models (§5.5), we obtain usages of $20 - 40\%$ at accuracies higher than 98% with respect to the cloud. Further, we outperform existing methods in usage by factors of $1.2 - 1.4$ at these high accuracies.

Related Work

A common approach is to simply learn *local classifiers with no cloud usage.* If the cloud model is available, one can use methods such as distillation [HVD15], and in general one can train classifiers in a resource aware way (e.g. [KGV17; Gup+17; Wu+19]). The main limitation of this approach is that if the setting is complex enough for a cloud to be needed, then in general such methods cannot attain a similar accuracy level.

Top-Down and Sequential Approaches are based on successively learning classifiers of increasing complexity, incorporating the previously learned classifiers (see [Xu+14; TS13; WTS15; NWS16; BWDS17]). This approach suffers a combinatorial explosion in the complexity of the learning problems. Recent efforts utilise reinforcement learning methods to rectify this (e.g. [JPL19b; JPL19a; PTLC18]).

The BL problem is an instance of selective classification, with two specialasions - we assume that a noiseless ground truth, i.e., the 'cloud classifier' exists, and the class of locally implementable models is much weaker than the class known to contain the cloud model, while the LwA literature is generally not concerned with 'simple' classifiers. We refer to the discussion of selective classification in the previous chapter for greater context.

Plug-in methods utilise a pre-trained low complexity model, and learn a gate by estimating its low-confidence regions. This includes the entropy-thresholding method discussed in the previous chapter, that serves as a very strong baseline [GE17; GE19].

A number of methods aim at *jointly learning gating and prediction* functions. Some of these belong to the SC literature - [GE19] proposes to ignore the non-convexity, and use SGD to optimise a loss of the form $\widehat{\mu}(\pi \neq g|\gamma = 1)$ subject to a budget constraint, while [CDM16] instead proposes the relaxation $\pi\gamma \leq (\pi + \gamma)/2$, and optimise this upper bound via convex relaxations. In the BL literature, [NS17a; NS17b] propose to relax the problem by introducing an auxiliary variable to decouple π and γ, and then perform alternating minimisation with a KL penalty between the gate and the auxiliary. Note that while each of these papers further specifies algorithms to train classifiers, their main conceptual contribution is the method they take to ameliorate the essential non-convexity of the gating setup. In contrast, our new formulation sidesteps this issue entirely.

Our approach to one-sided prediction is related to *Neyman-Pearson classification* [CHHS02; SN05], with the difference that instead of studying the conditional risks, we are concerned with restricting the total risk subject to one-sided constraints. This leads to the generalisation errors of one-sided prediction scaling with the total sample size, as opposed to the per-class sample sizes (see §5.3.1).

Bracketings are important in empirical process theory - for instance, 'bracketable' classes characterise the universal Glivenko-Cantelli property [vHan13]. While there are generic estimates of the bracketing entropies of various function classes (e.g. Ch2 of [vdVW96]), these typically do not constrain for complexity of the resulting brackets, and thus their application in our setting is limited. Instead, we explicitly aim to bracket functions by *simple* function classes (see §5.4.2). We note, however, that our results towards this are preliminary.

5.2 Definitions and Formulations

We will restrict discussion to binary functions on the domain $\mathcal{X}$, which is assumed to be compact[2]. $\mathcal{H}$ denotes the class of local classifiers, and $\mathcal{G}$ the class of cloud classifiers. We use $g \in \mathcal{G}$ to denote the high-complexity 'cloud' classifier. The training set is taken to be $\{(X_i, g(X_i)\}$, where the X_i are assumed to have been sampled

[2] Issues of measurability, and of existence of minimisers of optimisation problems posed as infima are suppressed, as is common in learning theory.

independently and identically from an unknown probability measure μ on $\mathcal{X}$.[3] For feasibility of various programs (particularly Def. 2), we assume that $\{0,1\} \subset \mathcal{H}$, and that $h \in \mathcal{H} \iff 1 - h \in \mathcal{H}$.

The main problem is to learn approximations to g in $\mathcal{H}$, with the option to 'fall back' to g. We aim at retaining high accuracy w.r.t. g while minimising usage of g itself.

5.2.1 Bracketing for Budget Learning

Definition *Given a measure μ and functions, $h_1 \le h_2$, the bracket $[h_1, h_2]$ is the set of all $\{0,1\}$-valued functions f such that $h_1 \le f \le h_2$ μ-a.s. The μ-size of such a bracket is $|[h_1, h_2]|_\mu := \mu(h_1 \neq h_2)$.*

As an example, on $[0,1]$, the functions $0(x)$ and $\mathbb{1}x > {}^1\!/\!_2$ induce the bracket containing all functions that are 0 on $[0, {}^1\!/\!_2]$. This bracket has size $\mu(X > {}^1\!/\!_2)$.

Notice that if $h_1 \neq h_2$ in the above, it is forced that $h_1 = 0, h_2 = 1$. We will be concerned with the brackets that can be built using hs from the local class $\mathcal{H}$.

Definition *The set of brackets generated by a class $\mathcal{H}$ is $\{[h_1, h_2] : h_1 \le h_2, h_1, h_2 \in \mathcal{H}\}$. We also say that these are $\mathcal{H}$-brackets.*

Suppose we can find a bracket $[h^-, h^+]$ in $\mathcal{H}$ that contains g. Since $h^- = h^+$ forces g to take the same value, we offer the classifier

$$c_{[h^-,h^+]}(x) = \begin{cases} h^+(x) & \text{if } h^+(x) = h^-(x) \\ g(x) & \text{if } h^+(x) \neq h^-(x) \end{cases}.$$

The above has the *usage* $|[h^-, h^+]|_\mu$. The budget needed by a class $\mathcal{H}$ to bracket (g, μ) is the smallest such usage,

$$\mathsf{B}(g, \mu, \mathcal{H}) := \inf_{\mathcal{H}-\text{brackets}} \{|[h^-, h^+]|_\mu : g \in [h^-, h^+]\}.$$

[3] If instead we have a raw dataset and no g, we assume that g is obtained by training a function in $\mathcal{G}$ over this set.

This extends naturally to bracketing of sets.

Definition 1 *A set of function-measure pairs $\mathcal{S} = \{(g_i, \mu_i)\}$ is bracket-approximable by a class $\mathcal{H}$ if for every (g, μ), there exists a $\mathcal{H}$-bracket containing g. The budget required for bracket approximation of $\mathcal{S}$ by $\mathcal{H}$ is*

$$\mathsf{B}(\mathcal{S}, \mathcal{H}) := \sup_{(g,\mu)\in\mathcal{S}} \mathsf{B}(g, \mu, \mathcal{H}).$$

This is a very weak notion of approximation - all it demands is that for every g, we can find some $\mathcal{H}$-bracket. Typical study of bracketings concentrates on real valued functions, and studies how many brackets, or how large an $\mathcal{H}$, we need to make the loss B smaller than some given value. We defer such explorations to §5.3.2, where we define a notion of budget learning.

For the following discussion, it is useful to define a relaxed version of brackets.

Definition *Let $\alpha \in [0, 1]$, and h_1, h_2 be $\{0, 1\}$-valued functions such that $\mu(h_1 \le h_2) \ge 1 - {}^{\alpha}\!/_{2}$. The α-approximate bracket $[h_1, h_2]$ with respect to μ is the set of functions f such that $\mu(h_1 \le f \le h_2) \ge 1 - \alpha$. We call $1 - \alpha$ the accuracy of the bracketing.*

The above brackets are approximate in two ways: the order of h_1 and h_2 may be reversed, and the functions in the $[h_1, h_2]$ may leak out from within them.

5.2.2 One-sided Approximation and Decoupled Optimisation of Brackets

In order to discuss the decoupled optimisation of brackets, we introduce the notion of *one-sided approximation.*

Definition 2 *For a function-measure pair (g, μ), an* approximation from below *to g in a class $\mathcal{H}$ is any minimiser of the following optimisation problem*

$$\mathsf{L}(g, \mu, \mathcal{H}) := \inf\{\mu(h \ne g) : h \in \mathcal{H}, h \le g\}.$$

We refer to L as the inefficiency of approximation from below of (g, μ) by $\mathcal{H}$. We analogously define approximation from above *as $1 - h$, where h is an approximation of $1 - g$ from below.*

We use 'one-sided approximation' to refer to both approximation from above and below.

If we let h^- be an approximation of a function g from below, and h^+ an approximation from above, then it follows that $h^- \leq g \leq h^+$. Thus, the bracket $[h^-, h^+]$ is well-defined. Further, for any bracket containing g,

$$\begin{aligned}\mu(h^+ \neq h^-) &= \mu(h^+ \neq h^-, g = 1) + \mu(h^+ \neq h^-, g = 0)\\ &= \mu(h^- = 0, g = 1) + \mu(h^+ = 1, g = 0)\\ &= \mu(h^- \neq g) + \mu(h^+ \neq g).\end{aligned}$$

Thus, if h^+ and h^- are respectively the minimisers of the right hand side, they must also be minimisers of the left hand side. Immediately, we have

$$\mathsf{B}(g, \mu, \mathcal{H}) = \mathsf{L}(g, \mu, \mathcal{H}) + \mathsf{L}(1 - g, \mu, \mathcal{H}),$$

and the respective minimisers of the Ls form a μ-optimal $\mathcal{H}$-bracketing of g!

This means that in order to bracket g optimally, it suffices to *separately* learn approximations to g from above and below. This decouples the optimisation problems inherent in learning these, and allows easy convex relaxations of both the above problems.

Note that the reverse direction trivially holds - the optimal bracket containing g provides two functions which upper and lower approximate g. These functions are optimal for the respective OSP problems.

Comment The above one-sided prediction structure is of course, the same as that of Chapter 4. In the same vein, the methodology described therein applies unchanged to budget learning, and we will omit explicity description of the same.

5.3 Learnability

As mentioned in the previous paragraph, we define notions of one-sided and budget learnability.

5.3.1 One-sided Learnability

The following parallels §4.2.4, and essentially presents the same results, adapted to the functional presentation in this chapter.

With only finite data, it is impossible to certify that $h \le f$ for most h, rendering the one-sided constraint tricky. We take the PAC approach, and relax this condition by introducing a 'leakage parameter' λ.

Definition 3 *A class $\mathcal{H}$ is* one-sided learnable *if for all $(\varepsilon, \delta, \lambda) \in (0,1)^3$, there exists a $m(\varepsilon, \delta, \lambda, \mathcal{H}) < \infty$, and a scheme $\mathscr{A} : (\mathcal{X} \times \{0,1\})^m \to \mathcal{H}$ such that for any function-measure pair (g, μ), given m samples of $(X_i, g(X_i))$, with $X_i \overset{i.i.d.}{\sim} \mu$, $\mathscr{A}$ produces a function $h \in \mathcal{H}$ such that with probability at least $1-\delta$:*

$$\mu(g(X) = 0, h(X) = 1) \le \lambda$$
$$\mu(g(X) = 1, h(X) = 0) \le \mathsf{L}(g, \mathcal{H}, \mu) + \varepsilon.$$

The above definition closely follows that of PAC learning in the agnostic setting, with the deviations that leakage is explicitly controlled, and that the excess risk control, ε, is on L, i.e. it is only with respect to *entirely* non-leaking functions. A key shared feature is that one-sided learnability is a property only of the class $\mathcal{H}$, and is agnostic to (g, μ).

If the class $\mathcal{H}$ is learnable, then with $m(\varepsilon, \lambda, \delta, \mathcal{H})$ samples we may learn a approximate bracketing of any g with usage at most $\mathsf{B}(g, \mu, \mathcal{H}) + 2(\varepsilon + \lambda)$ and accuracy at least $1 - 2\lambda$

Let us distinguish the above from the Neyman-Pearson classification setting of [CHHS02; SN05]. The latter can be seen as learning from below, but with explicit control on the *conditional probability* $\mu(h = 1|g = 0)$.[4] This is too strong for our needs - we are only interested in emulating the behaviour of g *with respect to* μ, and so if $\mu(g = 0) < \lambda$, then it is fine for us to learn any h. This induces the difference that the error rates in the cited papers decay with $\min(n_0, n_1)$, while our setting is simpler and PAC guarantees follow the entire sample size. Nevertheless, our claims on the sample complexity(§5.4.1) are derived similarly to the setting of 'NP-ERM' in these papers, including a testing and an optimisation phase.

5.3.2 Budget Learnability

The bracket-approximation of Def. 1 suffers from two problems in the ML context. Firstly, approximation by classes that are *not* one-sided learnable is irrelevant. Secondly, the definition does not control for effectiveness: a bracket-approximation with $\mathsf{B}(\mathcal{S}, \mathcal{H}) = 1$, is not useful - indeed, the trivial class $\mathcal{H} = \{0(x), 1(x)\}$ attains this for every $\mathcal{S}$. We propose the following to remedy these.

Definition 4 *We say that a set of function-measure pairs* $\mathcal{S} = \{(g, \mu), \dots\}$ *is budget-learnable by a class* $\mathcal{H}$ *if* $\mathcal{H}$ *is one-sided learnable and* $\mathsf{B}(\mathcal{S}, \mathcal{H}) < 1$.
We also, say that $\mathcal{H}$ *can budget learn* $\mathcal{S}$, *adding "with budget* B*" if* $\mathsf{B}(\mathcal{S}, \mathcal{H}) \leq \mathsf{B}$.

Learning theoretic settings usually require measure independent guarantees, leading to

Definition 5 *A function class* $\mathcal{G}$ *on the measurable space* $(\mathcal{X}, \mathscr{F})$ *is said to be budget learnable by a class* $\mathcal{H}$ *if the set* $\mathcal{S} := \mathcal{G} \times \mathcal{M}$ *is budget learnable by* $\mathcal{H}$, *where* $\mathcal{M}$ *is the set of all probability measures on* $(\mathcal{X}, \mathscr{F})$.

Notice that strict inequality is required in Def. 4. This is the weakest notion that is relevant in an ML context. Also note the trivial but useful regularity property that if $\mathcal{H}$ is one-sided learnable, then $\mathsf{B}(\mathcal{H} \times \mathcal{M}, \mathcal{H}) = 0$ - indeed, every $h \in \mathcal{H}$ is bracketed by $[h, h]$.

[4] In addition, the targeted control on this is some level $\alpha > 0$, not 0, and a relaxation of the form we use to $\alpha + \lambda$ is also utilised. Further, the property of only comparing against the best classifier at the target level of leakage (α in their case, 0 in ours) is also shared.

5.4 Theoretical Properties

This section details some useful consequences of the above definitions, which serve to highlight their utility.

5.4.1 One-sided learnability

Standard PAC-learning is intrinsically linked to the VC-dimension. The same holds for one-sided learnability.

Theorem 5.4.1. *If $\mathcal{H}$ has finite VC-dimension d, then it is one-sided learnable with*

$$m(\varepsilon, \lambda, \delta, \mathcal{H}) = \widetilde{O}\left(\left(\frac{1}{\lambda} + \frac{1}{\varepsilon^2}\right)(d + \log(1/\delta))\right).$$

Conversely, if $\mathcal{H}$ is one-sided learnable and has VC-dimension $d > 1$, then for $\delta < 1/100$,

$$m(\varepsilon, \lambda, \delta, \mathcal{H}) > \frac{d-1}{32(\lambda + \varepsilon)}.$$

Particularly, one-sided learnable classes must have finite VC-dimension.

The proof is left to §D.1.1. The lower bound is proved via a reduction to realisable PAC learning, while the upper bound's proof is similar to that for agnostic PAC learning, with the modification of adding a test that eliminates functions that leak too much.

The point of the Theorem 5.4.1 is to illustrate that sample complexity analyses for our formulation can be derived via standard approaches in learning theory. Alternate analyses via, e.g., Rademacher complexty or covering numbers are also straightforward (§D.1.1.1).

5.4.2 Budget Learnability

The key question of budget learning is one of bias: what classes of functions can be budget learned by low complexity classes? This section offers some partial results towards an answer.

Before we begin, the (big) question of how one measures complexity itself remains. We take a simple approach - since one-sided learnability itself requires finite VC-dimension, we call $\mathcal{H}$ low complexity if $\textsc{vc}(\mathcal{H})$ is small. Certainly VC dimension is a crude notion of complexity. Nevertheless this study leads to interesting bounds, and outlines how one may give theoretical analyses for more realistic settings that may be pursued in further work.

Importantly, we do not expect any one class to be able to meaningfully budget learn *all* classes of a given complexity. This follows since the definition of budget learnability implies that if sets $\mathcal{S}_1, \mathcal{S}_2$ of function-measure pairs are budget learnable, then so is $\mathcal{S}_1 \cup \mathcal{S}_2$. Such unions can lead to arbitrary increase in complexity, which must weaken the budget attained.[5] Thus, at the very least, the classes $\mathcal{H}$ must depend on $\mathcal{G}$, although we would like them to not depend on the measure.

5.4.2.1 Budget Learnability of Regular Classes

The class of Hölder smooth functions is a classical regularity assumption in non-parametric statistics. In this section, we define a natural analogue for $\{0,1\}$-valued functions, and discuss its budget learnability by low VC dimension classes. For simplicity, we restrict the input domain to the compact set $\mathcal{X} = [0,1]^p$. We use Vol to denote the Lebesgue measure on $\mathcal{X}$.

Definition *Let g be a $\{0,1\}$-valued function. A partition $\mathscr{P}$ of $\mathcal{X}$ is said to be aligned with g if each set $\Pi \in \mathscr{P}$ has connected interior, and if g is a constant on each such set.*

We define a notion of regularity for partitions below. Recall that a p-dimensional rectangle is a p-fold product of 1-D intervals.

[5] Formally, this finite union property and the lower bound Thm. 5.4.3 part (i) indicate that if $\mathcal{H}$ can budget learn *all* classes of VC dimension D on all measures with budget $1-c$ for any $c > 0$ that depends only on D or $\mathcal{X}$, but not on $\mathcal{H}$, then $\forall k \in \mathbb{N}, \textsc{vc}(\mathcal{H}) \geq CkD$ for a constant C.

Definition *A partition $\mathscr{P}$ is said to be V-regular if every part $\Pi \in \mathscr{P}$ contains a rectangle R_Π such that* $\mathrm{Vol}(R_\Pi) \ge V$ *and* $\mathrm{Vol}(\Pi \setminus R_\Pi) < V$.

The above partitions are well aligned with rectangles in the ambient space. The notion of regularity for function classes we choose to study demands that each function in the class has an associated 'nice' partition.

Definition 6 *We say that a class of functions $\mathcal{G} = \{g : [0,1]^p \to \{0,1\}\}$ is V-regular if for each $g \in \mathcal{G}$, there exists a V-regular partition aligned with g.*

Essentially the above demands that the local structure induced by any g can be neatly expressed. This condition is satisfied by many natural function classes on the bulk of their support - An important example is the class of g of the form $\mathbb{1}\{G(x) > 0\}$ for some Hölder smooth G that admit a margin condition with respect to the Lebesgue measure (see, e.g. [MT99; Tsy04]). Indeed, if $\{G\}$ satisfies the margin condition $\mathrm{Vol}(|G| < t) \le \eta$, and is L-Lipschitz, then $\{\mathbb{1}G > 0\}$ is V-regular on a region of mass $\ge 1 - \eta$ with $V \ge (2t/L)^p$.

We offer the obvious class that can budget learn V-regular functions over sufficiently nice measures - rectangles. For $\kappa \in \mathbb{N}$, we define the class $\mathcal{R}^{0,1}_\kappa$ to consist of functions h that may be parametrised by k rectangles $\{R_i\}$ and a label $s \in \{0,1\}$, and take the form

$$h(x; \{R_i\}, s) = s\mathbb{1}x \in \cup R_i + (1-s)\mathbb{1}x \notin \cup R_i.$$

The class $\mathcal{R}^{0,1}_\kappa$ above has VC dimension at most $2p(\kappa+1)$. The theorem below offers bounds on the budgets required to learn V-regular classes in p dimensions:

Theorem 5.4.2. *Let $\kappa := \lfloor d/2p - 1 \rfloor \le 1/V$. Suppose $\mu \ll \mathrm{Vol}$, and $\frac{\mathrm{d}\mu}{\mathrm{d}\,\mathrm{Vol}} \ge \rho$, and $\mathcal{G}$ is V-regular. Then $\textsc{vc}(\mathcal{R}^{0,1}_\kappa) = d$, and it can budget learn $\mathcal{G} \times \{\mu\}$ with*

$$\mathsf{B}\left(\mathcal{G} \times \{\mu\}, \mathcal{R}^{0,1}_\kappa\right) \le 1 - \rho \left\lfloor \frac{d}{2p} - 1 \right\rfloor V/3.$$

Conversely, for $V \le 1/2$, there exists a V-regular class $\mathcal{G}'$ such that if $\textsc{vc}(\mathcal{H}) \le d$, *then*

$$\mathsf{B}(\mathcal{G}' \times \{\mathrm{Vol}\}, \mathcal{H}) \geq 1 - \sqrt{3Vd\log(2e/V)}.$$

For the Lipschitz functions with margin discussed above, V scales as $\widetilde{\Theta}((C/L)^{-p})$, where L is the bound on the gradient, and C is some constant. The above shows that all such classes are learnable with budget $1 - \Omega(1)$ and VC-dim. d iff $d \gtrsim L^{-p+O(\log p)}$

5.4.2.2 Budget Learnability of bounded VC classes

Typically the function classes $\mathcal{G}$ that a cloud can implement are not nearly as rich as the set of all V-regular functions. This merits the investigation of classes with bounded (but large) complexity. Following the lines of study above, we investigate the budget learnability of finite VC classes, assuming $\textsc{vc}\,(\mathcal{G}) = D$ for large D.

Unlike covering numbers, bracketing numbers do not, in general, admit control for VC classes (e.g. constructions of [vHan13] and [Mal12]). This renders the budget learnability problem for bounded VC classes difficult. This is further complicated by the fact that we are interested in whether such classes can be meaningfully bracketed by *low-complexity* classes. Such questions are non-trivial to answer, and, frankly speaking, we do not solve the same. However, we offer two lower bounds, illustrating that if one wishes to non-trivially budget learn such classes with budget $1 - \Omega(1)$, and with VC dim. d, then d must grow as $\Omega(D)$. Further, we a present a few simple, natural cases where one *can* budget learn, irrespective of measure, with budget $\approx 1 - d/D$. We briefly discuss an open question that these classes stimulate.

5.4.3 Lower Bounds

For simplicity, we assume that $\mathcal{X} = [1:N]$ for some $N \gg 1$, and that $\mathscr{F} = 2^{\mathcal{X}}$. The classes $\mathcal{G}, \mathcal{H}$ can then be identified as members of $2^{\mathscr{F}}$. Our lower bounds are captured by the following statements

Theorem 5.4.3.

(i) *(Varying measure) Let $\mathcal{G}$ be any class with* $\textsc{vc}(\mathcal{G}) = D$*, and $\mathcal{H}$ with* $\textsc{vc}(\mathcal{H}) = d$*. Then there exists a measure μ such that*

$$\mathsf{B}(\mathcal{G} \times \{\mu\}, \mathcal{H}) \geq 1 - \sqrt{3\frac{d}{D}\log\frac{eD}{d}}.$$

(ii) *(Uniform measure) Let $N \in \mathbb{N}$, be a multiple of D such that $D \leq N/8e$. There exists a class $\mathcal{G}$ of VC-dimension D on* $[1:N]$*such that for any class $\mathcal{H}$, if* $\mathsf{B}\left(\mathcal{G} \times \{\mathrm{Unif}([1:N])\}, \mathcal{H}\right) \leq \mathsf{B} \in (D/N, 1/4e)$*, then*

$$\textsc{vc}\left(\mathcal{H}\right) \geq D\frac{\log(1/4e\mathsf{B})}{\log(eN)}.$$

The above bounds, while not very effective, indicate that to get small budget it is necessary that d grows linearly with D.

5.4.4 Some natural budget learnable classes

We present three simple examples:

- Sparse VC class: on the space $\mathcal{X} = [1:N]$, let $\mathcal{G} = \binom{\mathcal{X}}{\leq D}$. Then this $\mathcal{G}$ can be budget learned by the class $\binom{\mathcal{X}}{\leq d}$ of VC dimension d with budget $1 - d/D$.

- Convex Polygons in the plane: Let $\mathcal{X} = \mathbb{R}^2$, and $\mathcal{P}_D$ be the set of concepts defined by marking the convex hull of any D points as $b \in \{0,1\}$, and its exterior by $1-b$. [Tak07] shows that $\mathcal{P}_D$ has VC dimension $2D+2$. For $d \geq 4$, the class $\mathcal{P}_d$ (of VC dimension $2d+2$) can budget learn $\mathcal{P}_D$ with budget $1 - \lceil\frac{D}{d-2}\rceil^{-1} \approx 1 - (d/D)$ for $D \gg d$.

- Tensorisation of thresholds: Let $\mathcal{X} = [1:N]$, and let $\mathcal{G}$ be defined as the following class: Let $\mathcal{G}_0$ be the class on $[1:N/D]$ of the form $\mathbb{1}x \geq k$ for some k. We let $\mathcal{G} = \sum_{i=1}^{D} g_i$ where $g_i : [1 + iN/D, (i+1)N/D] \to \{0,1\}$ are of the form $g_i(x) = g_i'(x - iN/D)$ for some $g_i' \in \mathcal{G}_0$. Again, there exists a $\mathcal{H} \subset \mathcal{G}$ of VC dimension d that can budget learn $\mathcal{G}$ with budget $1 - d/D$.

Proofs for the above claims are left to Appendix D.1.4. There are two important features of the above classes, and their budget approximation

1. For each of the classes, there is a *subset* of these classes that has small VC dimension and can budget learn at (roughly) the budget $1 - d/D$. This subclass can be chosen irrespective of measure.

2. These classes are all extremal in the sense of satisfying the sandwich lemma with equality. In the first two cases they are maximal, while the third class is ample (see, e.g. [CCMW18]).

Maximal classes are known to admit unlabelled compression schemes of size equal to their VC dimension, and have many regularity properties - for instance, subclasses formed by restricting the class to some subset of the input are also maximal (see [CCMW18] and references within). It is an interesting open question whether maximal classes of VC dimension D can be budget learned by *subclasses* of VC dimension d with usage $1 - cd/D$ for some constant c.

5.5 Experiments

This section presents empirical work implmenting the BL via bracketing schema on standard machine learning data. We explore three binary classification tasks

1. A simple *synthetic* task in $\mathbb{R}^2$ that allows easy visualisation of the models found by the various strategies, as described in Figure 5·2.

2. The *MNIST odd/even* task, which requires discrimination between odd and even MNIST digits.

3. The *CIFAR random pair* task, which requires discrimination between a pair of

randomly chosen CIFAR-10 classes.[6]

The models considered are presented in Table 5.1. Each of the local classes chosen are far sparser than the corresponding cloud classes, which are taken to be the state of the art models for these tasks.

Task	**Cloud Classifier**	**Cloud Accuracy**	**Local Classifier**	**Local Accuracy**
Synthetic	4th order curve	1.00	Axis-aligned Conic Sections (2nd order curves)	0.840
MNIST Odd/Even	LeNet 2conv + maxpool layers 43.7K params	0.995	Linear 1.57K params	0.898
CIFAR Random Pair	RESNET-32 0.46M params	0.984	Narrow LeNet 2conv + maxpool layers 1.63K params	0.909

Table 5.1: Classification tasks studied, and the corresponding cloud and local classifier classes selected. Cloud accuracy is reported with respect to true labels, but local accuracy is with respect to cloud labels.

Bracketing is implemented essentially using the method of Chapter 4. See §D.2 for detailed descriptions. We compare the bracketing method to four existing approaches.

1. *Sum relaxation* (Sum Relax.) [CDM16], which relaxes the gating formulation to a sum as $\pi\gamma \le (\pi + \gamma)/2$, and then further relaxes this to real valued outputs and convex surrogate losses.

2. *Alternating Minimiation* (Alt. Min.) [NS17a], which introduces an auxiliary function u to serve as proxy for γ during training, replacing $\gamma\pi$ by $u\pi$. The algorithm then optimises a loss over (γ, π, u) via alternating minimisation over (γ, π) and then u, using a KL penalty $D(u\|\gamma)$ to promote $u \approx \gamma$.

3. *Selective Net* (Sel. Net.) [GE19], which is an architectural modification for deep networks that essentially optimises the raw gating setup without any relaxation via SGD.

[6] Note: supervision is provided *after* this choice. That is, if class a and b are chosen, then the algorithms are provided the class a and class b data.

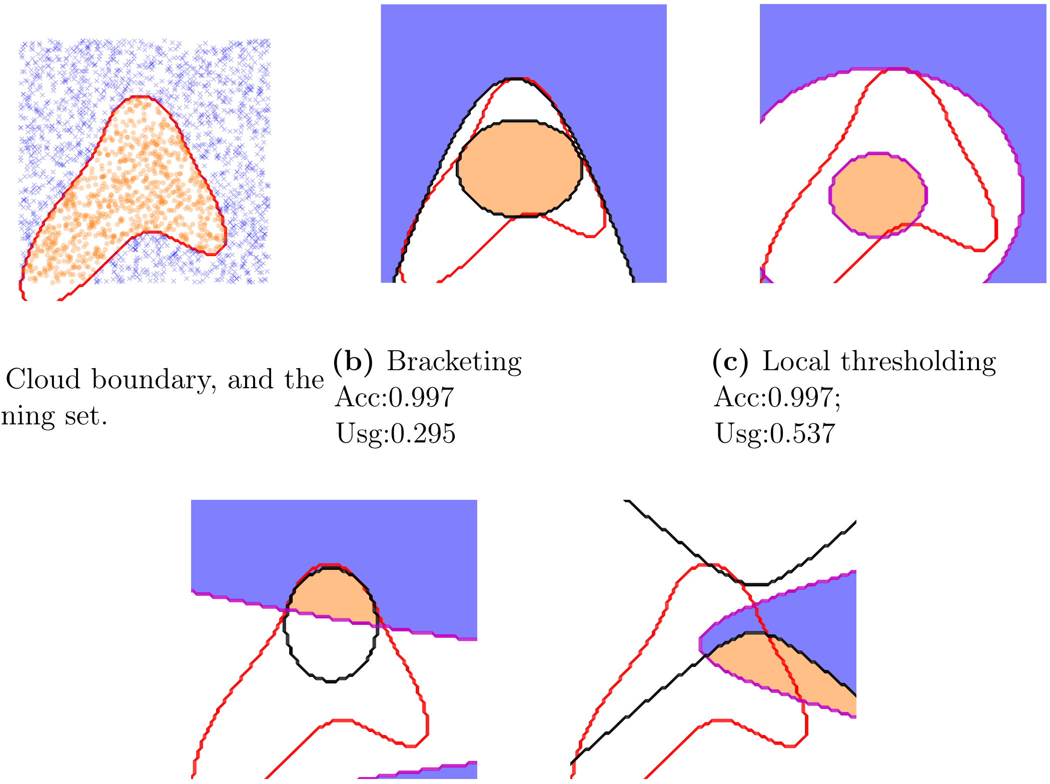
Task	Target	Bracketing		Local Thr.		Alt. Min.		Sum relax.		Sel. Net.		Gain
	Acc.	Usg.	ROL	Usg.	ROL	Usg.	ROL	Usg.	ROL	Usg.	ROL	
MNIST Odd/Even	0.995	0.457	2.19	0.653	1.53	0.830	1.20	0.785	1.27	0.658	1.52	1.431×
	0.990	0.387	2.58	0.515	1.94	0.740	1.35	0.651	1.54	0.544	1.84	1.332×
	0.980	0.299	3.35	0.358	2.79	0.604	1.66	0.651	1.54	0.423	2.37	1.199×
CIFAR Random Pair	0.995	0.363	4.01	0.510	2.25	0.854	1.19	0.620	2.07	0.436	3.04	1.280×
	0.990	0.294	5.66	0.399	3.41	0.754	1.40	0.488	3.31	0.347	4.30	1.265×
	0.980	0.214	9.97	0.276	6.38	0.611	1.87	0.345	5.81	0.257	11.67	1.195×

Table 5.2: Performances on BL tasks studied. Usage (usg.) and *relative operational lifetimes* (ROL), a common metric in BL which is the inverse of usage, are reported. In each case, the models attain the target accuracy (with respect to cloud) to less than 0.5% error - see Table D.2 in §D.2.6. *Gain* is the factor by which the bracketing usages are smaller than the best competitor. The CIFAR entries are averaged over 10 runs of the random choices. The results for all runs for the best two methods for are reported in Table D.1 in §D.2.6. Note that these are averages of each entry for each run and so average ROL is *not* the same as the inverse of the average usage.

4. *Local Thresholding* (Local Thresh.). This is a naïve baseline - one learns a local classifier, and then rejects points if the entropy of its (soft) output at the point is too high.

It is important to contextualise these usage numbers. In our choice of cloud and edge models, we are demanding that the edge models punch far above their weight when we try to budget learn the stated cloud classifiers - indeed, the edge models do not come even close to the clouds in standard accuracy. However, in Table 5.2, we see usages of 20-40% at high accuracies, and relative operational lifetimes (inverse of usage, see, e.g. [Zhu+19]) of 2.5-5. For settings like IoT devices, where communication dominates energy costs, this is a significant gain in operational lifetimes of the prediction pipeline at near SOTA accuracy.

These results demonstrate that the bracketing methodology is practically implementable and effective, with the resulting budget learners clearly outperforming existing methods on the studied tasks.



(a) Cloud boundary, and the training set.

(b) Bracketing
Acc:0.997
Usg:0.295

(c) Local thresholding
Acc:0.997;
Usg:0.537

(d) Alt. Min.
Acc:0.996;
Usg:0.563

(e) Sum Relax.
Acc:0.948;
Usg:0.819

Figure 5·2: Visualisation of classifiers resulting from the various approaches on a synthetic dataset. The red curve indicates the decision boundary of the cloud classifier, and figure (a) indicates this, and also shows the training set used as coloured dots. Figures (b)-(e) depict the budget learners learnt by various approaches. In these, the white region is the set of inputs on which the cloud is queried, while the orange and blue regions describe the decisions of the local predictor when it is queried. The black lines indicate decision boundaries of the various classifiers, and in figures (c)-(e), the magenta line indicates the boundary of the gate. Minimum usage solutions with accuracy at least 99.5% (when found) are presented.

5.6 Directions for Future Work

We think that the bracketing formulation of BL described above is rather nice. It shows practical promise, and gives a clean framework in which to theoretically study BL. Perhaps the biggest open direction in this vein is to study practically relevant classes, and explicitly model of the constraints at the edge and the power of the cloud to determine practical settings of $\mathcal{H}, \mathcal{G}$, and then, with these in hand, develop bracket approximability results that are practically relevant, and, hopefully, more optimistic even in the worst case than the above.

Another important direction is to break the black-box access to the cloud assumption made in the above study. Practically speaking, the cloud model is likely available to the learner during training time, and so it's output can be exploited in a fine grained way to guide the training, and potentially improve the resulting budgeted classifiers. Perhaps more prosaically this question can be posed as how the techniques of distillation and budget learning can be fused. Even beyond this, in practical situations where neither the cloud nor the edge models are perfect, one may synthetically generate even greater accuracy than that of a good cloud model by encouraging the local models to *not* abstain in regions where the cloud model is also wrong (even if this incurs a mistake), which again retains accuracy but further reduces the cloud usage.

Finally, let us mention a couple of technical problems that are insufficiently dealt with in the above. First, we remind the reader about the intriguing question about budget learnability of maximal VC classes by their subclasses. Second, our lower bounds are loose - there's a square root in them that we don't think belongs. They are also not very effective. The square-root comes from the fact that our analysis for these proceeds via covering numbers. Can bounds on bracketing numbers be given more directly, at least in simple cases? Can one remove the dependence on N?

Chapter 6

Online Selective Classification with Limited Feedback

6.1 Introduction

Consider a low-power or battery-limited edge device, such as a sensor or a smart-speaker that receives a stream of classification requests. Due to the resource limitations, such a device cannot implement modern models that are needed for accurate decisions. Instead the device has access (e.g. via an internet connection) to an accurate but resource-intensive model implemented on a cloud server, and may send queries to the cloud server in order to retain accuracy. Of course, this incurs costs such as latency and battery drain due to communication. The ideal operation of such a device should thus be to learn a rule that classifies 'easy' instances locally, while sending harder ones to the cloud, thus maintaining accuracy whilst minimising the net resource consumption [Xu+14; NS17a].

Selective classification [Cho57; Cho70] is a classical paradigm of relevance to such settings. The setup allows a predictor to abstain from classifying some instances (without incurring a mistake). This abstention models adaptive decisions to invoke more resource-intensive methods on subtle cases, like in the above example. The solution

This chapter is a reproduction of the paper [GKCS21], written with Anil Kag, Ashok Cutkosky, and Venkatesh Saligrama. Thanks are due to Tianrui Chen for helpful discussions.

concept is relevant widely - for instance, it is relevant to adaptively recommending further (and costly) tests rather than offering a diagnosis in a medical scenario, or to recommending a human review instead of an alarm-or-not decision in security contexts. Two aspects of such settings are of particular interest to us. Firstly, the cheaper methods are typically not sufficient to realise the true labels, due to which abstention may be a long-term necessity. Secondly, a-priori reliable labels can only be obtained by invoking the resource intensive option, and thus feedback on whether a non-abstaining decision was correct is unavailable.

We propose online selective classification, with an emphasis on ensuring very few mistakes, to account for the need for very accurate decisions. Concretely, an adversary sequentially produces contexts and labels (X_t, Y_t), and the learner uses the X_ts to produce a decision $\widehat{Y}_t$ that may either be one of K classes, or an abstention, which we represent as $\perp$. Feedback in the form of Y_t is provided if and only if $\widehat{Y}_t = \perp$, and the learner incurs a mistake if $\widehat{Y}_t$ was non-abstaining and did not equal Y_t.

With the emphasis on controlling the total number of mistakes, we study regrets achievable when compared to the behaviour of the best-in-hindsight error-free selective classifier from a given class - that is, one that makes no mistakes, while abstaining the fewest number of times. Notice that our situation is non-realisable, and therefore this competitor may abstain in the long-run. The two metrics of importance here are the number of mistakes the learner makes, and its excess abstention over this competitor. An effective learner must control both abstention and mistakes, and it is not enough to make one small, e.g. a learner that makes a lot of mistakes but incurs a very negative excess abstention is no good. This *simultaneous* control of two regrets raises particular challenges.

We construct a simple scheme that, when competing against finite classes, simultaneously guarantees $O(T^{\mu})$ mistakes and $O(T^{1-\mu})$ excess abstentions against adaptive adversaries (for any $\mu \in [0,1]$), and show that these rates are Pareto-tight [OR94]. We further show that against stochastic adversaries, the same rates can be attained with improved dependence of the regret bounds on the size of the class, and we also describe schemes that enjoy similar improvements against adaptive adversaries, but at the cost of the T-dependence of the regret bounds. The main schemes randomly abstain at a given rate in order to gain information, and otherwise play $\widehat{Y}_t$ consistent with the 'version space' of classifiers that have not been observed to make mistakes. For the adversarial case, the analysis of the scheme relies on a new 'adversarial uniform law of large numbers'(ALLN) to argue that such methods cannot incur too many mistakes. This ALLN uses a self-normalised martingale concentration bound, and further yields an adaptive continuous approximation guarantee for the Bernoulli-sampling sketch in the sense of Ben-Eliezer & Yogev [BY20; Alo+21]. The theoretical exploration is complemented by illustrative experiments that implement our scheme on two benchmark datasets.

6.1.1 Related Work

Selective classification has been well studied in the batch setting, and many theoretical and methodological results have appeared [e.g. HW06; BW08; EW10; WE11; KKM12; CDM16; Lei14; GE19; GKS21]. These batch results do not have strong implications for the online setting.

Cortes et al. have studied selective classification in the online setting [Cor+18], but with two differences from our setting. Firstly, rather than individually controlling mistakes and abstentions, the regret is defined according to the Chow loss, which adds up the number of mistakes and c times the number of abstentions, where c is a fixed cost parameter. Secondly (and more importantly) it is assumed that feedback is provided only when the learner *does not* abstain, rather than only when it does. This difference arises from the underlying situations being modelled - Cortes et al. view the abstention as a decision given to a user in which case no feedback is forthcoming, while we view it as a decision to invoke further processing. Both of the scenarios are reasonable, and so both of these explorations are valid, however it is unclear what implications one set of results have for the other.

A similar decision and feedback model as ours was proposed by Li et al. in the 'knows what it knows' (KWIK) framework [LLWS11]. The KWIK model, however, fundamentally views abstentions as a short term action, typically arguing that only a finite number of these are made. This is viable since Li et al. study this model in an essentially realisable setting, wherein the optimal labels are known to be essentially realised by a given class - notice that in such a case, a single abstention at an instance x determines what value should be played there in the long run. Our interest however lies in the situation where this data cannot be represented in such a way, and such strategies are not viable since the labels may be noisy. Our work thus generalises the KWIK setting to non-realisable data, and to situations wherein abstention is a valid long-term action, as motivated in the introduction, by studying behaviour against competitors that may abstain.[1]

While Szita and Szepesvári have extended the KWIK formulation to the agnostic case in a regression setting [SS11], this work also focuses of limiting the number of abstentions to be finite rather than long-run abstentions. Concretely it is assumed that $Y_t = g(X_t) + \text{noise}$, for some function g, and the learner knows a class $\mathcal{H}$, and a bound Δ such some $h \in \mathcal{H}$ is Δ-close to g (in an appropriate norm). Using the knowledge of Δ, they describe schemes that have limited abstention, but at the cost of mistakes, by producing responses $\hat{Y}_t$ that are up to $(2 + o(1))\Delta$ separated from Y_t. In contrast, in our formulation, contexts X_t for which no function in $\mathcal{H}$ can

[1] The KWIK model also bears other significant differences. It posits an input parameter ε, and requires that the learner either abstains, or produces an ε-accurate response. A notion of competitor is not invoked, and rather than studying regret, the number of abstentions needed to achieve this ε-accuracy is studied.

represent the ground truth g well would always be abstained upon. In addition to this work, trade-offs between mistakes and abstentions in a relaxed version of the KWIK framework have been considered [ZC16; SZB10; DZ13], and in particular the agnostic case has been explored by Zhang and Chaudhuri [ZC16], but unlike our situation this relaxed KWIK model requires full-feedback to be available whether or not the learner abstains. Neu and Zhivotovskiy [NZ20] also work in this relaxed model, and show that when comparing the standard loss of a *non-abstaining* classifier against the Chow loss of an abstaining learner, regrets independent of time can be obtained.

Due to the limited feedback, our setting is related to partial-monitoring [LS20, Ch. 37]. Viewing actions as choices over functions, our setting has feedback graphs [MS11] that connect abstaining actions to every other action and themselves. The novelty with respect to partial-monitoring arises from the fact that we individually control two notions of losses, rather than a single one. It's unclear how to apply the generic partial-monitoring setup to this situation - indeed, naïvely, our game is only weakly observable in the sense of Alon et al.[ACDK15], and one would expect $\Omega(T^{2/3})$ regrets, while we can control both mistakes and excess abstention to $\tilde{O}(\sqrt{T})$. A limited feedback setting where two 'losses' *are* individually controlled is label-efficient prediction [CLS05], where a learner must query in order to get feedback. However, in our setting, abstentions are both a way to gather feedback, and also necessary to prevent mistakes. That is, our competitor may abstain regularly, but makes few mistakes, while in this prior work the competitor does not abstain, but may make many mistakes. The resulting scenario is both qualitatively and quantitatively distinct, e.g. in label-efficient prediction, the smallest symmetric rate of number of queries and excess mistakes is again $\Theta(T^{2/3})$.

6.2 Setting, and Problem Formulation

Setup Let $\mathcal{X}$ be a feature space, $\mathcal{Y}$ a finite set of labels, and $\mathcal{F}$ a finite class of selective classifiers, which are $\mathcal{Y} \cup \{\perp\}$ valued. For simplicity, we assume that $\mathcal{F}$ contains the all abstaining classifier (i.e. the function $f_\perp$ such that $\forall x, f_\perp(x) = \perp$). We will denote $|\mathcal{F}| = N$. The setting may be described as a game between a learner and an adversary (or more prosaically, a data generating mechanism) proceeding in T rounds. Also for simplicity, we will assume that T is known to both the learner and the adversary in advance. The objects in this game are the context process, $X_t \in \mathcal{X}$, the label process $Y_t \in \mathcal{Y}$, the action process $\widehat{Y}_t \in \mathcal{Y} \cup \{\perp\}$ and the feedback process $Z_t \in \mathcal{Y} \cup \{*\}$, where $* \notin \mathcal{Y}$ is a trivial symbol. The information sets of the adversary and learner up to the tth round are respectively $\mathscr{H}^{\mathfrak{A}}_{t-1} := \{(X_s, Y_s, \widehat{Y}_s) : s < t\}$, and $\mathscr{H}^{\mathfrak{L}}_{t-1} := \{(X_s, \widehat{Y}_s, Z_s) : s < t\}$.

The Game For each round $t \in [1 : T]$, the adversary produces a context and a label (X_t, Y_t) on the basis its history $\mathscr{H}^{\mathfrak{A}}_{t-1}$. The learner observes only the context, X_t, and on the basis of this and its history $\mathscr{H}^{\mathfrak{L}}_{t-1}$, produces an action $\widehat{Y}_t$. We will say that this action is an abstention if $\widehat{Y}_t = \perp$, and that it is a prediction otherwise. If the action was an abstention, set $Z_t = Y_t$, and otherwise to $*$. The learner then observes Z_t, and the round concludes. Notice that since Z_t is a deterministic function of Y_t and $\widehat{Y}_t$, and since the adversary observes both, $\mathscr{H}^{\mathfrak{L}}_{t-1}$ can be determinstically generated from $\mathscr{H}^{\mathfrak{A}}_{t-1}$. Due to the same reason, $\widehat{Y}_t$ and Y_t are conditionally independent given $(X_t, \mathscr{H}^{\mathfrak{A}}_{t-1})$.

Adversaries are characterised by a sequence of conditional laws on (X_t, Y_t) given $\mathscr{H}^{\mathfrak{A}}_{t-1}$ (and $T, \mathcal{F}$). In the following we will explicitly consider two classes of such laws:

- Stochastic Adversary: (X_t, Y_t) are drawn according to a fixed law, P, unknown to the learner, independently of $\mathscr{H}^{\mathfrak{A}}_{t-1}$.

- Adaptive Adversary: (X_t, Y_t) are arbitrary random variables with a $\sigma(\mathscr{H}_{t-1}^{\mathfrak{A}})$ measurable joint law.

We will denote a generic class of adversaries as $\mathscr{C}$.

Performance Metrics The two principal quantities of interest are the number of mistakes made by the learner, and the number of times it has abstained. We will denote these as

$$M_T := \sum_{t \le T} \mathbb{1}\{\widehat{Y}_t \notin \{\perp, Y_t\}\}, \quad \textit{and} \quad A_T := \sum_{t \le T} \mathbb{1}\{\widehat{Y}_t = \perp\}.$$

As previously discussed, the performance of a learner is measured in terms of regret with respect to the best-in-hindsight abstaining classifier from $\mathcal{F}$ that makes no mistakes, that is

$$f^* \in \arg\min_{f \in \mathcal{F}} \sum_{t \le T} \mathbb{1}\{f(X_t) = \perp\} \quad \text{s.t.} \quad \sum_{t \le T} \mathbb{1}\{f(X_t) \notin \{\perp, Y_t\}\} = 0.$$

Note that such an f^* is always realised, since the class is finite, and since it contains the all abstaining classifier. Let $A_T^* := \sum_{t \le T} \mathbb{1}\{f^*(X_t) = \perp\}$ denote the value of the minimum above. The principal metrics of interest to us are the *abstention regret* $A_T - A_T^*$, and the *total mistakes* M_T.

Solution Concept The two performance metrics naturally involve a tradeoff - for instance, making some mistakes may allow a learner to drastically reduce its abstention regret to the point that it is negative. We pursue the trade-off between the worst possible behaviour of either regret.

Definition (Regret Achievability) *For functions $\varphi, \psi : \mathbb{N}^2 \to \mathbb{R}$, we say that expected regret bounds of (φ, ψ) are achievable against a class of adversaries $\mathscr{C}$ if there exists* *a learner such that for every adversary in $\mathscr{C}$, $\mathbb{E}[A_T - A_T^*] \le \varphi(T,N)$ and $\mathbb{E}[M_T] \le \psi(T,N)$.*

As is common, we are interested in the growth rates of achievable bounds with T. We thus define

Definition (Achievable rates) *we say that asymptotic expected-regret rates of $(\alpha, \mu) \in [0,1]^2$ are achievable against a class of adversaries $\mathscr{C}$ if an expected regret bound of (φ, ψ) can be achieved against it for functions φ, ψ, said to be witnesses for the rate, such that*

$$\limsup_{T\to\infty} \frac{\log \varphi(T,N)}{\log T} \le \alpha \quad \textit{and} \quad \limsup_{T\to\infty} \frac{\log \psi(T,N)}{\log T} \le \mu.$$

Notice that if (α, μ) is an achievable rate, so is (α', μ') for $\alpha' \ge \alpha, \mu' \ge \mu$. As a result, the lower boundary of the set of achievable rates is well defined, and we will refer to this as the *Pareto frontier of achievable rates.* This is equivalently characterised by the function $\underline{\alpha}(\mu) := \inf\{\alpha : (\alpha, \mu) \text{ is an achievable rate}\}$. This is well defined since $\forall \mu, (1, \mu)$ is achievable by always abstaining.

6.3 The Adversarial Case

We begin with the adversarial case. The scheme, called the 'versioned uniform explorer' (VUE) is described below, and we discuss both the motivation of the scheme, and its analysis.

The main idea underlying VUE is that any function f that is observed to make a mistake on an instance X_t (due to the learner abstaining on this instance) can be removed from future consideration, since we are only trying to match the behaviour of the competitor f^*, and clearly $f \neq f^*$ as it has made a mistake. This motivates setting up a 'version space,'

$$\mathcal{V}_t := \left\{ f : \sum_{s<t} \mathbb{1}\{Z_s \neq *, f(X_s) \notin \{\perp, Y_s\}\} = 0 \right\},$$

the set of functions that are consistent with the observations made up to time t. Notice that $f^* \in \mathcal{V}_t$ for all t. Given $\mathcal{V}_t$, we can restrict to playing an action in the set $\widehat{\mathcal{Y}}_t := \{f(X_t) : f \in \mathcal{V}_t\}$ - $f^*(X_t)$ lies in this set, and thus any action outside of it can be eliminated. Of course, if $\widehat{\mathcal{Y}}_t$ is a singleton, then it contains $f^*(X_t)$, and we can just play it.

Algorithm 2 VUE

```
Inputs: $\mathcal{F}$, Exploration rate $p$.
Initialise: $\mathcal{V}_1 \leftarrow \mathcal{F}$.
for $t \in [1:T]$ do
    $\widehat{\mathcal{Y}}_t \leftarrow \{f(X_t) : f \in \mathcal{V}_t\}$.
    if $|\widehat{\mathcal{Y}}_t| = 1$ then
        $\widehat{Y}_t \leftarrow f(X_t)$ for any $f \in \mathcal{V}_t$.
        $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_t$.
    else
        Sample $C_t \sim \mathrm{Bern}(p)$.
        if $C_t = 1$ then
            Set $\widehat{Y}_t = \perp$, observe $Y_t$.
            $\mathcal{U}_t \leftarrow \{f : f(X_t) \in \{\perp, Y_t\}\}$
            $\mathcal{V}_{t+1} = \mathcal{V}_t \cap \mathcal{U}_t$.
        else
            Pick $\widehat{Y}_t \in \widehat{\mathcal{Y}}_t \setminus \{\perp\}$.
            $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_t$.
```

Next, since we are incentivised to minimise the total number of abstentions, it behooves us to play non-abstaining actions whenever possible. However, this puts us in a bind, since feedback is produced only when we play an abstaining action. Taking inspiration from [CLS05], we abstain at a rate p by tossing a biased 'exploratory coin', C_t, abstaining when $C_t = 1$, and otherwise playing any non-abstaining action in $\widehat{\mathcal{Y}}_t$. Clearly, such a strategy can incur at most pT excess abstention regret in expectation. Mistakes made by this strategy are controlled via the following 'adversarial law of large numbers' (ALLN).

Lemma 6.3.1. *Let $\{\mathscr{F}_t\}_{t=1}^\infty$ be any filtration, and $\{U_t\}_{t=1}^\infty, \{B_t\}_{t=1}^\infty$ be $\{\mathscr{F}_t\}$-adapted binary processes, such that $B_t \sim \mathrm{Bern}(p)$, $p < 1/2$ is jointly independent of $\mathscr{F}_{t-1}, U_t$*

for each t. Let $W_t = \sum_{s\le t} U_s$, and $\widetilde{W}_t = \sum_{s\le t} U_s B_s$. For any $\delta \in (0, {}^1\!/\!\sqrt{e})$,

$$\mathbb{P}\left(\exists t : \widetilde{W}_t \le 1, W_t > \frac{8\log(1/\delta)}{p}\right) \le \delta.$$

The above is argued in §E.1 using a self-normalised martingale tail inequality [HRMS20]. We note that this self-normalisation is critical, and without this techniques such as Freedman's inequality yield an extraneous $\sqrt{T}$ factor in the bounds that is untenable for our purposes. The same argument, along with the shaping technique of Howard et al. [HRMS18] yields a Bernstein-type law of iterated logarithms that controls $|W_t - {}^{\widetilde{W}_t}\!/\!{}_p|$ at a level $\tilde{O}(1/p + \sqrt{W_t/p \log\log t})$, which should be useful more broadly. This full version (presented in §E.1) further shows that the 'Bernoulli-sampler' [BY20; Alo+21] offers a continuous approximation in the sense of Ben-Eliezer & Yogev [BY20], but with the error for sets of low incidence flattened as expected due to Bernstein's inequality.

For our purposes, the point of Lemma 6.3.1 is to allow us to argue that no matter what the adversary does, if we uniformly abstain at a rate p, then we will 'catch' any mistake-prone function before it makes $O(1/p)$ mistakes. Exploiting a union bound, this in turn means that with high probability, any such function will fall out of the version space $\mathcal{V}_t$ before it has incurred much more than $\log N/p$ mistakes. Since the label produced by Algorithm 2 must equal $f(X_t)$ for *some* f in the version space, we can infer that the number of mistakes the learner makes is at most the number of times any function in the version space is wrong. Using the Lemma yields a bound of $\widetilde{O}(1/p)$ on the number of mistakes that any functions in the version space can have ever made, and since there are only N possible functions, in total the number of mistakes the learner can make is bounded as $\widetilde{O}(N/p)$. More formally, the argument, presented in §E.2, argues this for a single function $f \in \mathcal{F}$ by instantiating the lemma with $\mathscr{F}_t = \sigma(\mathscr{F}_t = \sigma(\mathscr{H}_t^{\mathfrak{A}}), B_t = C_t$, and $U_t^f := 1\{f(X_t) \notin \{\perp, Y_t\}\}$. The resulting

$\widetilde{W}_t^f$ is the number of mistakes f is *observed* to have made, and $f \in \mathcal{V}_t$ if and only if $\widetilde{W}_t^f = 0$[2]. Along with a use of Bernstein's inequality to control A_T this yields the result below.

Theorem 6.3.2. *Algorithm 2 instantiated with $p < {}^1\!/_2$, and run against an adaptive adversary, attains the following with probability at least $1-\delta$ over the randomness of the learner and the adversary:*

$$M_T \le \frac{9N\log(2N/\delta)}{p}$$
$$A_T - A_T^* \le pT + \sqrt{2p(1-p)T\log(2/\delta)} + 2\log(2/\delta).$$

In particular, taking $p = \sqrt{N/T}$ yields the symmetric regret bound

$$\max(M_T, A_T - A_T^*) \lesssim \sqrt{NT}\log(N/\delta).$$

We conclude with a few remarks.

Achievable rates Taking $\delta = {}^1\!/_T$, and varying p in $({}^{\log T}\!/_T, 1]$ gives the rates attainable by VUE

Corollary 6.3.3. *All rates (α, μ) such that $\alpha > 0, \alpha + \mu > 1$ are achievable against adaptive adversaries.*

These rates are tight - as expressed in Corollary 6.4.3, rates such that $\alpha + \mu < 1$ are not achievable even against stochastic adversaries. The Pareto frontier is therefore the line $\alpha + \mu = 1$.

Dependence on N It should be noted that the dependence on the number of functions, N, in Thm. 6.3.2 is polynomial, as opposed to the more typical logarithmic dependence on the same in online classification. The problem of characterising this

[2] This argument only needs control for the case $\widetilde{W}_t = 0$. The ≤ 1 in Lemma 6.3.1 is exploited in §6.5.2.

dependence appears to be subtle, and we do not resolve the same. In the following section, we explore schemes that improve this aspect, but at a cost - §6.4 yields logarithmic dependence against stochastic adversaries, while §6.5 gives a scheme that has a logarithmic dependence against adaptive adversaries, but worse dependence with T.

It is worth stating that the analysis above is tight for Algorithm 2 - consider the domain $\mathcal{X} = [1 : N]$, and the class $\mathcal{F} = \{f_t : t \in [0 : N]\}$ such that $f_t(x) = \perp$ if $x \le t$ and $= 1$ if $x > t$. Now consider an adversary that chooses a t^* in advance, and presents the contexts 1 T/N times, 2 T/N times and so on, labelling contexts smaller than t^* as 0, and contexts larger than t^* as 1. Notice that in each case, there is exactly one function in $\mathcal{V}_t$ that does not abstain. The scheme above incurs $\Omega(pT(1 - t^*/N))$ excess abstention, and $\Omega(t^*/p)$ mistakes, and linearly large t^* form a tight example. Of course, this is not a lower bound on this problem, and the question of the optimal dependence on N remains open.

Hedge-Type Schemes The natural approach of proceeding by weighing the cost of abstention versus a mistake, and running a hedge-type scheme on an importance-estimate of the resulting loss does not lead to tight rates - the scheme MIXED-LOSS-PROD of §6.5 pursues precisely this strategy, and the worse case symmetric regret bounds that standard analyses lead to scale as $T^{2/3}$ instead of as $T^{1/2}$ as for VUE (Cor. 6.5.2). This may be due to the fact certain-error prone classifiers in $\mathcal{F}$ may have very low abstention rates, and thus overall large weight, and it is unclear how to eliminate this behaviour.

6.4 The Stochastic Case

Algorithm 3 VUE-PROD

```
**Inputs**: $\mathcal{F}, p$, Learning rate $\eta$.
**Initialise**: $\mathcal{V}_1 \leftarrow \mathcal{F}, \forall f, w_1^f \leftarrow 1$.
**for** $t \in [1:T]$ **do**
    Sample $f_t \sim \pi_t = \frac{w_t^f \mathbb{1}\{f \in \mathcal{V}_t\}}{\sum_{f \in \mathcal{V}_t} w_t^f}$.
    Toss $C_t \sim \mathrm{Bern}(p)$.
    $\widehat{Y}_t \leftarrow \begin{cases} \perp & C_t = 1 \\ f_t(X_t) & C_t = 0 \end{cases}$.
    $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_t$.
    **if** $C_t = 1$ **then**
        $\mathcal{U}_t \leftarrow \{f : f(X_t) \in \{\perp, Y_t\}\}$
        $\mathcal{V}_{t+1} = \mathcal{V}_t \cap \mathcal{U}_t$.
    **for** $f \in \mathcal{V}_{t+1}$ **do**
        $a_t^f \leftarrow \mathbb{1}\{f(X_t) = \perp\}$
        $w_{t+1}^f \leftarrow w_t^f \cdot (1 - \eta a_t^f)$.
```

This section argues that the regret bounds of Thm. 6.3.2 can be improved to behave logarithmically in N in the stochastic setting. There are a couple of issues with Algorithm 2 that impede a better analysis in the stochastic case. The first, and obvious, one is that how $\widehat{Y}_t$ is chosen is not specified. More subtly, the fact that the scheme insists on playing non-abstaining actions whenever possible makes it difficult to control the number of mistakes without a polynomial dependence on N.

We sidestep these issues in Algorithm 3 by maintaining a law π_t on functions in $\mathcal{V}_t$ that only depends on $\mathscr{H}_{t-1}^{\mathfrak{L}}$, and predicting by setting $\widehat{Y}_t = f(X_t)$ for $f_t \sim \pi_t$. Notice that playing this way it is possible that we abstain on X_t even if the exploratory coin comes up tails. We control mistakes by arguing that very error-prone functions are all quickly eliminated (due to the stochasticity), and using the property that π_t does not depend on X_t to limit the mistakes incurred up to such a time. Abstention control follows by choosing π according to a strategy that favours fs with small overall abstention rate over the history. In Algorithm 3, we use a version of the PROD scheme of [CMS07] to set weights, analysed with shrinking decision sets. The following is shown along these lines in §E.3.

Theorem 6.4.1. *Algorithm 3, run against stochastic adversaries with $\eta = p$, attains the regret bounds*

$$\mathbb{E}[M_T] \le 8\frac{\log T \log(NT)}{p}, \quad \textit{and} \quad \mathbb{E}[A_T - A_T^*] \le pT + \frac{\log N}{p}.$$

We note that VUE-PROD also enjoys favourable bounds in the adversarial case - mistakes are bounded as $\tilde{O}(N/p)$, and abstention regret as in the above result. This is in contrast to simpler follow-the-versioned-leader type schemes that also satisfy similar bounds as Thm. 6.4.1 in the stochastic case. Also note that the above cannot attain rates such that $\alpha \le 1/2$, an inefficiency introduced due to the conditional independence of π_t and X_t.

Finally, we show a lower bound. The statement equates stochastic adversaries with their laws.

Theorem 6.4.2. *If $\mathcal{F}$ contains two functions f_1, f_2 such that there exists a point x for which $f_1(x) = \perp \neq f_2(x)$, then for every $\gamma \in [0, 1/2]$, there exists a pair of laws P_1^γ, P_2^γ such that any learner that attains $\mathbb{E}_{P_1^\gamma}[A_T - A_T^*] = K$ must incur $\mathbb{E}_{P_2^\gamma}[M_T] \ge \gamma(e^{-2\gamma K}T - K)$.*

Thus, if a (φ, ψ) regret bound with $\sup \frac{\varphi}{T} < \frac{1}{2e^2}$ is achievable, then $\varphi \cdot \psi = \Omega(T)$. Indeed, using the above with $\gamma = 1/\varphi(T, N)$, gives $\mathbb{E}_{P_1}[A_T - A_T^*] = K \le \varphi(T, N)$, and so $\psi(T, N) \ge \mathbb{E}_{P_2}[M_T] \ge \frac{T}{\varphi(T,N)} e^{-2K/\varphi(T,N)} - 1$. This proves the following.

Corollary 6.4.3. *If $(\alpha, \mu) \in [0, 1]^2$ is such that $\alpha + \mu < 1$, then an (α, μ) regret rate is not achievable against stochastic adversaries, and, a fortiori, against adaptive adversaries.*

6.5 Reducing the dependence of regret bounds on N in the adversarial case

This section concentrates on improving the N-dependence of regret bounds in the adversarial case via two avenues. The first improves this dependence to $\log(N)$ by running PROD with a weighted loss, but at the cost of increasing T dependence. This holds greatest relevance when T is bounded as a polynomial of N, which is of interest because N can be quite large even in reasonable settings - e.g., a discretisation of d-dimensional hyperplanes induces $N = \exp(Cd)$. The second approach considers the case when the set of possible contexts, i.e. $\mathcal{X}$ is not too large. While in this case, N can be as large as $(|\mathcal{Y}| + 1)^{|\mathcal{X}|}$, we show bounds depending only linearly on $|\mathcal{X}|$.

6.5.1 Weighted PROD

Algorithm 4 MIXED-LOSS-PROD

```
Inputs: $\mathcal{F}$, Exploration rate $p$, Learning rate $\eta$.
Initialise: $\forall f \in \mathcal{F}, w_1^f \leftarrow 1$.
for $t \in [1 : T]$ do
    Sample $f_t \sim \pi_t = {}^{w_t^f}\!/\!{}_{\sum w_t^f}$.
    Toss $C_t \sim \mathrm{Bern}(p)$.
    if $C_t = 1$ then
        $\widehat{Y}_t \leftarrow \perp$
    else
        $\widehat{Y}_t \leftarrow f_t(X_t)$
    $\forall f \in \mathcal{F}$, evaluate $\ell_t^f$
    $w_{t+1}^f \leftarrow w_t^f(1 - \eta \ell_t^f)$.
```

We continue the uniform exploration, but play according to the PROD method, with the loss

$$\ell_t^f := C_t \mathbb{1}\{f(X_t) \notin \{\perp, Y_t\}\} + \lambda \mathbb{1}\{f(X_t) = \perp\},$$

where λ both trades-off the relative costs of mistakes and abstentions, in the vein of the fixed cost Chow loss, and accounts for the sub-sampling of the mistake loss.

The analysis of this scheme, presented in §E.4, exploits the quadratic bound of PROD due to [CMS07] to control the sum $\mathbb{E}[pM_T + \lambda(A_T - pT)]$ by $\min_g \log N/\eta + \sum \eta(\ell_t^g)^2$, where the expectation is only over the coins C_t, and the $-pT$ term is due to the extra abstentions due to the exploratory coin. The key observation is that since f^* makes no mistakes, $\sum(\ell_t^{f^*})^2 = \lambda^2 A_T^*$, and so taking $g = f^*$, and exploiting the weight allows us to separately control the regrets in terms of A_T^*.

Theorem 6.5.1. *Algorithm 4, when run against adaptive adversaries with $\eta = {}^1\!/\!_2, \lambda \leq p$, attains*

$$\mathbb{E}[M_T] \leq \frac{2\log N}{p} + \frac{2\lambda}{p}\mathbb{E}[A_T^*], \quad \textit{and} \quad \mathbb{E}[A_T - A_T^*] \leq pT + \frac{2\log N}{\lambda}.$$

6.5.1.1 Rates

Theorems 6.4.1 and 6.5.1 show regret bounds with logarithmic dependence in N. The following concept separates rates attainable with this advantageous property from those with worse N-dependence.

Definition (Logarithmically Achievable Rates) *We say that rates (α, μ) are logarithmically achievable against adversaries from a class $\mathscr{C}$ if there exists a learner that attains a (ψ, φ)-regret against such adversaries for ψ, φ that witness the rate (α, μ), and satisfy that for every fixed T, $\max(\varphi(T, N), \psi(T, N)) = O(\mathrm{polylog}(N))$.*

Since $A_T^* \leq T$, choosing $p = T^{-u}, \lambda = T^{-(u+v)}$ in MIXED-LOSS-PROD for any $(u, v) \in [0, 1]^2, u + v \leq 1$ allows us to attain rates of the form $(\alpha, \mu) = (\max(1 - u, u + v), 1 - v)$. Notice that for any fixed v, the smallest α so attainable is ${}^{1+v}\!/\!_2$. This shows

Corollary 6.5.2. *Any rate (α, μ) such that $\alpha + \mu/2 > 1$ is logarithmically achievable against adaptive adversaries.*

The following figure illustrates the worst case achievable rate regions in the three cases considered.

Adaptive Rates Observe that if $A_T^* \asymp T^{\alpha^*}$ for some $\alpha^* < 1$, then nominally, the achievable rates can be improved. Indeed, with the parametrisation $p = T^{-u}, \lambda = T^{-u+v}$, we may attain rates of the form $(\alpha, \mu) = (\max(1-u, u+v), \max(u, \alpha^* - v))$. Further, a given mistake rate μ can be attained by setting $u = \mu$, and $\alpha^* - v \le \mu$. With these constraints, the smallest abstention rate attainable is

$$\widetilde{\alpha}(\mu; \alpha_*) = \max\left(1 - \mu, (1 + (\alpha^* - \mu)_+)/2\right),$$

achieved by setting $v = (\alpha_* - \mu)_+, u = \min(1 - (\alpha^* - \mu)_+, 2\mu)/2$. Such rates can in fact be attained adaptively, without prior knowledge of α^*. The main bottleneck here is that the quantity A_T^* is not observable. However, every function g that is never *observed* to make a mistake satisfies $\sum (\ell_t^g)^2 = \lambda^2 \sum \mathbb{1}\{g(X_t) = \bot\}$, and such functions are identifiable given $\mathscr{H}_t^{\mathfrak{L}}$. Let

$$B_t^* := \min \sum_{s \le t} \mathbb{1}\{g(X_s) = \bot\} \quad \text{s.t.} \quad \sum_{s \le t} C_s \mathbb{1}\{g(X_s) \notin \{\bot, Y_t\} = 0.$$

Note that B_t^* grows monotonically, and is always smaller than $A_t^* = \sum_{s \le t} \mathbb{1}\{f^*(X_t) = \bot\}$. We show the following in §E.4.1 via a scheme that adaptively sets p, λ according to B_t^*.

Theorem 6.5.3. *For any $\alpha^*, \mu, \varepsilon \in (0, 1]$, Algorithm 6 attains, without prior knowledge*

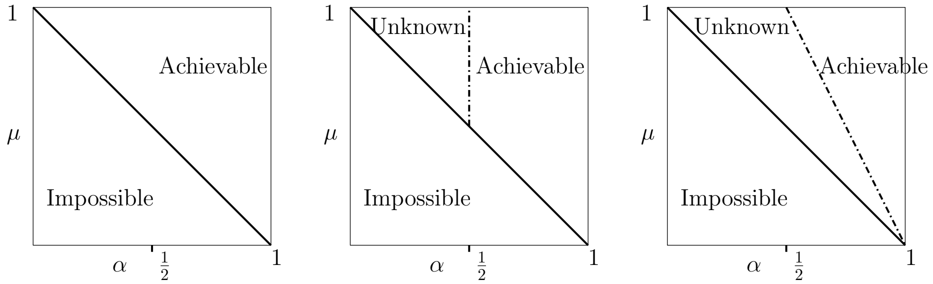


Figure 6·1: Left shows rates achievable against adaptive adversaries. Middle and right show logarithmically achievable rates against stochastic and adaptive adversaries respectively.

of α^*, any rate of the form $(\widetilde{\alpha}(\mu, \alpha^*) + \varepsilon, \mu + \varepsilon)$ against adaptive adversaries that induce $A_T^* \le T^{\alpha^*}$ almost surely.

The rates $\widetilde{\alpha}$ essentially interpolate between the second and third panels of Fig. 6·1. Concretely the region achieved consists of the intersection of the regions $\{\alpha > 1/2\}$, $\{\alpha + \mu > 1\}$ and $\{2\alpha + \mu > 1 + \alpha^*\}$, with the last set being active only when $\alpha^* \ge {}^1\!/_2$.

6.5.2 A $|\mathcal{X}|$-dependent analysis of VUE

We give an alternate mistake analyse for VUE over finite domains. The analysis is slightly stronger: let $\mathbf{y} \in ([1:K] \cup \{\perp\})^{|\mathcal{F}|}$ be indexed by elements of $\mathcal{F}$, with the 'fth' entry $\mathbf{y}_f$ reprsents a value that f might take. Consider the resulting partition of $\{\mathcal{X}_\mathbf{y}\}_{\mathbf{y} \in ([1:K] \cup \{\perp\})^{\mathcal{F}}}$, where each part $\mathcal{X}_\mathbf{y} \subset \mathcal{X}$ contains points that have the same pattern of function values, that is $\mathcal{X}_\mathbf{y} = \{x : \forall f \in \mathcal{F}, f(x) = \mathbf{y}_f\}$. The following argument can be run unchanged by replacing single xs in the following by all xs in one $\mathcal{X}_\mathbf{y}$. That is, we may replace $|\mathcal{X}|$ in the following Theorem 6.5.4 with $|\{\mathcal{X}_\mathbf{y}\}|$. For simplicity, we present the argument for $|\mathcal{X}|$ only.

Denote $\widehat{\mathcal{Y}}_t^x := \{f(x) : f \in \mathcal{V}_t\}$. Notice that after the first time t such that $X_t = x, \widehat{Y}_t = \perp$, we will remove from the version space all classifiers that did not abstain or output the correct classification at time t. Thus if we define $y^x \in [1:K]$ to be Y_t, then for all subsequent times, $\widehat{\mathcal{Y}}_t^x \subset \{\perp, y^x\}$. As a result, if we observe two mistakes at any given x, then we cannot make any more mistakes at a subsequent time t' with $X_{t'} = x$, because the only remaining decision in $\widehat{\mathcal{Y}}_{t'}^x$ must be $\perp$.

We may now proceed in much the same way as §6.3 - instantiate $U_t^x = \mathbb{1}\{X_t = x, \widehat{Y}_t \notin \{\perp, Y_t\}\}$, $B_t = C_t$, and union bound over the xs. Then $|\widehat{\mathcal{Y}}_t^x| \ge 2$ if and only if $\widetilde{W}_t^x \le 1$, and, invoking Lemma 6.3.1, up to such a time at most $W_t^x = O({}^{\log|\mathcal{X}|}\!/_p)$ mistakes may be made on instances such that $X_t = x$. But then totting up, we make at most $O(|\mathcal{X}| \log |\mathcal{X}|/p)$ mistakes, as encapsulated below

Theorem 6.5.4. *Algorithm 2 instantiated with $p \le 1/2$ and run against an adaptive adversary, attains the following with probability at least $1-\delta$ over the randomness of the learner and the adversary:*

$$M_T \le \frac{9|\mathcal{X}|\log(2|\mathcal{X}|/\delta)}{p}$$
$$A_T - A_T^* \le pT + \sqrt{2p(1-p)T\log(2/\delta)} + 2\log(2/\delta).$$

Along with the bound itself, the above result makes a couple of points regarding the characterisation of N-dependence of the regrets in online selective classification. Firstly, it suggests that efficient analyses, and possibly schemes, must incorporate the structure of $\mathcal{X}$; and secondly it shows that constructions that attempt to show superlogarithmic in N lower bounds must have both N and $|\mathcal{X}|$ large, and thus typical strategies placing a very rich class on a small domain will not be effective.

6.6 Experiments

We evaluate the performance of Algorithm 3 on two tasks - CIFAR 10 [Kri09], and GAS [Ver+12] - see §E.5 for details of implementation, and here for the relevant code. The former represents a setting where an expert can be adaptively invoked, which we treat by providing the true labels of the classes upon abstention. The second case is more explicitly an adaptive feature selection task - the GAS dataset has features from 16 sensors, and we train one model, g, on all of this data, while the selective classification task operates on data from the first 8 sensors only, and receives the output of g when abstaining. The standard accuracies of the model classes we implement are $\sim 90\%$ on CIFAR-10, and $\sim 77\%$ on GAS. In both cases, a training set is used to learn a parameterized family of selective classifiers, $f_{\mu,t}$. The hyperparameters (μ, t) provide control over various levels of accuracy and abstention. For training, we leverage a recent method [GKS21] that yields such a parameterisation, which is discretised to get $N = 600$ of these functions to form our class $\mathcal{F}$. We then sequentially classify the test datasets of each of the tasks.

One subtlety with the setting is that none of the selective classifiers in $\mathcal{F}$ actually make no mistakes. To avoid the trivialities emerging from this, we relax the versioning condition to only drop classifiers that are seen to make mistakes on at least $\varepsilon N_t + \sqrt{2\varepsilon N_t}$ mistakes at time t, where N_t is the number of times feedback was received up to time t, and the second term handles noise. Additionally, if it turns out that all functions in $\mathcal{V}_t$ are wrong on a particular observed instance, we ignore this feedback (since such an error is unavoidable). Such variations of 'relaxed versioning' are natural ideas when extending the present problem to the one where the competitor may be allowed to make non-zero mistakes, although its analysis is beyond the scope of this dissertation. The scheme's viablility in this extended setting with only simple modifications indicates the practicality of such strategies.

Below, we take the competitor to be the function that makes the fewest mistakes, denoted as M_T^*. If there is more than one such function, we take the one that makes the fewest abstention to get A_T^*. We measure *excess mistakes* $M_T - M_T^*$ and excess abstentions $A_T - A_T^*$ with respect to this competitor.

Behaviour of regrets with the length of the game Fig. 6·2 presents the excess mistakes as a fraction of T for the two datasets, i.e. ${}^{M_T - M_T^*}\!/_T$, as T, is varied. The learners are all instantiated with the exploration rate $p = {}^1\!/\!\sqrt{T}$. We observe that the excess abstentions are negative (or near-zero) over this range (see Fig. E·1 in §E.5). Therefore we do not plot these below (the orange line is MMEA, see below). We note that the relative mistakes stay below $\sqrt{{}^{2\log N}\!/_T}$, bearing out the theory.

Achievable Operating Points of Mistakes and Abstentions Fig. 6·3 shows the mistake and abstention rates attainable by varying p and ε, while holding T fixed at 500 (which is large enough to show long-run structure, but small enough allow fast experimentation). Concretely, we vary these linearly for 20 values of $p \in [0.015, 0.285]$, and 10 values of $\varepsilon \in [0.001, 0.046]$. The resulting values represent operating points that can be attained by a choice of p, ε. The same plot includes lines that represent the operating points when the scheme is run with $\varepsilon = 0.001$, the smallest value we take. Note that in practice, the best choices of ε, p may be data dependent, and choosing them in an online way is an interesting open problem (also see §E.5.6).

The Price of Being Online We characterise this in two ways beyond the excess mistakes.

- In Fig. 6·2, we also plot the 'mistake-matched excess abstention' (MMEA). This is defined as follows - if the scheme concludes with having made M_T mistakes, we find, in hindsight, the classifier that minimises the number of abstentions, subject to making at most M_T mistakes. The MMEA is the excess abstention of the learner over those of this relaxed competitor, and represents how many fewer abstentions a batch learner would make if allowed to make as many mistakes as the online learner. Notice that this MMEA remains well controlled in Fig. 6·2, and appears to scale as $O(\sqrt{T})$.
- In Fig. 6·3, we also plot the post-hoc operating points of the classifiers in $\mathcal{F}$ as black triangles. This amounts to plotting the optimal abstentions amongst classifiers that make at most m mistakes, varying m.[3] We note that the red operating points of the scheme get close to the black frontier, illustrating that

[3] Observe that the MMEA corresponds to the horizontal distance between a red-point with m mistakes, and the left-most black point with y-coordinate under m.

the inefficiency due to being online is limited. As the time-behaviour of MMEA in Fig.6·2 illustrates, the inefficiency is expected to grow sublinearly with T, and to thus vanish under amortisation.

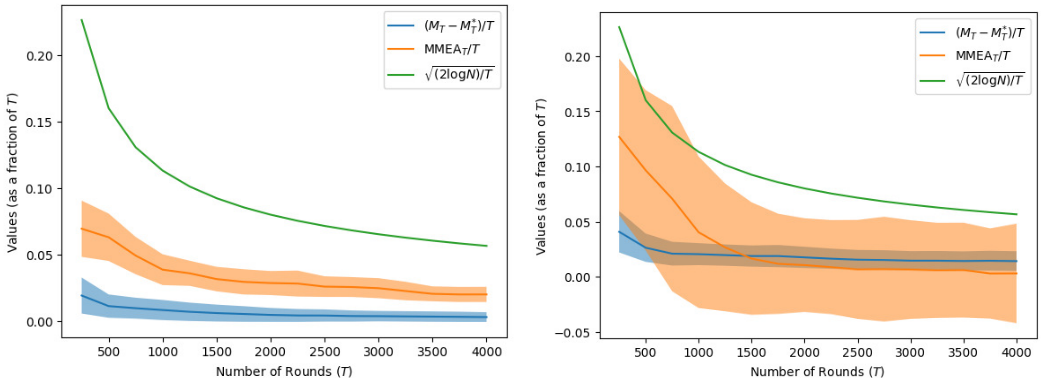


Figure 6·2: $M_T - M_T^*$, and MMEA as fractions of T, as the number of rounds T is varied for CIFAR-10 (left) and GAS (right). The plots are averaged over 100 runs, and one-standard-deviation error regions are drawn.

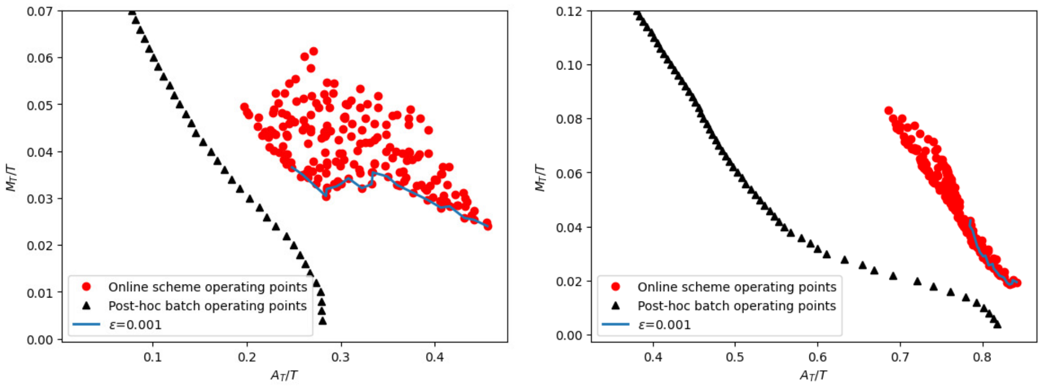


Figure 6·3: Operating points for our scheme as ε and p are varied are represented as red dots (for CIFAR-10 in the left, and GAS in the right). The black triangles represent operating points obtained by batch learning with the benefit of full feedback. The blue lines interpolate points obtained by varying p for $\varepsilon = 0.001$ Points are averaged over 200 runs. Note that the values are raw mistakes and abstentions, and not regrets.

6.7 Discussion

Online selective classification offers a primitive that has relevance to both safety-critical and resource-limited settings. In this chapter, we highlighted the role of long-term abstentions in such situations, and studied this problem under the feedback limitation that labels are only provided when the system abstains, which is the only time high-complexity evaluation would be invoked in a selective classification system. When working with a finite class of model, we identified a simple scheme that provides a tight (in terms of T) trade-off between mistakes and excess abstentions against adaptive adversaries. We further discussed two schemes that improve upon the dependence of the same on the size of the model class - tightly against stochastic adversaries, and at the cost of some rate performance against adaptive adversaries. Together, these schemes and analyses provide some basic foundations for the problem when competing against no-mistake models. Additionally, we carried out empirical studies that validate the scheme in the stochastic case, and demonstrate that with minor modifications, the scheme is resilient to the situation where no selective classifier in the model class is mistake-free. A number of interesting questions remain open, and we discuss a few of these below.

Perhaps the most basic question left open by the above study is how the minimax regrets against adaptive adversaries depend on N. Along with being a basic scientific question, this issue has implications for whether the results can be extended to infinite classes. Indeed, under assumptions of bounded combinatorial dimensions, the VUE-PROD and MIXED-LOSS-PROD schemes can be extended to infinite model classes, but the basic technique to do so yields trivial bounds for VUE due to the linear dependence on N. If this dependence could be improved to logarithmic, the extension to model classes with finite (multiclass versions of) Littlestone dimension would be immediate.

A practically relevant and theoretically interesting direction is online SC but where the competitor can make non-zero mistakes. This can be set up in at least two ways - either an error parameter ε is given to the learner, which must ensure that both notions of regret are small against competitors that make at most εT mistakes; or, no explicit error parameter is specified, and the learner is required to compete against the least mistake-prone model in a given set (similarly to §6.6). Both settings raise new challenges, since one must relax the notion of versioning used in the above work for related schema to be viable. The latter setting raises a further issue of how one can adapt to the mistake rate of the competitor. Also of practical relevance is the case where abstentions are not equally penalised, but have some variable cost. Here too, one can study variants of signalling regarding whether the cost of abstention is available before or only after an abstaining decision is made.

Finally, we observe that while tight, the random exploration technique is somewhat unsatisfying, and practically a context-adapted abstention strategy is likely to offer meaningful advantages over it. In analogy with the exploration in label-efficient prediction, one direction towards exploring context-aware methods is to study more concrete structured situations, such as linear models with noisy feedback that are popular in the investigation of online selective sampling.

Part III

Appendix: Details and Proofs

Appendix A

Appendix to Chapter 2

A.1 Proofs omitted in Section 2.2

A.1.1 Proof of Achievability in Theorem 2.2.1

We will restrict attention to the case $a > b$ below. The $b > a$ case follows identically. Recall the test in this setting:

$$N_a^{x_0}(G) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{bn}{4} + C_1 \max(\sqrt{nb \log(2/\delta)}, \log(2/\delta)),$$

where C_1 is the constant implicit in Lemma A.1.1 below.

Under the null distribution, $N_a^{x_0}(G)$ is distributed as $\mathrm{Bin}(n^2/4, b/n)$, while under the alternate, it is distributed as $\mathrm{Bin}((n-s)^2/4 + s^2/4, b/n) * \mathrm{Bin}(s(n-s)/2, a/n)$. These distributions can be separated by Bernstein concentration bounds [CL06, Ch. 2], as summarised by the following Lemma, which is proved in subsequent sections.

Lemma A.1.1. *There exist constants $C_0, C_1 > 1$ such that, if $nb + s(a-b) > C_0 \log(1/\delta)$, then with probability at least $1 - \delta/2$*

(α) Under H_0: $N_a^{x_0}(G) \le \dfrac{bn}{4} + C_1 \max\left(\sqrt{nb \log(2/\delta)}, \log(2/\delta)\right).$

(β) Under H_1: $N_a^{x_0}(G) \ge \dfrac{bn}{4} + \dfrac{s(a-b)}{4} - C_1\sqrt{(nb + s(a-b)) \log(2/\delta)}.$

As the proof of the above lemma discusses, results of the above type hold in the more generic situation where both the communities and the changes can be unbalanced, so long as each community is of at least linear in n size. This allows one to extend the entirety of this theorem to the setting $n^+n^- = \Omega(n^2)$ on replacing $bn/4$ above with $\mathbb{E}_{\mathrm{Null}}[N_a^{x_0}(G)]$, where n^+ and n^- are the sizes of the two communities, i.e., the number of i such that $x_i = +1$ and $x_i = -1$ respectively.

Since $s|a-b| \geq s\Lambda \geq C\log(2/\delta)$, the lemma above holds in our setting on picking C large enough. (α) in Lemma A.1.1 indicates that the false alarm error of test is $\leq \delta/2$. Further, since $(nb + s(a-b))\log(2/\delta) > \log^2(2/\delta)$, part ($\beta$) shows that missed detection error is $\leq \delta/2$ if

$$\frac{1}{4}s(a-b) > 2C_1\sqrt{(nb+s(a-b))\log(2/\delta)} \iff \frac{(a-b)^2}{nb+s(a-b)} > C\frac{\log(2/\delta)}{s^2}.$$

The argument is concluded by some casework:

(i) If $nb \leq s(a-b)$, then the left hand side of the condition above can be bounded from below by $s(a-b)/2$, and thus $s(a-b) \geq 2C_1\log(2/\delta)$ is sufficient. But $s(a-b) \geq s(a-b)^2/(a+b) = s\Lambda$ is larger than $C\log(1/\delta)$, and choosing C large enough is sufficient.

(ii) On the other hand, if $nb > s(a-b)$, the left hand side is instead lower bounded by $s^2(a-b)^2/2nb \geq s^2\Lambda/2n$, and thus $s^2\Lambda \gtrsim n\log(2/\delta)$ is sufficient to satisfy the same. □

A.1.1.1 Proof of Lemma A.1.1

The proof proceeds by establishing the centres of the statistic $N_a^{x_0}$ under the null and alternate distributions, and then invoking Bernstein-type bounds [CL06, Ch 2] to show the claimed statements separately.

(α) For the null, $N_a^{x_0}(G)$ is distributed as $\mathrm{Bin}(n^2/4, b/n)$. Thus, clearly $\mathbb{E}_{\mathrm{Null}}[N] = bn/4$. Further, by Bernstein's inequality for the upper tail,

$$P_{\mathrm{Null}}(N_a^{x_0}(G) > \mathbb{E}_{\mathrm{Null}}[N_a^{x_0}(G)] + nt) \leq \exp\left(-\frac{n^2/4 \times t^2/2}{n^2/4 \times (b/n) + nt/3}\right)$$
$$\leq \exp\left(-\frac{3}{2}\frac{nt^2}{b+4t}\right) \leq \exp\left(-\frac{3}{8}\frac{nt^2}{t+b}\right).$$

Thus, if

$$\frac{nt^2}{b+t} \geq \frac{8}{3}\log(2/\delta),$$

then this tail has mass at most $\delta/2$. We may now consider the two cases

(i) If $nb \leq 16/3 \log(2/\delta)$, then plugging in $t = 16/3\frac{\log(2/\delta)}{n}$ above yields that the the condition is satisfied, since then

$$\frac{nt^2}{b+t} \geq \frac{nt^2}{2t} = \frac{nt}{2} = \frac{8}{3}\log(2/\delta).$$

(ii) If $nb \geq 16/3\log(2/\delta)$, then setting $t = \sqrt{(16/3)\frac{b}{n}\log 2/\delta}$ we can bound

$$\frac{nt^2}{b+t} = \frac{16/3\log(2/\delta)}{1+\sqrt{(16/3)\log(2/\delta)/nb}} \geq \frac{16/3\log(2/\delta)}{2}.$$

As a consequence, picking $nt = \max(\sqrt{(16/3)nb\log(2/\delta)}, 16/3\log(2/\delta))$ implies that the probability in question is at most $\delta/2$.

We note that this calculation can be made more robust, in that if the communities are unbalanced but linearly sized with n, then the number of edges crossing is $n^+(n-n^+) = \Omega(n^2)$ in the above, and essentially the same goes through with $n^2/4$ replaced by n^2/C for some constant C.

(β) This proof proceeds in much the same way as the above. With the modification that the distribution of $N_a^{x_0}(G)$ is now $\mathrm{Bin}(n^2/4 - s(n-s)/2, b/n) * \mathrm{Bin}(s(n-s)/2, a/n)$, since $2 \times s(n-s)/4$ of the edges are now between nodes of the same communities. The centre of this is easily seen to be $\frac{nb}{4} + \frac{s(n-s)}{2}\frac{a-b}{n}$. Further invoking the Bernstein lower tail, we find that

$$\begin{aligned} P_{\mathrm{Alt}}(N_a^{x_0}(G) \leq \mathbb{E}_{\mathrm{Alt}}[N_a^{x_0}(G)] - nt) &\leq \exp\left(-\frac{1}{2}\frac{n^2t^2}{\frac{s(n-s)}{2}\cdot\frac{a}{n} + \frac{n^2-2s(n-s)}{4}\cdot\frac{b}{n}}\right) \\ &\leq \exp\left(-\frac{n^2t^2}{nb + s(a-b)}\right) \end{aligned}$$

The required claim now follows directly by setting $t = \sqrt{\frac{(nb+s(a-b))\log(2/\delta)}{n}}$.

Again, the above can also be rendered more robust to imbalance. Suppose that the communities and the changes are both imbalanced, and let n^+, n^- be the sizes of the communities in x_0, and s^+, s^- be the number of nodes that are moved from $+$ to $-$ and vice-versa according to the alternate x. Then the number of edges which behave according to a/n in the alternate is $\tau = s^+(n^- - s^-) + s^-(n^+ - s^+)$. But $\tau \leq s(n^+ + n^- - s^+ - s^-) = sn$, so the concentration results go through with a weakening of a factor of 2. Further, assume wlog that $s^+ \geq s^-$. since $s^+ + s^- = s$, and $n^+ + n^- = n$, we have that

$$\tau = s^+(n - s) + (s - 2s^+)(n^+ - s^+).$$

Minimising the above subject to $s^+ \in [s/2 : s]$, we find that the minima can be uniformly lower bounded by $\min(s \min(n^+, n^-), s(n-s)/2)$. So long as each community is of linear size, this is $\Omega(sn)$, and thus the centre of the statistic moves by $\Omega(s(a-b))$ with respect to the null statistic.

Putting the two effects above together, we can write that under the alternate distribution, with probability $\geq 1 - \delta/2$,

$$N_a^{x_0}(G) \geq \mathbb{E}_{\text{Null}}[N_a^{x_0}(G)] + \frac{1}{C_1} s(a-b) - C_2\sqrt{(nb + s(a-b))\log(2/\delta)}.$$

In conjunction with the discussion for unbalanced but linearly sized communities in case (α), the above allows the claims of the achievability part of Theorem 2.2.1 to hold for the case where both communities are of linear size and changes are not constrained to be balanced without any change other than a weakening of the constants implicit in the same. The only modification required for this is to update the tests to threshold at $\mathbb{E}_{\text{Null}}[N_a^{x_0}(G)] +$ (fluctuation term) instead of at $bn/4$ as presented in the main text.

A.1.2 Proofs of converse bounds from Theorem 2.2.1

This section begins with an exposition of Le Cam's method, which is the general proof strategy we employ to show both these converse bounds. This is followed by separate subsections devoted to each converse bound claimed in Theorem 2.2.1.

A.1.2.1 Le Cam's method.

The generic lower bound strategy is constructed by noting that the minimax risk of the goodness-of-fit problem is lower bounded by the risk of the same with any given prior on the alternate communities, i.e. the risk of the problem

$$H_0 : x = x_0 \quad \text{vs} \quad H_1 : x \sim \pi$$

for a π supported on $\{x : d(x, x_0) \geq s\}$ (or some restriction of the same, as in the following sections), and the Bayes risk

$$R_\pi := \inf_{\varphi : G \to \{H_0, H_1\}} P(\varphi = H_1 | x_0) + \sum_{x : d(x, x_0) \geq s} \pi(x) P(\varphi = H_0 | x).$$

By classical Neyman-Pearson theory [see, e.g., LR06], the likelihood ratio test is optimal under the above risk, and

$$\underset{\pi}{R} = 1 - d_{\mathrm{TV}} \left(P(G|x_0), \sum \langle P(G|x) \rangle_\pi \right),$$

where $\langle P(G|x) \rangle_\pi := \sum_x \pi(x) P(G|x)$, and recall the total variation distance

$$d_{\mathrm{TV}}(P, Q) := \frac{1}{2} \|P - Q\|_1.$$

We proceed by bounding d_{TV} by an f-divergence more conducive to tensorisation in order to exploit the (conditional) independence of the edges in an SBM, and then by choosing an appropriate π. The f-divergence inequalities we use are

1. χ^2 bound: Recall that

$$D_{\chi^2}(Q\|P) = \sum_x P(x) \left(\frac{Q(x) - P(x)}{P(x)}\right)^2 = \mathbb{E}_P[L^2(X)] - 1,$$

where $L(x) := Q(x)/P(x)$ is the likelihood ratio. It holds that

$$d_{\mathrm{TV}}(P, Q) \le \sqrt{\frac{1}{2} \log(1 + D_{\chi^2}(Q\|P))},$$

which follows from Pinsker's inequality and the fact that

$$D_{KL}(Q\|P) \le \log(1 + D_{\chi^2}(Q\|P)),$$

which is a consequence of Jensen's inequality applied to the log (or, equivalently, the monotonicity of Rényi divergences).

Invoking the above inequality and Le Cam's method, we find that if for some π, and for $L(G) := \frac{\langle P(G|x)\rangle_\pi}{P(G|x_0)}$, then the following is necessary for the minimax risk of the GOf problem to be bounded by δ :

$$\mathbb{E}_{x_0}[L^2(G)] \ge \exp\left(2(1-\delta)^2\right).$$

For $\delta \le 1/4$, this yields a lower bound of $\mathbb{E}_{x_0}[L^2] > 3.08$.

2. Hellinger bound: The Bhattacharya coefficient of P, Q is defined as

$$\mathrm{BC}(P, Q) := \sum_x \sqrt{P(x)Q(x)},$$

and the Hellinger divergence as

$$d_H(P, Q) := \sqrt{1 - BC(P, Q)} = \frac{1}{\sqrt{2}} \|\sqrt{P} - \sqrt{Q}\|_2.$$

We exploit the classical inequality

$$d_{\mathrm{TV}}(P,Q) \leq \sqrt{d_H^2(P,Q)(2 - d_h^2(P,Q)} = \sqrt{1 - \mathrm{BC}^2(P,Q)},$$

which is a consequence of the Cauchy-Schwarz inequality.

Again plugging this in with $Q = \langle P(G|x) \rangle_\pi$, we find that in order for the risk to be smaller than δ, we must have that

$$\delta \geq 1 - \sqrt{1 - \mathrm{BC}^2} \geq \frac{\mathrm{BC}^2}{2} \implies \mathrm{BC} \leq \sqrt{2\delta},$$

where $\mathrm{BC} = \mathrm{BC}(\langle P(G|x) \rangle_\pi, P(G|x_0))$.

We now proceed to show the claimed bounds. Recall that we are required to show that if $R_{\mathrm{GoF}} < \delta \leq 1/4$, them

$$\Lambda \gtrsim \log(1 + n/s^2) \tag{A.1}$$

$$s\Lambda \gtrsim \log(1/\delta). \tag{A.2}$$

A.1.2.2 Proof of (A.1)

For convenience, we let

$$\nu := (a-b)^2 \left(\frac{1}{a(1-a/n)} + \frac{1}{b(1-b/n)} \right). \tag{A.3}$$

Since $a, b \leq n/2$, and since $a/b = \Theta(1)$, we have $\Lambda \asymp \nu$, and it suffices to show the same bound on the latter.

We invoke Le Cam's method with a χ^2-bound. Let $m := n/2, t := s/2$ and let x_0 be the partition $([1:m], [m+1:2m])$.

The alternate prior is chosen to be the uniform prior on the set of alternate partitions constructed as follows. For each $T \subset [1:m]$, we define the partition

$$\begin{aligned} y_T(+) &= [1:m] \cup (m+T) \sim T \\ y_T(-) &= [m+1:2m] \cup T \sim (m+T), \end{aligned}$$

where $(m+T) = \{i+m : i \in T\}$. Let $\mathcal{Y}_t := \{y_T : T \subset [1:m], |T| = t\}$. For conciseness, we define the measures on $\mathcal{G}$:

$$P_{y_T}(\cdot) := P_T(\cdot) := P(\cdot \mid y_T),$$

and set $P_0 = P(\cdot \mid x_0)$. Further, for convenience, we set $p = a/n$ and $q = b/n$.

For a graph G, we find that $L(G) := \frac{1}{|\mathcal{Y}_t|} \sum_{x \in \mathcal{Y}_t} \frac{P_x(G)}{P_0(G)}$. To invoke Le Cam's method (§A.1.2.1), we need to upper bound $\mathbb{E}_{P_0}[L^2(G)]$.

To this end, we will define for an edge $e = (u, v)$, and a graph G (which is implicit in the notation)

$$f_e(q,p) := (q/p)^e((1-q)/(1-p))^{1-e}. \tag{A.4}$$

Above, $f_e(q,p)$ arises as a ratio of the probabilities of a Bern(q) and a Bern(p) random variable. Thus, it is the likelihood ratio of an edge being between nodes in the different and in the same community.

First observe that

$$\begin{aligned} \frac{P_T}{P_0} = &\left(\prod_{\substack{i \in [1:m] \sim T, \\ j \in m+T}} f_{ij}(p,q) \right) \left(\prod_{\substack{i \in [m+1:2m] \sim m+T, \\ j \in T}} f_{ij}(p,q) \right) \\ &\times \left(\prod_{\substack{i \in [1:m] \sim T, \\ j \in T}} f_{ij}(q,p) \right) \left(\prod_{\substack{i \in [m+1:2m] \sim m+T, \\ j \in m+T}} f_{ij}(q,p) \right). \end{aligned} \tag{A.5}$$

An important feature of the setup above is that every term in the above product is independently distributed, and wherever $f_{ij}(p,q)$ appears, the corresponding e_{ij} is

Bern(q), and similarly with $f_{ij}(q,p)$ and Bern(p).

Note that

$$\mathbb{E}_{P_0}[L^2(G)] = \sum_{T_1,T_2 \subseteq [1:m] \text{ of size } t} \mathbb{E}_{P_0}\left[\frac{P_{T_1}(G)P_{T_2}(G)}{P_0^2(G)}\right],$$

and so we must control expectations of this form in order to apply Le Cam's method. Let us fix T_1 and T_2 for now, and partition the nodes into groups as described by the Figure A·1[1].

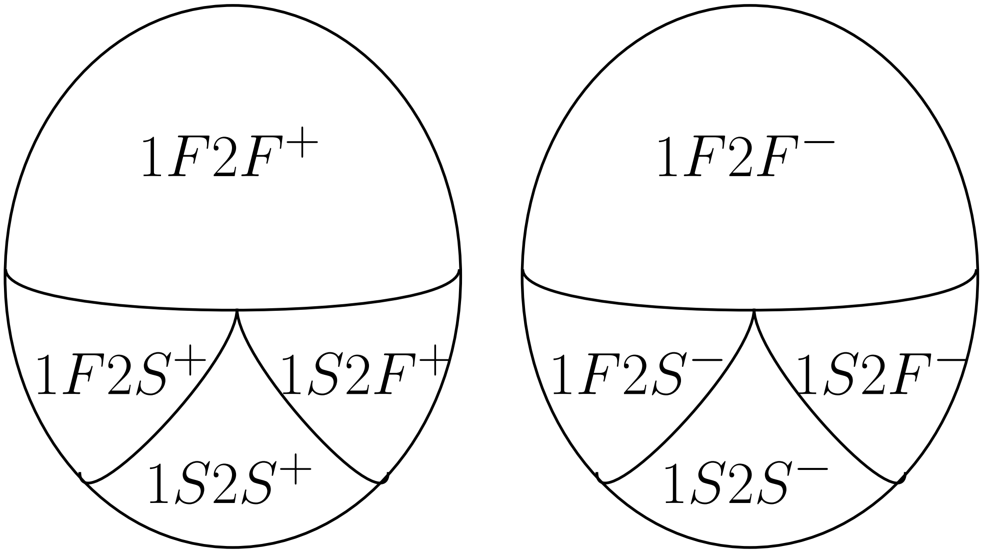


Figure A·1: A schematic of the nodes, partitioned according to their labellings in x_0, y_{T_1}, y_{T_2}. The two ovals denote the partition induced by x_0 into groups marked $+$ and $-$. The section $1F2F^+$ denotes the nodes in the $+$ group whose labels remain *fixed* to $+$ in both y_{T_1}, y_{T_2}. The section marked $1S2F^+$ denotes the nodes in the $+$ group whose labels are *switched* to $-$ in y_{T_1} but remain *fixed* to $+$ in y_{T_2}. Other labels are analogously defined.

Note that in the figure, $1F2F^+ = [1:m] \sim (T_1 \cup T_2), 1S2S^- = (m+T_1) \cap (m+t_2)$ and so on. Also, importantly, the size of groups with the same number of Ss and Fs in

[1] The argument, while simple, gets a little notationally hairy at this point. We recommend that the reader consults Figure A·1 frequently, preferably a printed copy that allows one to sketch the various types of connections on it.

the above representation is identical (i.e., $|1F2S^+| = |1F2S^-| = |1S2F^+| = |1S2F^-|$ and so on.)

We consider how the terms relating to the edge (u, v) for any $u, v \in [1 : 2m]$ appear in the product $\frac{P_{T_1} P_{T_2}}{P_0^2}$. Below,

- Clearly, if u and v are both in the same group in both settings, the behaviour of the edge (u, v) under the alternate distributions and the null distribution is identical, and these terms will not appear in the product.

- If both $(u, v) \in 1F2F^+ \times 1F2F^- \cup 1S2S^+ \times 1S2S^-$, then again, the edge (u, v) has identical distribution under both alternates and the null, and these terms do not appear in the product.

- If $(u, v) \in 1F2F^+ \times 1F2S^+$, then the (u, v) term does not appear in P_{T_1}/P_0, but appears once in P_{T_2}/P_0. Since likelihoods must average to 1, and since the distributions of the edges are independent, any term which appears just once is averaged out when we take expectations with respect to P_0. Thus, even though these terms appear in the product, we may ignore them due to our eventual use of the expectation operator. A quick check will show that the same effect happens for $(u, v) \in \Gamma_1 \times \Gamma_2$, where Γ_1 can be obtained by inverting one instance of an F to a S or vice versa, and possibly changing the sign (e.g. $1F2S^- \times 1S2S^+$.) Thus, all such pairs can be safely ignored.

- This leaves us with edges of the form $\{1F2F^\pm \times 1S2S^\pm\} \cup \{1F2S^\pm \times 1S2F^\pm\}$. In these cases, if the signs of the two choices match - i.e.

$$(u, v) \in \Gamma^+ \times \widetilde{\Gamma}^+ \text{ for } (\Gamma, \widetilde{\Gamma}) \in \{(1F2F, 1S2S), (1S2F, 1F2S)\},$$

  then we will obtain a contribution of $f_{uv}(q, p)^2$ to the product. On the other hand, if they differ, then we will obtain a contribution of $f_{uv}(p, q)^2$

Accounting for the above, and taking expectation, we have that

$$\mathbb{E}\left[\frac{P_{T_1}P_{T_2}}{P_0^2}\right] = (\Psi)^{|1F2F^+|\cdot|1S2S^+|+|1F2F^-|\cdot|1S2S^-|+|1S2F^+|\cdot|1F1S^-|+|1F2S^+|\cdot|1S2F^-|}, \tag{A.6}$$

where

$$\Psi := \mathbb{E}_{e\,\sim\,\mathrm{Bern}(p)}[f_e(q,p)^2]\mathbb{E}_{e\,\sim\,\mathrm{Bern}(q)}[f_e(p,q)^2] \tag{A.7}$$

Further, since in our choice of the alternate communities the groups with the same number of Ss and Fs have identical size, and thus we may rewrite (A.6) above as

$$\mathbb{E}\left[\frac{P_{T_1}P_{T_2}}{P_0^2}\right] = \Psi^{2(|1F2F^+||1S2S^+|+|1S2F^+|^2)}.$$

For convenience, let $|1S2S^+| = |T_1 \cap T_2| = k$. We then have that $|1S2F^+| = t - k$ and $|1F2F^+| = m + k - 2t$.

We thus have that

$$\begin{aligned}
\mathbb{E}_{P_0}\frac{P_{T_1}P_{T_2}}{P_0^2} &= \exp\left((\log\Psi)(2k(m+k-2t)+2(t-k)^2)\right) && \text{(A.8)}\\
&= \exp\left((\log\Psi)(2mk+2k^2-4kt+2k^2+2t^2-4kt)\right) && \text{(A.9)}\\
&= \exp\left((\log\Psi)(2mk+4k^2-8kt+2t^2)\right) && \text{(A.10)}\\
&\le \exp\left((\log\Psi)((2m-4t)k+2t^2)\right), && \text{(A.11)}
\end{aligned}$$

where we have used that $k \le t$.

Now, for $(p,q) = (a/n, b/n)$,

$$\begin{aligned}
\Psi &= \left(\frac{q^2}{p}+\frac{(1-q)^2}{(1-p)}\right)\left(\frac{p^2}{q}+\frac{(1-p)^2}{(1-q)}\right) && \text{(A.12)}\\
&= \left(1+\frac{(p-q)^2}{p(1-p)}\right)\left(1+\frac{(p-q)^2}{q(1-q)}\right) && \text{(A.13)}\\
&= \left(1+\frac{(a-b)^2}{na(1-a/n)}\right)\left(1+\frac{(a-b)^2}{nb(1-b/n)}\right) && \text{(A.14)}\\
&= 1+\frac{\nu}{n}+O(n^{-2}) \le 1+2\frac{\nu}{n}. && \text{(A.15)}
\end{aligned}$$

As a consequence, using $2m = n$, and the development above,

$$\mathbb{E}_{P_0} \frac{P_{T_1} P_{T_2}}{P_0^2} \le \exp\left(\frac{4t^2}{n}\nu\right) \exp\left(2k\nu(1 - 4t/n)\right). \tag{A.16}$$

The above is insular to the precise identities of T_1, T_2. Further, for a given T_1, the number of partitions T_2 such that $|T_1 \cap T_2| = t$ is $\binom{t}{k}\binom{m-t}{t-k}$. Feeding this into the expression for $\mathbb{E}[L^2(G)]$ and some simple manipulations yield that

$$\mathbb{E}_{P_0}[L^2(G)] \le \frac{e^{\frac{4t^2}{n}\nu}}{\binom{m}{t}} \sum_{k=0}^{t} \binom{t}{k}\binom{m-t}{t-k} \exp\left(2k\nu(1 - 4t/n)\right), \tag{A.17}$$

where we remind the reader that $t = s/2, m = n/2$.

Recall from §A.1.2.1 that if $\mathbb{E}_{P_0}[L^2] < 3$, then the risk exceeds 0.25. Thus, we will aim to upper bound (A.17) by 3.

We begin by rewriting

$$\mathbb{E}_{P_0}[L^2(G)] \le \frac{e^{\frac{4t^2}{n}\nu}}{\binom{m}{t}} \sum_{k=0}^{t} \binom{t}{k}\binom{m-t}{t-k} \exp\left(2k\nu(1 - 4t/n)\right), \tag{A.18}$$

$$= e^{\frac{4t^2}{n}\nu} \mathbb{E}[\xi^Z], \tag{A.19}$$

where $\xi := \exp\left(2\nu(1 - 4t/n)\right) > 1$ and $Z = \sum_{i=1}^t Z_i$, where Z_i are sampled without replacement from the collection of t $(+1)$s and $m - t$ (0)s. Note that $z \mapsto \xi^z$ is continuous and convex for $\xi \ge 1$. By Theorem 4 of [Hoe63],

$$\mathbb{E}[\xi^Z] \le \mathbb{E}[\xi^{\widetilde{Z}}],$$

for $\widetilde{Z} = \sum_{i=1}^t \widetilde{Z}_i$, where $\widetilde{Z}_i$ are drawn by sampling with replacement from the same collection. But $\widetilde{Z}$ is just a Binomial random variable with parameters $(t, t/m)$. Thus, we have that

$$\begin{aligned}\mathbb{E}_{P_0}[L^2(G)] &\le e^{\frac{2t^2}{m}\nu}\left(1+\frac{t}{m}\left(\exp\left(2\nu(1-2t/m)\right)-1\right)\right)^t && \text{(A.20)}\\ &\le \exp\left(2\frac{t^2}{m}\nu+\frac{t^2}{m}\left(\exp\left(2\nu(1-2t/m)\right)-1\right)\right) && \text{(A.21)}\\ &\le \exp\left(\frac{t^2}{m}\left(2\nu+e^{2\nu}-1\right)\right) && \text{(A.22)}\\ &\le \exp\left(2\frac{t^2}{m}\left(e^{2\nu}-1\right)\right), && \text{(A.23)}\end{aligned}$$

where the final inequality uses $u < e^u - 1$. Using the above, and noting that $m/2t^2 = n/s^2$, we find that

$$\nu \le \frac{1}{2}\log\left(1+\frac{\log(3)n}{s^2}\right) \implies \mathbb{E}_{P_0}[L^2(G)] \le 3,$$

finishing the argument. □

A.1.2.3 Proof of the converse bound A.2

Recall that this part of the theorem claims that if $R_{\text{Gof}} \le \delta \le 1/4$, then $s\Lambda \ge C\log(1/\delta)$.

We will again use Le Cam's method (§A.1.2.1), this time controlling the total variation distance by a Hellinger bound.

Let $x_0 = ([1:n/2],[n/2+1:n])$ be the null partition, and $\mathcal{Y} := \{y\}$, with $y := ([1:n/2-s/2]\cup[n/2+1:n/2+s/2],[n/2-s/2+1:n/2]\cup[n/2+s/2+1:n])$. We let $P_{x_0}(G) := P(G|x_0)$, and similarly P_y. Recall from the section on Le Cam's method that the following is a necessary condition for the risk to be smaller than δ

$$\text{BC}(P_{x_0},P_y) \le \sqrt{2\delta}.$$

The Bhattacharya Coefficient can be estimated directly in this setting. (We omit the derivation below)

$$\mathrm{BC}(P_y, P_{x_0}) = \left(\sqrt{\frac{ab}{n^2}} + \sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}\right)^{s(n-s)} \tag{A.24}$$

For $u, v < 3/4$,

$$\sqrt{(1-u)(1-v)} \geq 1-(u+v)/2-2(u-v)^2.$$

Thus

$$\begin{aligned}
\mathrm{BC}(P_y, P_{x_0}) &\geq \left(1-\frac{a+b}{2n}+\frac{ab}{n}-2\frac{(a-b)^2}{n^2}\right)^{s(n-s)} && \text{(A.25)}\\
&= \left(1-\frac{(\sqrt{a}-\sqrt{b})^2}{2n}-2\frac{(a-b)^2}{n^2}\right)^{s(n-s)} && \text{(A.26)}\\
&\geq \left(1-\frac{(\sqrt{a}-\sqrt{b})^2}{n}\right)^{s(n-s)} && \text{(A.27)}\\
&\geq \exp\left(-2s(\sqrt{a}-\sqrt{b})^2\right), && \text{(A.28)}
\end{aligned}$$

where the third inequality uses $(a+b) < n/4$, and the final uses used $1-u \geq e^{-2u}$ for $0 < u \leq 0.75$—which applies since $0 < (\sqrt{a}-\sqrt{b})^2 < \max(a,b) < n/4$—and $n-s \leq n$.

Now note that

$$(\sqrt{a}-\sqrt{b})^2 = \frac{(a-b)^2}{(\sqrt{a}+\sqrt{b})^2} \leq \frac{(a-b)^2}{a+b} = \Lambda,$$

and thus,

$$\mathrm{BC}(P_y, P_{x_0}) \geq \exp\left(-2s\Lambda\right).$$

Invoking the condition for $R_{\text{GoF}} \le \delta$ above, we have

$$\begin{aligned} &\exp\left(-2s\Lambda\right) \le \sqrt{2\delta} \\ \iff\ & s\Lambda \ge \frac{1}{4}\log\frac{1}{2\delta}. \end{aligned}$$

For $\delta \le 1/4$, we may further lower bound the above by $(\log(1/\delta))/8$. □

A.1.3 A comment on the role of Λ when $a/b \neq \Theta(1)$

The main text concentrates on the setting where a/b is a constant. Here, we briefly comment on the setting where the ratio $\rho := \frac{\max(a,b)}{\min(a,b)}$ is *diverging* with n. In the setting of balanced communities and divergent ρ, the behaviour of the goodness-of-fit problem is no longer described by the quantity $\Lambda = \frac{(a-b)^2}{a+b}$, but instead depends on

$$\mu := \frac{(a-b)^2}{\min(a,b)}.$$

Specifically, our proofs can, with minimal changes, be adapted to say that for balanced GoF, R_{GoF} can be solved with vanishing risk if the following hold:

$$\begin{aligned} s\Lambda &= \omega(1) \\ \mu &= \omega(n/s^2), \end{aligned}$$

and further, to attain the same, it is necessary to have

$$\begin{aligned} s\Lambda &= \omega(1) \\ \mu &\gtrsim \log(1 + n/s^2). \end{aligned}$$

Indeed, for the lower bounds, $\mu \le \nu \le 4\mu$ uniformly, where ν is the SNR quantity in the previous section, and the upper bounds naturally feature μ.

Together, the above offer a tight characterisation of the GoF problem in the setting of balanced communities and *large* s. Note that ${}^{\mu}/_{\Lambda} = 1 + \rho$ diverges with ρ, and thus the above indicate that GoF testing becomes much easier as this ratio blows up - something to be expected.

Despite the above developments, we concentrated on the setting $\rho = \Theta(1)$ in the main text. This is largely because the majority of the literature on the SBM focuses on this regime, as this is the hardest setting for inference about the planted structure. Thus, in order to compare to existing work, we examined the $a \asymp b$ setting.

As an aside, we note that unlike the above GoF results, the TST results do not alter in the setting of divergent ρ. Theorem 2.3.1, and in particular the converse bound $\Lambda \gtrsim 1$, continues to hold for this setting.

On the whole, this line of work is still under investigation, particularly whether the behaviour of GoF for large ρ continues to be driven by μ in the setting of small changes. We plan to explore this question in later work.

A.2 Proofs omitted from section 2.3

A.2.1 Proof of Achievability in Theorem 2.3.1

Recall that the scheme in Algorithm 1 utilises a partial recovery routine. For the purposes of the following argument, we invoke the method of [CRV15], which provides a procedure that, under the conditions of the theorem, that attains with probability at least $1 - 1/n$ recovery with at most $\varepsilon_{\max} n$ errors, where $\varepsilon_{\max} = \min(1/2, 2e^{-C\Lambda})$ for an explicit constant C. We choose Λ large enough so that $\varepsilon_{\max}$ is bounded strictly below $1/2$ - for convenience, say by $1/3$.

Let $G' \sim P(\cdot|x)$ be an independent copy of G, useful in the analysis, and recall the definition of $\widetilde{G}, G_1$ from Algorithm 1. We define the following events that we will condition on in the sequel:

$$\mathcal{E}(G_1) = \{\text{Number of edges in } G_1 \ \le an/2\} \qquad \mathcal{E}(\hat{x}) = \{d(\hat{x}, x) \le \varepsilon_{\max} n\}$$

For succinctness, we let $\mathcal{E} := \mathcal{E}(G_1) \cap \mathcal{E}(\hat{x})$. The analysis proceeds in four steps:

(L1) **Lemma A.2.1.** $P(\mathcal{E}) \ge 1 - 4/3n$.

(L2) **Lemma A.2.2.** $\left|\mathbb{E}[2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G') \mid \mathcal{E}]\right| \le a^2.$

(L3) **Lemma A.2.3.** *If $d(x, y) \ge s$, then for $\kappa := (1 - 2\varepsilon_{\max})^2 - 1/(n-1)$,*

$$\mathbb{E}[T^{\hat{x}}(G') - T^{\hat{x}}(H) \mid \mathcal{E}] \ge \kappa \frac{(a-b)}{n}(n-s)s.$$

(L4) **Lemma A.2.4.** *Let $\xi := a^2 + 5\sqrt{2na\log(6n)}$. Then*

$$\begin{aligned} &P_{\text{Null}}\left(T \ge \xi | \mathcal{E}\right) \le 2/3n \\ &P_{\text{Alt.}}\left(T \le \kappa(a-b)s/2 - \xi | \mathcal{E}\right) \le 4/3n \end{aligned}$$

We briefly describe the functional roles of the above, and relegate their proofs to the following sections.

(L1) allows us to make use of the typicality of G_1 and the recovery guarantees of $\hat{x}$. The former is primarily useful for (L2), while the latter induces (L3).

(L2) lets us avoid the technical issues arising from the fact $\widetilde{G}$ and $G_1, \hat{x}$ are correlated, and allows us to work with the simpler G'. It also shows that under the null, the mean of T is small. This lemma is likely loose, and introduces the nuisance condition $a \le n^{1/3}$.

(L3) shows that under the alternate, the centre of T linearly grows with s despite the weak recovery procedure's errors.

(L4) serves to control the fluctuations in T. The $\sqrt{n}$-level term arises from the randomness in $\widetilde{G}, H, G'$, and the a^2 term from our use of G' and (L2).

Putting the above together, we find that the risk is bounded by $4/3n + 2/3n + 4/3n \le 4/n$ if

$$\kappa(a-b)s \ge 4(a^2 + 5\sqrt{2na\log(6n)}).$$

Since $a^2 = a^{3/2}\sqrt{a} \le \sqrt{na}$, and for Λ a large enough constant, $\varepsilon_{\max} \le 1/3 \implies \kappa \ge (1/3 - 1/(n-1))^2 \ge 1/36$ for $n \ge 7$, the above condition is equivalent to

$$(a-b)s \ge C'\sqrt{na\log(6n)}$$

for a large enough C'. Rearranging and squaring, this is equivalent to

$$\frac{(a-b)^2}{a} \gtrsim \frac{n\log(6n)}{s^2}.$$

For $s \geq n^{1/2+c}$ as in the statement, the quantity on the right hand side is decaying with n. Further, Λ is smaller than the left hand side, so it being bigger than a constant forces the above to hold.

Note that the threshold in Algorithm 1 alters the fluctuation range above from $\sqrt{na}$ to $\sqrt{n(a+b)}$. The reason for this is that this relaxation allows Algorithm 1 to be agnostic to the knowledge of (a, b) - generic spectral clustering schemes do not require this knowledge, and the threshold of our scheme depends only on $n(a+b)$, which can be robustly estimated in our setting since the number of edges in the graph is proportional to this. In addition, invoking the bounds of [CRV15] allows explicit control on κ above, and thus provides an explicit value of the constant C in Algorithm 1. □

A.2.1.1 Proof of Lemma A.2.1

We first note that by the work of [CRV15], or [FC19], under the conditions of the theorem, $\mathcal{E}(\hat{x})$ holds with probability at least $1 - 1/n$. By a union bound, it suffices to show that $P(\mathcal{E}(G_1)) \geq 1 - \frac{1}{3n}$. Recall that

$$P((e, v) \in G_1 | x) = \begin{cases} \frac{a}{2n} & x_u = x_v \\ \frac{b}{2n} & x_u \neq x_v, \end{cases} \tag{A.29}$$

and that edges are independent. Thus the number of edges in G_1 is a sum of Bernoulli random variables of parameter $\leq a/2n$. The factor of 2 arises since G_1 is sub-sampled at rate 1/2. Let $\#G_1$ be the number of edges in G_1. We have

$$\mathbb{E}[\#G_1] \leq \binom{n}{2} \frac{a}{2n} \leq \frac{na}{4} \tag{A.30}$$

$$P(\#G_1 \geq \mathbb{E}[\#G_1] + \sqrt{na \log(3n)}) \leq 1/3n, \tag{A.31}$$

where the first bound follows from inspection, and the second follows from the Bernstein upper tail bound of [CL06, Ch. 2] and the condition $a \geq 16 \log(6n)/n$. Further invoking this condition we find that $\sqrt{na \log(3n)} \leq na/4$, and thus

$$P(\mathcal{E}(G_1)) = P(\#G_1 \leq na/2) \geq 1 - \frac{1}{3n}.$$

□

A.2.1.2 Proof of Lemma A.2.2

Let
$$c_{uv} := \frac{(a+b)+(a-b)x_u x_v}{2} \le a.$$
Recall that c_{uv}/n is the probability under x of the edge (u,v) existing.

Also note that for a graph Γ and a partition z,
$$T^z(\Gamma) = \sum_{1\le u<v\le n} z_u z_v \Gamma_{uv},$$
where $\Gamma_{uv} := \mathbf{1}\{(u,v)\in\Gamma\}$.

We're interested in controlling
$$T = T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G') = \sum \hat{x}_u\hat{x}_v(2\widetilde{G}_{uv} - G'_{uv}).$$
Since $\hat{x}$ is a deterministic function of G_1, $\widetilde{G}$ is independent of $\hat{x}$ given G_1. Further, G' is independent of $(G_1, \widetilde{G})$. Lastly observe that
$$P((u,v)\in\widetilde{G} \mid G_1) = \frac{c_{uv}/2n}{1-c_{uv}/2n}(1-(G_1)_{uv}).$$

As a consequence,
$$\begin{aligned}
\mathbb{E}[T \mid G_1] &= \sum \hat{x}_u\hat{x}_v\left(2\cdot\frac{c_{uv}/2n}{1-c_{uv}/2n}(1-(G_1)_{uv}) - \frac{c_{uv}}{n}\right) && \text{(A.32)}\\
&= \sum \hat{x}_u\hat{x}_v\frac{c_{uv}^2/2n^2}{1-c_{uv}/2n} - \sum \hat{x}_u\hat{x}_v\frac{c_{uv}/n}{1-c_{uv}/2n}(G_1)_{uv} && \text{(A.33)}\\
\implies |\mathbb{E}[T \mid G_1]| &\le \sum_{u<v}\frac{c_{uv}^2/2n^2}{1-c_{uv}/2n} + \sum_{u<v}\frac{c_{uv}/n}{1-c_{uv}/2n}(G_1)_{uv} && \text{(A.34)}\\
&\le \frac{a^2/2n^2}{1-a/2n}\binom{n}{2} + \frac{a/n}{1-a/2n}\#G_1. && \text{(A.35)}
\end{aligned}$$

where recall that $\#G_1$ is the number of edges in G_1. Note that we may condition on $\mathcal{E}$, the occurrence of which is a deterministic function of G_1. Since under $\mathcal{E}$ we have $\#G_1 \le an/2$, we find that
$$|\mathbb{E}[T \mid G_1, \mathcal{E}]| \le \frac{1}{1-a/2n}\left(\frac{a^2}{2n^2}\frac{n^2}{2} + \frac{a}{n}\frac{an}{2}\right) \le a^2, \tag{A.36}$$

where the final inequality uses that $1/(1-a/2n) \le 4/3$, which follows from $a \le (n/2)^{1/3}$, and $n \ge 2$.

Finally observe that the right hand side of the equation above does not depend on G_1. Thus, we may integrate over $P(G_1 \mid \mathcal{E})$ to find that

$$|\mathbb{E}[T \mid \mathcal{E}]| \le a^2.$$

Remark : This lemma is likely rather weak. In particular, the upper bound on $|\mathbb{E}[T|G_1]|$ completely ignores the relationship between $\hat{x}$ & G_1, and that between G_1 & c_{uv}. Indeed, (A.32) may also be rewritten as

$$\mathbb{E}[T \mid G_1] = \sum \frac{c_{uv}/n}{1 - c_{uv}/2n} \hat{x}_u \hat{x}_v \left(\frac{c_{uv}}{2n} - (G_1)_{uv} \right).$$

Since $(G_1)_{uv} \sim \text{Bern}(c_{uv}/2n)$, and $\hat{x}$ is a clustering derived from G_1, it may be possible to control the above to something much smaller than a^2. This may require nontrivial use of the $\mathcal{E}(\hat{x})$ conditioning here, which is unused in the above argument. Unfortunately it seems that such control would closely depend on the scheme used to obtain $\hat{x}$, which tend to be complex - most schemes involve non-trivial regularisation of G_1, as well as some amount of quantisation of the solution to an optimisation problem to produce $\hat{x}$, due to which the covariance of G_1 and $\hat{x}$ is difficult to understand. For completeness' sake we point out that an upper bound on the same of $O(a^2/n)$ would remove the nuisance condition of $a \le n^{1/3}$ present in Theorem 2.3.1. □

A.2.1.3 Proof of Lemma A.2.3

We proceed by first developing some intuition behind the proof of Lemma A.2.3 instead of launching straight into the same. Further, we assume throughout that $d(x, y) \ge s$.

Let

$$\begin{aligned} \text{Incorrect} &:= \{u \in [1:n] : x(u) \neq \hat{x}(u)\} \\ \text{Unchanged} &:= \{u \in [1:n] : x(u) = y(u)\}. \end{aligned}$$

and the sets 'Correct' and 'Changed' be their respective complements. We show in Appendix A.2.1.4 the following lemma

Lemma A.2.5.

$$\mathbb{E}[T^{\hat{x}}(G') - T^{\hat{x}}(H) \mid \hat{x}] = \frac{(a-b)}{n}\Big(n(\text{Unchanged}) - 2n(\text{Incorrect, Unchanged})\Big) \times \Big(n(\text{Changed}) - 2n(\text{Incorrect, Changed})\Big), \tag{A.37}$$

where

$$n(\text{Unchanged}) = |\text{Unchanged}|$$
$$n(\text{Incorrect, Unchanged}) = |\text{Incorrect} \cap \text{Unchanged}|,$$

and the other terms are defined analogously.

Suppose $n(\text{Incorrect}) = k$. Due to the exchangability of the nodes when $|\{u : x(u) = +\}| = |\{u : x(u) = -\}|$, the incorrectly labelled nodes in $\hat{x}$ correspond to a choice of $k \in [0 : n/2]$ nodes picked without replacement from $[1 : n]$ uniformly at random. Further, since the changes made in y are chosen independently of the graphs, they are independent of $\hat{x}$. Thus, the number of correct and incorrect nodes changed forms hypergeometric distribution. The expected number of Incorrect nodes changed is precisely $\frac{s}{n} \cdot k$, where s is the number of changes made, and similarly for Incorrect nodes unchanged.

Further invoking the results of [FC19], if $\Lambda \geq C \log(1/\varepsilon_{\max})$, then $k \leq \varepsilon_{max} n$ with probability at least $1 - 1/n$. As a consequence, the bound in Lemma A.2.5 remains large in magnitude even on integrating over the randomness in $\hat{x}$. This was the subject of Lemma A.2.3 from the text, reproduced below for convenience.

Lemma A.2.3

$$\mathbb{E}[T^{\hat{x}}(G') - T^{\hat{x}}(H) \mid \mathcal{E}] \geq \left((1 - 2\varepsilon_{\max})^2 - \frac{1}{n-1}\right)\frac{(a-b)}{n}(n-s)s,$$

the proof of which is the subject of Appendix A.2.1.5. □

A.2.1.4 Proof of Lemma A.2.5

We will require explicit counting of a number of groups of nodes. Let us first define them:

Let

$$\begin{aligned}
S^{++} &:= \{u \ \in [1:n] : \hat{x}(u) = +1,\, x(u) = +1\}, \quad n^{++} := |S^{++}|,\\
S^{+-} &:= \{u \ \in [1:n] : \hat{x}(u) = +1,\, x(u) = -1\}, \quad n^{+-} := |S^{+-}|,\\
S^{--} &:= \{u \ \in [1:n] : \hat{x}(u) = -1,\, x(u) = -1\}, \quad n^{--} := |S^{--}|,\\
S^{-+} &:= \{u \ \in [1:n] : \hat{x}(u) = -1,\, x(u) = +1\}, \quad n^{-+} := |S^{-+}|.
\end{aligned}$$

The sets above encode the partitions induced by $\hat{x}$ and x, with the first symbol in the superscript denoting the label given by $\hat{x}$. Observe that S^{+-}, S^{-+} are the sets of nodes mislabelled in $\hat{x}$.

Lastly, for $(\mathfrak{i}, \mathfrak{j}) \in \{+,-\}^2$, let

$$\begin{aligned}
C^{\mathfrak{i},\mathfrak{j}} &:= S^{\mathfrak{i},\mathfrak{j}} \cap \{u \in [1:n] : x(u) \neq y(u)\}\\
n_C^{\mathfrak{i},\mathfrak{j}} &:= |C^{\mathfrak{i},\mathfrak{j}}|
\end{aligned}$$

These are the nodes that change their labels in y. Note that the values of each of the above objects is a function of $\hat{x}$. For now we will fix $\hat{x}$, and compute expectations over the randomness in G', H alone.

We first study N_w: $N_w^{\hat{x}}(G) = N_w^{\hat{x}}(G[+]) + N_w^{\hat{x}}(G[-])$, where $G[+]$ is the induced subgraph on the nodes $\{u \in [1:n] : \hat{x}(u) = +\}$ and similarly $G[-]$.

By simple counting arguments,

$$\mathbb{E}[N_w^{\hat{x}}(G'[+]) \mid \hat{x}] = \binom{n^{++} + n^{+-}}{2} \frac{a}{n} - \frac{(a-b)}{n} n^{++} n^{+-}. \tag{A.38}$$

Under H, the nodes in C^{++} behave as if they were in S^{+-} and those in C^{+-} as if they were in S^{++}. Computations analogous to before lead to

$$\mathbb{E}[N_w^{\hat{x}}(G'[+]) - N_w^{\hat{x}}(H[+]) \mid \hat{x}] = \frac{a-b}{n} \left((n^{++} - n_C^{++}) - (n^{+-} - n_C^{+-}) \right) (n_C^{++} - n_C^{+-}) \tag{A.39}$$

By symmetry, we can obtain the above for $G[-]$s by toggling the group labels above. Thus, conditioned on a fixed $\hat{x}$, we have

$$\begin{aligned}\mathbb{E}[N_w^{\hat{x}}(G') - N_w^{\hat{x}}(H) \mid \hat{x}] =& \frac{(a-b)}{n}\Big(\big((n^{++} - n_C^{++}) - (n^{+-} - n_C^{+-})\big)\,(n_C^{++} - n_C^{+-}) \\ & + \big((n^{--} - n_C^{--}) - (n^{-+} - n_C^{-+})\big)\,(n_C^{--} - n_C^{-+})\Big). \end{aligned} \tag{A.40}$$

Similar calculations can be performed for N_a. Since in edges across the true partitions, the edges in the same group appear with probability a/n and in different groups with b/n, the roles of a and b will be exchanged in this case, leading to a factor of $+(a-b)$ instead of $-(a-b)$. We will suppress the tedious computations, and simply state that

$$\begin{aligned}\mathbb{E}[N_a^{\hat{x}}(G') - N_a^{\hat{x}}(H) \mid \hat{x}] =& \frac{(a-b)}{n}\Big(\big((n^{++} - n_C^{++}) - (n^{+-} - n_C^{+-})\big)\,(n_C^{--} - n_C^{-+}) \\ & + \big((n^{--} - n_C^{--}) - (n^{-+} - n_C^{-+})\big)\,(n_C^{++} - n_C^{+-})\Big). \end{aligned} \tag{A.41}$$

For convenience, we define

$$\begin{aligned} n(\text{Correct, Unchanged}) &:= (n^{++} + n^{--}) - (n_C^{++} + n_C^{--}) \\ n(\text{Correct, Changed}) &:= (n_C^{++} + n_C^{--}) \\ n(\text{Incorrect, Unchanged}) &:= (n^{+-} + n^{-+}) - (n_C^{+-} + n_C^{-+}) \\ n(\text{Incorrect, Changed}) &:= (n_C^{+-} + n_C^{-+}) \end{aligned}$$

where 'correctness' corresponds to the nodes u such that $\hat{x}(u) = x(u)$, while 'unchangedness' to u such that $x(u) = y(u)$.

Subtracting (A.41) from (A.40) then yields that for fixed $\hat{x}$

$$\mathbb{E}[T^{\hat{x}}(G') - T^{\hat{x}}(H) \mid \hat{x}] \tag{A.42}$$

$$\begin{aligned} =& \frac{(a-b)}{n}\Big(n(\text{Correct, Unchanged}) - n(\text{Incorrect, Unchanged})\Big) \\ & \times \Big(n(\text{Correct, Changed}) - n(\text{Incorrect, Changed})\Big). \end{aligned} \tag{A.43}$$

The lemma now follows on observing that

$$n(\text{Unchanged}) = n(\text{Correct, Unchanged}) + n(\text{Incorrect, Unchanged}),$$

and similarly $n(\text{Changed})$. □

A.2.1.5 Proof of Lemma A.2.3

Effectively, we are considering the following process: we have a bag of n balls - corresponding to the nodes - of two colours (types), Changed and Unchanged, and we are picking $k \leq n/2$ of them uniformly at random without replacement. Let

$$\eta_1 := n(\text{Unchanged, Incorrect}) \tag{A.44}$$
$$\eta_2 := n(\text{Changed, Incorrect}) \tag{A.45}$$

and

$$\begin{aligned} \zeta :=& (n(\text{Unchanged}) - 2n(\text{Incorrect, Unchanged})) & \text{(A.46)} \\ & \times (n(\text{Changed}) - 2n(\text{Incorrect, Changed})) \\ =& (n - s - 2\eta_1)(s - 2\eta_2). & \text{(A.47)} \end{aligned}$$

We now condition on the number of errors being k, which imposes the condition that $\eta_1 + \eta_2 = k$. Recall the sampling without replacement distribution, which implies that

$$P(\eta_1 = k - j, \eta_2 = j \mid d(\hat{x}, x) = k) = \frac{\binom{n-s}{k-j}\binom{s}{j}}{\binom{n}{k}}. \tag{A.48}$$

Thus,

$$\begin{aligned} \mathbb{E}[\eta_1 | d(x, \hat{x}) = k] &= \frac{k}{n}(n - s) \\ \mathbb{E}[\eta_2 | d(x, \hat{x}) = k] &= \frac{k}{n}(s) \\ \mathbb{E}[\eta_1\eta_2 | d(x, \hat{x}) = k] &= (n - s)(s)\frac{k(k-1)}{n(n-1)} = s(n - s)\left(\frac{k^2}{n^2} - \frac{k(n-k)}{n^2(n-1)}\right). \end{aligned}$$

As a consequence, we obtain that

$$\begin{aligned}\mathbb{E}[\zeta|d(x,\hat{x})=k] &= s(n-s)\left(1-4\frac{k}{n}+4\frac{k^2}{n^2}-4\frac{k(n-k)}{n^2(n-1)}\right)\\ &= s(n-s)\left(\left(1-2\frac{k}{n}\right)^2-4\frac{k(n-k)}{n^2(n-1)}\right) \qquad \text{(A.49)}\end{aligned}$$

Note that the above is decreasing as k increases for $k \le n/2$.

Note further that the Markov chain ζ–$d(\hat{x},x)$–G_1 holds. Thus the above also holds for $\mathbb{E}[\zeta \mid \mathcal{E}(G_1), d(x,\hat{x})=k]$.

We now condition on $\mathcal{E}(\hat{x})$ to find that

$$\begin{aligned}\frac{\mathbb{E}[\zeta \mid \mathcal{E}(\hat{x}),\mathcal{E}(G_1)]}{s(n-s)} &\ge \left((1-2\varepsilon_{\max})^2-4\frac{\varepsilon_{\max}(1-\varepsilon_{\max})}{n-1}\right) \qquad \text{(A.50)}\\ &\ge (1-2\varepsilon_{\max})^2-\frac{1}{n-1} \qquad \text{(A.51)}\end{aligned}$$

where we have used $\varepsilon_{\max} \le 1/2$, and the (unstated but obvious) condition that $n \ge 2$.

Applying the above to the result of Lemma A.2.5, we find that

$$\mathbb{E}[T^{\hat{x}}(G')-T^{\hat{x}}(H) \mid \mathcal{E}] \ge \left((1-2\varepsilon_{\max})^2-\frac{1}{n-1}\right)\frac{(a-b)}{n}(n-s)s. \qquad \square$$

A.2.1.6 Proof of Lemma A.2.4

Recall the notation from Appendix A.2.1.2. Under the null $H \overset{\text{law}}{=} G'$. Below, we will use G' as a proxy for H in the null distribution, and use H only in the alternate.

To begin with, observe that both G', H are independent of $G_1, \widetilde{G}, \hat{x}$, and that $\widetilde{G}$ is independent of $\hat{x}$ given G_1. Now, $T^{\hat{x}}$ is a signed sum of independent Bernoulli random variables with parameters smaller than a/n given G_1.. Thus, invoking results from Ch. 2 of [CL06] (and using that for $a \ge C$ for some large enough C implies that $a \ge 16\log(6n)/n \iff 1/6n \le \exp(-na/16))$), we find that for $\Gamma \in \{\widetilde{G}, G', H\}$,

$$P\left(\left|T^{\hat{x}}(\Gamma)-\mathbb{E}[T^{\hat{x}}(\Gamma) \mid G_1,\mathcal{E}]\right| \ge \sqrt{2na\log(6n)} \mid G_1,\mathcal{E}\right) \le \frac{1}{3n},$$

where we have used that $\mathcal{E}$ is determined given G_1 (i.e. $\mathcal{E}$ lies in the sigma-algebra generated by G_1.)

We now control the null and alternate fluctuations given $\mathcal{E}$.

Null: By the union bound, we find that

$$P\left(2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G') \ge \mathbb{E}[2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G') \mid G_1, \mathcal{E}] + 3\sqrt{2na\log(6n)} \mid G_1, \mathcal{E}\right) \le \frac{2}{3n}.$$

Recall from equation (A.36) from the proof of Lemma A.2.2 that $\mathbb{E}[2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G') \mid G_1, \mathcal{E}] \le a^2$. Feeding this in, we find that

$$P\left(2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G') \ge a^2 + 3\sqrt{2na\log(6n)} \mid G_1, \mathcal{E}\right) \le \frac{2}{3n}.$$

The right hand side above does no depend on G_1, and neither does the fluctuation radius wihtin the probability. Thus integrating over $P(G_1 \mid \mathcal{E})$, we find that

$$P\left(T \ge a^2 + 3\sqrt{2na\log(6n)} \mid \mathcal{E}\right) \le \frac{2}{3n},$$

where we have used that $T = 2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(H) \overset{\text{law}}{=} T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G')$ under the null.

Alt: Following the above development again, this time with lower tails, we find that given G_1 with probability at least $1 - 2/3n$,

$$\begin{aligned} 2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G') &\ge -(\mathbb{E}[2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G') \mid G_1, \mathcal{E}]) - 3\sqrt{2na\log(6n)} \\ T^{\hat{x}}(G') - T^{\hat{x}}(H) &\ge +(\mathbb{E}[T^{\hat{x}}(G') - T^{\hat{x}}(H) \mid G_1, \mathcal{E}]) - 2\sqrt{2na\log(6n)} \end{aligned}$$

Further, given $(G_1, \mathcal{E})$, by Lemmas A.2.2, A.2.3 we have

$$\begin{aligned} 2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(G') &\ge -a^2 - 3\sqrt{2na\log(6n)} \\ T^{\hat{x}}(G') - T^{\hat{x}}(H) &\ge +\kappa(a-b)s(1-s/n) - 2\sqrt{2na\log(6n)}, \end{aligned}$$

where $\kappa = (1 - 2\varepsilon_{\max})^2 - 1/(n-1)$. Adding the above, we find by the union bound that

$$P\left(2T^{\hat{x}}(\widetilde{G}) - T^{\hat{x}}(H) \ge \kappa(a-b)s(1-s/n) - a^2 - 5\sqrt{2na\log(6n)} \mid G_1, \mathcal{E}\right) \ge 1 - \frac{4}{3n}.$$

The claim follows on noting that the right hand side and the fluctuation radius do not depend on G_1, and integrating the inequality over G_1. □

A.2.2 Proof of the converse bound from Theorem 2.3.1.

We restate the lower bound below as a proposition:

Proposition. *There exists a universal constant C, and another $c < 1$ that depends on C, such that if $\Lambda \le C$ and $s \le \frac{n}{2}(1-c)$, then reliable two-sample testing of balanced communities for s changes is impossible for large enough n.*

In particular, for $a + b < n/4$, the statement holds with $C = 1/8, c = 1/6$ for $n \ge 136$, and in this case, $R_{\mathrm{TST}} \ge 0.25$.

Proof. The proof proceeds by using a variation of Le Cam's method, and importing impossibility results for the so-called distinguishability problem [BMNN16]. In particular, suppose that in the null distribution, the communities are drawn according to the uniform prior on balanced communities, denoted by π. Further, assume that if a s-change is made, then the resulting community is chosen uniformly from all communities that are at least s far from the null community. We have the hypothesis test:

$$H_0 : (G, H) \sim \sum_{x \in \mathcal{B}} \pi_x P(G|x)P(H|x) \quad \text{vs } H_1 : (G, H) \sim \sum_{x,y \in \mathcal{B}} \pi_x \pi_{y|x} P(G|x)P(H|y),$$

where we use $\mathcal{B}$ to denote the set of balanced communities, and $\pi_{y|x}$ is the uniform distribution on $\mathcal{B} \cap \{y : d(x, y) \ge s\}$. For succinctness, let us denote the null and alternate distributions above as p_{null} and p_{alt} respectively.

Once again, by Neyman-Pearson theory,

$$R_{\mathrm{TST}} \ge R_\pi \ge 1 - d_{\mathrm{TV}}(p_{\mathrm{null}}, p_{\mathrm{alt}}) \ge 1 - d_{\mathrm{TV}}(p_{\mathrm{null}}, Q) - d_{\mathrm{TV}}(Q, p_{\mathrm{alt}}),$$

where Q is any distribution, and the last inequality is since d_{TV} is metric.

We choose Q to be the unstructured distribution induced by an Erdős-Rényi graph of parameter $(a + b)/2n$. The primary reason for this is that explicit control on the total variation distance between p_{null} and Q is then available - for instance, by [WX18, §3.1.2], we have

$$D_{\chi^2}(p_{\mathrm{null}} \| Q) + 1 \le \mathbb{E}\left[\exp\left(\tau \left(\frac{4\mathscr{H} - n}{\sqrt{n}}\right)^2\right)\right],$$

where $\mathscr{H}$ is a Hypergeometric$(n, n/2, n/2)$ random variable, and

$$\tau = \frac{(a-b)^2}{2n(a+b)} + \frac{(a-b)^2}{2n(2-a/n-b/n)}.$$

Notice the extra factor of 2 compared to the expressions in [WX18], which arises since we sum over two independent graphs G, H and not one. We observe that

$$\tau = \Lambda \frac{n}{2n-a-b},$$

and explicitly, if $a+b \le n/4$, then $\tau \le \frac{4}{7}\Lambda$.

We now consider the alternate term. As a preliminary, let

$$\gamma := \frac{\sum_{k=0}^{s-1} \binom{n/2}{k/2}^2}{\binom{n}{n/2}}.$$

Note that γ is the probability that two balanced communities chosen independently and uniformly, lie within distortion s. Indeed, since communities are formed by identifying antipodal points in the boolean cube, the probability of picking a community at distortion $< s$ coincides with that of picking a balanced vector at Hamming distance $< s$ from a given balanced vector in the cube $\{0,1\}^n$. The denominator in γ is clearly the number of balanced vectors in the cube, while the numerator is the number of balanced vectors at a distance of $< s$ from any given balanced vector - we choose $k < s$, and choose $k/2$ points marked 1 and $k/2$ marked 0, and flip them all.

As a consequence, we find that for any $x, y \in \mathcal{B}$,

$$\pi_{y|x} \le \frac{\pi_y}{1-\gamma}.$$

Thus, in the χ^2 expressions for p_{alt}, we have

$$\begin{aligned}
&\mathbb{E}_{(G,H)\sim Q^{\otimes 2}}\left[\left(\frac{p_{\text{alt}}}{Q}\right)^2\right] \\
&\quad= \sum_{x,y,x',y'} \mathbb{E}\left[\frac{P(G|x)P(G|x')}{Q^2(G)}\frac{P(H|y)P(H|y')}{Q^2(H)}\right]\pi_x\pi_{x'}\pi_{y|x}\pi_{y'|x'} \\
&\quad\le \frac{1}{(1-\gamma)^2}\sum_{x,y,x',y'} \mathbb{E}\left[\frac{P(G|x)P(G|x')}{Q^2(G)}\frac{P(H|y)P(H|y')}{Q^2(H)}\right]\pi_x\pi_{x'}\pi_y\pi_{y'} \\
&\quad= \frac{1}{(1-\gamma)^2}\left(1+\chi^2(\sum_{x\in\mathcal{B}}\pi_x P(G|x)\|Q(G))\right)^2
\end{aligned}$$

Since the final quantity is explicitly controlled in the cited section, we also have

$$\begin{aligned}
1 + D_{\chi^2}(p_{\text{alt}}\|Q) &\le \frac{1}{(1-\gamma)^2}\mathbb{E}\left[\exp\left(\frac{\tau}{2}\left(\frac{4\mathscr{H}-n}{\sqrt{n}}\right)^2\right)\right]^2 \\
&\le \frac{1}{(1-\gamma)^2}\mathbb{E}\left[\exp\left(\tau\left(\frac{4\mathscr{H}-n}{\sqrt{n}}\right)^2\right)\right],
\end{aligned}$$

the final relation arising from Jensen's inequality.

Since the quantity appears often, we let

$$\beta := \mathbb{E}\left[\exp\left(\tau\left(\frac{4\mathscr{H}-n}{\sqrt{n}}\right)^2\right)\right].$$

Invoking the inequality $d_{\text{TV}} \le \sqrt{\log(1+D_{\chi^2})/2}$, we find that

$$R_{\text{TST}} \ge 1 - \sqrt{\log(\beta)/2} - \sqrt{\log(\beta(1-\gamma)^{-2})/2} = 1 - \sqrt{\log(\beta/(1-\gamma))}.$$

Note that the only s-dependent term in the above bounds is γ. We first offer control on the γ, and claim that for $s/n < 1/2$, $\gamma \to 0$. Indeed, since $s \le n/2$, and by standard refinements of Stirling's approximation (for instance, we use [Gal68, Exercise

5.8] below),

$$\gamma \leq s\frac{\binom{n/2}{s/2}^2}{\binom{n}{n/2}} \leq s\frac{1}{2\pi}\frac{n/2}{s/2(n-s)/2}2^{nh_2(s/n)}\left(\sqrt{\frac{n}{8(n/2)^2}}2^n\right)^{-1}$$
$$\leq \sqrt{\frac{2n}{\pi^2}}2^{-n(1-h_2(s/n))},$$

where h_2 is binary entropy in bits.

At this point the argument in the limit as $n \to \infty$ is complete - since $4(\mathscr{H} - n)/\sqrt{2n} \overset{\text{Law}}{\to} \mathcal{N}(0,1)$, β is bounded as $n \to \infty$ by $\sqrt{1-2\tau}$ if $\tau < 1/2$, and since in this limit $\tau \to \Lambda/2$ (for $a, b = o(n)$), we obtain that if $\limsup s/n < 1/2$, and $\Lambda < 1$, then $\liminf R_{\text{TST}} > 0$.

Non-asymptotic bounds can be recovered by giving up space on the constants, leading to the statement we have claimed.

Concretely, to attain $R_{\text{TST}} > 1/4$, it suffice to show that $\beta(1-\gamma)^{-1} < e^{9/16}$. Now, for $s < n/3$, we have

$$(1-\gamma)e^{9/16} \geq \left(1 - \sqrt{2n/\pi^2}2^{-0.08n}\right)e^{9/16} > 1.75$$

for $n \geq 136$.[2] Thus, it suffices to control β to below 1.75 in this regime. To this end, note that $u \mapsto \exp\left(\tau((4u-n)/\sqrt{n})^2\right)$ is a continuous, convex map, and thus, by [Hoe63, Thm. 4],

$$\beta \leq \mathbb{E}\left[\exp\left(\tau\left(\frac{4\mathscr{B}-n}{\sqrt{n}}\right)^2\right)\right],$$

where $\mathscr{B} \sim \text{Bin}(n/2, 1/2)$.

[2] This is calculated using a computer algebra system. Analytically it is still easy to argue something similar - for, say, $n \geq 10000$, the expression is at least $\sqrt{e} * 0.99 > 1.4$, continuing along which leads to an analytic proof of the conclusion holding for $\Lambda < 1/16$ by following the next footnote.

By Chernoff's bound, $P(|\mathscr{B} - n/4| \geq \sqrt{n}u) \leq 2e^{-4u^2}$, and thus, we have

$$\begin{aligned}\beta &\leq \int\limits_0^\infty P\left(\exp\left(\tau\left(\frac{4\mathscr{B} - n}{\sqrt{n}}\right)^2\right) \geq u\right) \mathrm{d}u \\ &\leq \int\limits_0^\infty \min(1, 2u^{-1/4\tau})\, \mathrm{d}\tau \\ &= \frac{2^{4\tau}}{1 - 4\tau},\end{aligned}$$

the final equality holding so long as $1/4\tau > 1 \iff \tau < 1/4$. The original claim follows if

$$\frac{2^{4\tau}}{1 - 4\tau} \leq \frac{7}{4},$$

which is true for $\tau < 0.074$. Since $\tau \leq 4/7\Lambda, \Lambda < 1/8$ implies that $\tau < 4/56 < 0.072$.[3] □

A couple of quick comments are useful here:

1. Note that the above cannot be applied usefully to GoF. This is because in GoF, the null is explicitly available, and we do not have the benefit of averaging with π in the TV expressions. This causes the equivalent term $\chi^2(P(G|x_0)\|Q(G))$ to grow exponentially with $n\Lambda$.

2. The above characterises the tightness of our claimed bounds for TST of large changes - the method works if $\Lambda = \Omega(1)$ and $s \gg \sqrt{n \log n}$, and by the above argument, no test can work if $\Lambda \ll 1$, as long as the change is not extreme ($\limsup s/n < 1/2$).

3. While the above approach is wasteful in how it utilises s, this is actually a non-issue, since the bounds require a separate control on $d_{\mathrm{TV}}(p_{\mathrm{null}}\|Q)$, which

[3] The number 0.074 is calculated using a computer algebra system. Purely analytic calculations are straightforward as well - for example by using $2^{4\tau} \leq 1 + 4\tau$ for $\tau < 1/4$, which can be proved by noting that $1 + 4\tau - 2^{4\tau}$ is initially increasing, and then strictly decreasing after a point, and that $1/4$ is a root of this function. This implies that the conclusion holds so long as $\tau < 3/44$, which hold if $\Lambda < 21/176 \approx 0.119$.

can only be controlled if $\Lambda = O(1)$. In particular, we cannot pull out better bounds for the small s situation from the above.

A.3 Experimental Details

A.3.1 Experiments on SBMs

The experiemnts simulate an ensemble of GoF and TST test and evaluate the performance of the two schemes using the sum of false alarm and missed detection probabilities $(FA + MD)$.

While the GoF scheme is implemented precisely as in the main text, the experiments use a slightly modified version of Algorithm 1 for the TST:

(i) G_1 subsamples G at a rate η, and the test statistic T is appropriately modified: $T := \frac{1}{1-\eta} T^{\hat{x}_1}(\widetilde{G}) - T^{\hat{x}_1}(H)$. Intrinsically, the spectral clustering step is the more singal-sensitive part of the scheme 1. While splitting the graphs equally is fine for theoretical results, it is better in practice to devote more SNR to the clustering step, and less to compute the test statistic, which can be done by increasing η. In the following, we set $\eta = 0.85$. Other values of η are explored in Appendix A.3.1.2.

(ii) The constant factor in the threshold developed in the test is conservative, and we vary it to adjust for different values of η and to mitigate its suboptimality. In the experiments, we used the threshold $\frac{3}{4}\sqrt{n(a+b)\log(6n)}$.

As noted in the main text, the experiments are performed for various (s, Λ) for a fixed value of $a/b = 3$. Λ is varied between Λ_0 and $10\Lambda_0$ for $\Lambda_0 = 3/4log(n/100) \approx 1.7$. This is significantly below the theoretical threshold of 2 necessary for non-trivial recovery. Further, $8\Lambda_0 = 2\log(n)$, at which point recovery with constant order distortion becomes viable.

A.3.1.1 Implementation details

The experiment is setup as follows:

1. We fix a value of $\Lambda_0 = 3/4 \log(n/100)$ as above. Then, for some choice of $b/a = r$, we choose (a, b) satisfying $\Lambda = \alpha\Lambda_0$ and $\alpha \in [1, 10]$. r is set to be 1/3.

2. For a fixed number of nodes, n, and for $s \in [1 : n/2]$, we consider the balanced partition $x = [x_i]_{i=1}^n$ with

$$x_i = \begin{cases} 0, & 0 \leq i \leq n/2 \\ 1, & n/2 < i \leq n \end{cases}$$

for the null distribution, and the shifted balanced partition $y = [y_i]_{i=1}^n$ with

$$y_i = \begin{cases} 0, & s/2 < i \leq n/2 + s/2 \\ 1, & i \in (n/2, n] \cup [0, s/2] \end{cases}$$

for the alternate distribution. This ensures that $d(x, y) = s$. We take $\lfloor\cdot\rfloor$ whenever s or n are odd.

3. We sample $G, G' \sim P(\cdot\,|\,x)$ and $H \sim P(\cdot\,|\,y)$, where P represents drawing from an SBM with parameters n, a and b, as described in §2.1

GoF procedure. Recall that we are given a proposed partition x_0. Here we set $x_0 = x$. The results of running the tests on the graph G then serve to characterise size, and those on H serve to characterise power.

1. For the naïve scheme, we produce partitions $\hat{x}$ and $\hat{y}$ from G and H respectively via spectral clustering (see below for details), and declare for null in either case if $d(x_0, z) < s/2$, where z is respectively $\hat{x}$ and $\hat{y}$.

2. For the alternate scheme, we instead compute the statistic from §2.2, and reject on the basis of the threshold developed there.

TST procedure. Similarly to the above, runs on the pair (G, G') serve to characterise size, and on (G, H) serve to characterise power of the test. Precisely:

1. For the naïve two-sample test based on recovery and comparison, we estimate $\hat{x}, \hat{x}'$ and $\hat{y}$ from G, G' and H respectively. The structure is estimated using spectral clustering (see below for implementation details). We declare that a change has occurred if $d(\hat{x}, \hat{x}') \geq s/2$, and no change if $d(\hat{x}, \hat{y}) < s/2$. We get a false alarm every time we declare a change on the pair (G, G'), and we miss a detection whenever we declare no change on the pair (G, H). The false alarm and missed detection probabilities are estimated as an average over $M = 100$ samples.

2. For the two-sample test based on Algorithm 1, we follow the algorithm as stated, making only the modifications previously described. To be precise, we estimate $\hat{x}_1$ from G_1, a subsampling of (the edges of) G at rate η. Then, we compute the test statistics in the null and alternate distributions:

$$T_{\text{Null}} = \frac{1}{1-\eta} T^{\hat{x}_1}(\widetilde{G}) - T^{\hat{x}_1}(G')$$

and

$$T_{\text{Alt.}} = \frac{1}{1-\eta} T^{\hat{x}_1}(\widetilde{G}) - T^{\hat{x}_1}(H),$$

where $\tilde{G} = G - G_1$.

In both the above cases, the simulations are performed over a range of $\Lambda = \alpha \Lambda_0$ and s, where $\alpha \in [1, 10]$ and $s \in (0, 250)$. Performance is indicated using the sum of false alarm and missed detection rates.

Details associated with the implementation of the aforementioned schemes are given below:

1. All experiments were implemented in the Python language (v3.5+), using the Numpy (v1.12+) and Scipy (v0.18+) packages [Oli06; JOP01].
2. Structure learning was performed using the Spectral Clustering [vLux07] algorithm, as implemented by the Scikit-learn package (v0.19.1+) [Ped+11].
3. Spectral Clustering was regularized in the manner suggested by [JY16]. Effectively, if G was the adjacency matrix to be submitted to the Scikit-learn spectral clustering function, we performed pre-addition, and instead passed $G + \tau \mathbf{1}\mathbf{1}^\top$. We set $\tau = \frac{1}{10n}$, which proved sufficient to run the spectral clustering function with no errors or warnings.
4. All plots were generated using Matplotlib (v2.1+) [Hun07].

A.3.1.2 Modifications to η

For completeness, we demonstrate how the performance of the modified two-sample test based on Algorithm 1 varies as η is changed. Figure A·2 compares the naïve two-sample test against the scheme based on Algorithm 1, for three different values of η: 0.7, 0.8 and 0.9.

We use the following parameters: $n = 500$, $\mathrm{SNR}_0 = \frac{3}{8}\log(n/100) = \frac{3}{8}\log 5 \approx 0.5$, $\frac{b}{a} = r = 1/3$. For $\eta = 0.7$ and $\eta = 0.8$, the threshold used is $\sqrt{n(a+b)\log(6n)}$, while for $\eta = 0.9$, we used a higher threshold of $\frac{3}{2}\sqrt{n(a+b)\log(6n)}$.

While differences are rather subtle, a careful examination may reveal that as η increases, the failure region recedes, while the success region advances in the high-s, low-SNR regime. However, the cost of this is an increased threshold to maintain success at $\delta = 0.1$, and a wider transition region, indicating that different η might be optimal at different n.

A.3.2 Experiments on the Political Blogs dataset

While the original graph has 1490 nodes, we followed standard practice in selecting the largest (weakly) connected component of the graph, which contains 1222 nodes. We denote this graph as G.The true partition of the blogs according to political leaning is available, denoted x_{True} here. This also allows accurate estimates of the graph parameters (a, b) to be made, and we use these estimates for a, b for GoF, and for the semi-synthetic procedure for TST. We found that $\hat{a} \approx 49.5$, while $\hat{b} \approx 5.2$, giving a ratio $a/b \approx 10$. The communities, according to x_{True} are of sizes 636 and 586.

The regime of low Λ is explored via sparsification. Fixing a $\rho \in (0, 1]$, sparsification is performed by independently flipping coins for each edge in G, and keeping the edge with probability ρ. We refer to ρ as the rate of sparsification.

We lastly note that at no sparsification ($\rho = 1$), spectral clustering produces a partition $\hat{x}_1$ such that $d(x_{\text{True}}, \hat{x}_1) = 56$.

GoF Procedure.

1. The graph is sparsified at rate ρ. Let the sparsened graph be G_ρ.

2. For the naïve recovery based scheme, spectral clustering is performed on G_ρ as in the previous section to generate $\hat{x}_\rho$.

3. For the proposed test from §2.2, the statistic is computed on G_ρ.

4. The size of the test is estimated by running the GoF tests with $x_0 = x_{\text{True}}$. For the naïve scheme, we reject if $d(\hat{x}_\rho, x_0) \geq s/2$; for the proposed scheme, we use the test from §2.2.

5. To compute the power at distortion s, we first generate y by randomly inverting the community labels of s nodes in x_{True}. We then run the same procedure as in the previous line, but with $x_0 = y$. Note that the graphs are not edited in any way.

6. The precise implementation details are exactly as in Appendix A.3.1.1, with the minor difference that we use a regularizer of $\tau = 1$ for spectral clustering.

TST Proceudre.

1. Recall that TST requires two graphs as input. The experiment compares the political blogs graph against SBMs.

2. To compute the size, we require a graph with the same underlying communities as G. Thus we generate G', which is drawn as an SBM with the underlying partition x_{True}, and parameters a, b as estimated from the political blogs graph G.

3. To determine the power of the tests we need a graph with an s-far underlying community. For this, we first generate a y such that $d(x_{\text{True}}, y) = s$, as we did in the GoF Procedure. Next, we sample H as an SBM with underlying partition y.

4. The graphs G, G' and H are now all sparsified at rate ρ to get G_ρ, G'_ρ and H_ρ.

5. The size of each test is estimated using the TST procedures, as described in Appendix A.3.1.1 on the pair (G_ρ, G'_ρ). Power is similarly estimated using the TST procedures on the pair (G_ρ, H_ρ).

A.3.3 Experiments on the GMRFs

Following the heuristic detailed in §2.4.3, we naïvely generalise community recovery and testing to this setting, by replacing all instances of the graph adjacency matrix in previous settings with the sample covariance matrix.

The Gaussian Markov Random Field is described by its precision matrix Θ (i.e., the inverse covariance matrix of the Gaussian random vector on its nodes). We perform a preliminary examination of the possibility of testing changes in communities for an SBM-structured GMRF even when learning the structure is hard or impossible. As described in Section 2.4.3, we set

$$\Theta = I + \gamma G,$$

where G is the adjacency matrix of an SBM with known parameters. We generate samples from the GMRF as follows:

1. For a fixed number of nodes n, we fix an SNR for the SBM, Λ, and compute (a, b) satisfying this Λ so that $b/a = r$.

2. Here, we consider $n = 1000$ nodes and take $\Lambda = 30\Lambda_0$, where $\Lambda_0 = \frac{10}{11}\log(\frac{n}{100}) \approx 2.1$, as before, and $r = 1/10$. We find that $(a, b) \approx (12.34 \log n, 1.234 \log n)$. Note that since $\Lambda \approx 63 \approx 10 \log(n)$, recovery of the communities for a raw SBM at this SNR is trivial.

3. We fix a GMRF parameter γ. Here, we take $\gamma = 3/(a + b) \approx 0.032$.

4. We can now construct the precision matrix Θ after sampling G from the SBM. We re-sample to ensure that Θ is positive-definite, but in practice, for the value of γ quoted above, we did not encounter the need to re-sample.

5. To generate i.i.d. samples $\zeta \sim \mathcal{N}(0, \Theta^{-1})$ in a stable manner, we use the following algorithm:

   (a) Compute the lower-triangular Cholesky factor R of Θ, so that $\Theta = RR^\top$.

(b) Sample $\xi \sim \mathcal{N}(0, I)$ from a standard n-dimensional multivariate normal distribution.

(c) Solve for ζ in $R^\top \zeta = \xi$.

This suffices, since, $\zeta = (R^\top)^{-1}\xi$ would then have the covariance matrix $(R^\top)^{-1}R^{-1} = (RR^\top)^{-1} = \Theta^{-1}$.

6. In this manner, we generate samples from the null and alternate distributions: let ζ, ζ' and υ respectively denote samples drawn from a GMRF structured using G, G' and H respectively. Here, G, G' and H exactly are as described in Section A.3.1.1.

Next, we describe how each of the two schemes is evaluated:

1. Assuming we have t i.i.d. samples of ζ, generated as described above, we estimate the covariance matrix $\hat{\Sigma}$ of ζ using the standard estimator:

$$\hat{\Sigma} = \frac{1}{t-1}\sum_{i=1}^{t}(\zeta_i - \bar{\zeta})(\zeta_i - \bar{\zeta})^\top,$$

where $\bar{\zeta} = \frac{1}{t}\sum_{i=1}^{t}\zeta_i$. We then compute the correlation matrix,

$$\hat{C} : \hat{C}_{ij} = \frac{\hat{\Sigma}_{ij}}{\sqrt{\hat{\Sigma}_{ii}\hat{\Sigma}_{jj}}},$$

which will be used in place of the adjacency matrix for both two-sample testing schemes.

2. Similarly, we compute $\hat{C}$, $\hat{C}'$ and $\hat{D}$ from ζ, ζ' and υ respectively.

3. The naïve two-sample test based on recovery and comparison is evaluated exactly as described in Section A.3.1.1, except that $\hat{C}$, $\hat{C}'$ and $\hat{D}$ are used in place of G, G' and H respectively. False alarm and missed detection rates are also computed in exactly the same way.

4. The two-sample test based on Algorithm 1 has several important variations:

   (a) We use the test statistics

   $$T_{\text{Null}} = T^{\hat{x}}(\hat{C}) - T^{\hat{x}}(\hat{C}')$$
   $$T_{\text{Alt.}} = T^{\hat{x}}(\hat{C}) - T^{\hat{x}}(\hat{D}),$$

   for the null and alternate distributions respectively. Here, $\hat{x}$ has been estimated from $\hat{C}$.

   (b) The threshold for the test is estimated from data. That is, we simulate $M = 100$ samples of T_{Null} and $T_{\text{Alt.}}$ each, and fit a classifier to differentiate between the two distributions. The classifier used is a simplistic 1-dimensional Linear Discriminant Analysis.

   (c) We estimate false alarm and missed detection rates by applying the classifier to a hold-out dataset. To use the data as efficiently as possible, we use 10-fold repeated, stratified cross-validation, with 10 repetitions.

Remark on subsampling.

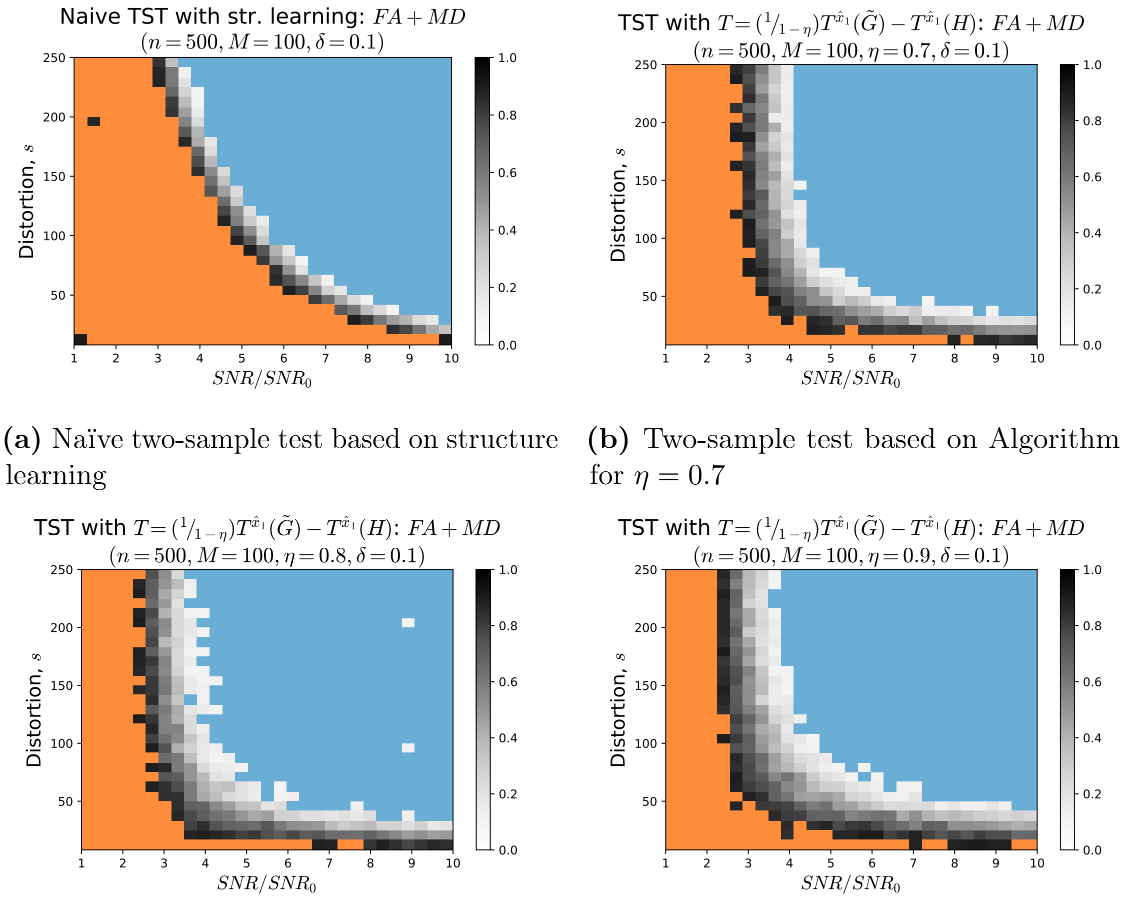
1. Note that in the two-sample test for GMRFs based on Algorithm 1, we do not subsample $\hat{C}$ as we did before in the case of SBMs.

2. While previously, we had subsampled G to create two subgraphs G_1 and $\tilde{G}$ that shared independence properties for ease of theoretical analysis, it should be noted that subsampling results in an effective loss of SNR. This is also the reason why we had to adjust the implementation using a different rate η.

3. However, it emerges empirically that skipping the subsampling entirely, with a completely dependent $\hat{x}$ and G, makes for better separation between the null and alternate distributions, providing a more powerful statistic.

4. Since we could not analytically derive a threshold for this statistic, we presented the subsampled test statistic for the first experiment.

5. Since in the case of GMRFs, we are estimating the threshold from data, we use the more powerful test statistic to show the full extent of possible gains when using a dedicated algorithm for change detection, instead of naïvely looking for changes by learning community structures first.



(a) Naïve two-sample test based on structure learning

(b) Two-sample test based on Algorithm 1 for $\eta = 0.7$

(c) Two-sample test based on Algorithm 1 for $\eta = 0.8$

(d) Two-sample test based on Algorithm 1 for $\eta = 0.9$

Figure A·2: A comparison between the naïve two-sample test based on structure learning, and the two-sample test we propose in Algorithm 1, for $\eta \in \{0.7, 0.8, 0.9\}$. Error rates lower than $\delta = 0.1$ have been shaded blue to represent "success", while those higher than $1 - \delta = 0.9$ have been shaded orange to represent "failure".

Appendix B

Appendix to Chapter 3

B.1 Appendix to §3.2

B.1.1 Proof of Ordering of Sample Complexities

The proposition is argued by direct reductions showing how a solver of a harder problem can be used to solve a simpler problem. The main feature of the definitions that allows this is that the risks of SL and EoF are defined in terms of a probability of error.

Proof of Proposition 3.2.1.

Reducing EoF to SL: Suppose we have a $(s-1/2)$-approximate structure learner with risk δ that uses n samples. Then we can construct the following EoF estimator with the same sample costs. Take a dataset from $Q^{\otimes n}$, and pass it to the structure learner. With probability at least $1-\delta$, this gives a graph $\widehat{G}$ that is at most $\lfloor s/2 \rfloor$-separated from $G(Q)$. Now compute $G(P)\triangle\widehat{G}$ ($G(P)$ is determined because P is given to the EoF tester). By the triangle inequality applied to the adjacency matrices of the graphs under the Hamming metric, this identifies $G(P)\triangle G(Q)$ up to an error of $(s-1)/2$, and so, the EoF risk incurred is also δ. Taking $\delta = 1/8$ concludes the argument.

Reducing GoF to EoF: Suppose we have a s-EoF solver that uses n samples with risk δ. Again, take a dataset from $Q^{\otimes n}$, and pass it to the EoF solver, along with P. With probability at least $1-\delta$, this yields a graph $\widehat{G}$ such that $|\widehat{G}\triangle(G(P)\triangle G(Q)| \le (s-1)/2$. But then, if $G(Q) = G(P)$, $\widehat{G}$ can have at most $(s-1)/2$ edges, while if $|G(P)\triangle G(Q)| \ge s$, then $\widehat{G}$ must have at least $(s+1)/2$ edges. Thus, thresholding on the basis of the number of edges in $\widehat{G}$ produces a GoF tester with both null and alternate risk controlled by δ, or total risk 2δ. Taking $\delta = 1/8$ then finishes the argument. □

B.1.2 Proof of Upper Bound on n_{SL}

This proof is essentially constructed by slightly improving upon the proof of [SW12, Thm 3a)] due to Santhanam & Wainwright, which analyses the maximum likelihood scheme. We use notation from that paper below.

Proof of Theorem 3.2.2. [SW12] shows, in Lemmas 3 and 4, that if the data is drawn from an Ising model $P \in \mathcal{I}_d$, and $Q \in \mathcal{I}_d$ is such that $G(P) \triangle G(Q) = \ell$, then

$$P^{\otimes n}(\mathscr{L}(P) \le \mathscr{L}(Q)) \le \exp\left(-n\ell\kappa/8d\right),$$

where $\mathscr{L}(P)$ denotes the likelihood of P, i.e. if the samples are denoted $\{X^{(k)}\}_{k \in [1:n]}$, then $\mathscr{L}(P) = \prod_{k=1}^n P(X^{(k)})$, and

$$\kappa = (3e^{2\beta d} + 1)^{-1} \sinh^2(\alpha/4) \ge \frac{\sinh^2(\alpha/2)}{4e^{2\beta d}}.$$

Now, for the max-likelihood scheme to make an error in approximate recovery, it must make an error of at least s - i.e., an error occurs only if $\mathscr{L}(Q) \ge \mathscr{L}(P)$ for some Q with $G(Q) \triangle G(P) \ge s$. Union bounding this as Pg. 4129 of [SW12], we may control this as

$$\begin{aligned} P(\text{err}) &\le \sum_{\ell=s}^{pd} \binom{\binom{p}{2}}{\ell} \exp\left(-n\ell\kappa/8d\right) \\ &\le \sum_{\ell=s}^{pd} \exp\left(\ell\left(\log\frac{ep^2}{2\ell} - n\kappa/8d\right)\right) \\ &\le \sum_{\ell=s}^{pd} \exp\left(\ell\left(\log\frac{ep^2}{2s} - n\kappa/8d\right)\right). \end{aligned}$$

Now, if $n\kappa/8d \ge 2\log {ep^2}/{2s} = 2\log {p^2}/{s} + 2(1 - \log(2))$, and if $\exp\left(-ns\kappa/8d\right) \le 1/2$ then the above is bounded as $2\exp\left(-ns\kappa/8d\right)$, which can be driven lower than any δ by increasing n by an $O(s^{-1}\log(2/\delta))$ additive factor. It follows that

$$n_{\mathrm{SL}}(s, \mathcal{I}) \le \frac{16d}{\kappa}\left(\log\frac{p^2}{s} + 2 + O(1/s)\right),$$

and the claim follows by expanding out the value of κ. □

B.2 Appendix to §3.3

B.2.1 Expanded Proof Technique

This section expands upon §3.3.1 in the main text, including a treatment of the method used for EoF lower bounds, giving an expanded version of Lemma 3.3.4, and a theorem collating the resulting method to construct bounds. Some of the text from §3.3.1 is repeated for the sake of flow of the presentation.

As discussed previously, the proofs proceed by explicitly constructing distributions with differing network structures that are statistically hard to distinguish. In particular, we measure hardness by the χ^2-divergence. We begin with some notation.

Definition *A s-change ensemble in $\mathcal{I}$ is a distribution P and a set of distributions $\mathcal{Q}$, denoted $(P, \mathcal{Q})$, such that $P \in \mathcal{I}, \mathcal{Q} \subseteq \mathcal{I}$, and for every $Q \in \mathcal{Q}$, it holds that $|G(P) \triangle G(Q)| \geq s$.*

Each of the testing bounds we show will involve a mixture of n-fold distributions over a class of distributions. For succinctness, we define the following symbol.

Definition *For a set of distributions $\mathcal{Q}$ and a natural number n, we define the mixture*

$$\langle \mathcal{Q}^{\otimes n} \rangle := \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} Q^{\otimes n}.$$

Consider the case of GoF testing, with the known distribution P. Suppose we provide the tester with the additional information that the dataset is drawn either from P, or from a distribution picked uniformly at random from $\mathcal{Q}$, where $(P, \mathcal{Q})$ for a s-change ensemble. Clearly, the Bayes risk suffered by any tester with this side information must be lower than the minimax risk of GoF testing. The advantage of this formulation is that the risks of these tests with the side information can be lower bounded by standard techniques - basically the Neyman-Pearson Lemma. The following generic bound, which is Le Cam's two point method [Yu97; IS12] captures this.

Lemma B.2.1. *(Le Cam's Method)*

$$R^{\mathrm{GoF}}(n,s,\mathcal{I}) \geq \sup_{(P,\mathcal{Q})} 1 - d_{\mathrm{TV}}(\langle \mathcal{Q}^{\otimes n}\rangle, P^{\otimes n}) \geq \sup_{(P,\mathcal{Q})} 1 - \sqrt{\frac{1}{2}\log(1+\chi^2(\langle \mathcal{Q}^{\otimes n}\rangle \| P^{\otimes n}))},$$

where the supremum is over s-change ensembles in $\mathcal{I}$.

Above, $\chi^2(\cdot\|\cdot)$ is the χ^2-divergence, which is defined for distributions P, Q as follows

$$\chi^2(Q\|P) := \begin{cases} \mathbb{E}_P\left[\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^2\right] - 1 & \text{if } Q \ll P \\ \infty & \text{if } Q \not\ll P \end{cases}.$$

Note that generally the method is only stated as the first bound, and the second is a generic bound on the total variation divergence which follows from Pinsker's inequality and the monotonicity of Rényi divergences. The χ^2-divergence is invoked becuase it yields a twofold advantage in that it both tensorises well, and behaves well under mixtures such as $\langle \mathcal{Q}^{\otimes n}\rangle$ above.

For the EoF bounds, more care is needed. Recall that the EoF problem only requires errors smaller than $s/2$. To address this, we introduce the following.

Definition *An (s', s)-packing change ensemble is an s-change ensemble $(P, \mathcal{Q})$ such that $\mathcal{Q}$ is an s'-packing under the Hamming metric on network structures, that is, for every $Q, Q' \in \mathcal{Q}$, $|G(Q) \triangle G(Q')| \geq s'$.*

Clearly, if one can solve the EoF problem, one can exactly recover the structures in a $(s/2, s)$-packing change ensemble. Thus, the following lower bound of Guntuboyina is applicable.

Lemma B.2.2. *[Gun11, Example II.5]*

$$R^{\mathrm{EoF}}(n,s,\mathcal{I}) \geq \sup_{(P,\mathcal{Q})} 1 - \frac{1}{|\mathcal{Q}|} - \sqrt{\frac{\sum_{Q\in\mathcal{Q}} \chi^2(Q\|P)}{|\mathcal{Q}|^2}},$$

where the supremum is taken over $(s/2, s)$-packing change ensembles in $\mathcal{I}$.

Note that [Gun11] shows a number of lower bounds of the above form. We use the χ^2-divergence here primarily for parsimony of effort, in that the bounds on χ^2-divergences we construct for the GoF setting can easily extended to the EoF case via the above.

Our task is now greatly simplified - we merely have to construct change ensembles such that $|\mathcal{Q}|$ is large, and $\chi^2(Q\|P)$ is small for every $Q \in P$. Since it is difficult to directly construct large degree bounded graphs with tractable distributions, we will instead provide constructions on a small number of nodes, and lift these up to the whole p nodes by the following lemma.

Lemma B.2.3. *(Lifting) Let P_0 and Q_0 be Ising models with degree $\le d$ on $\nu \le p$ nodes such that $|G(P_0)\triangle G(Q_0)| = \sigma$, and $\chi^2(Q_0^{\otimes n}\|P_0^{\otimes n}) \le a_n$. Let $m := \lfloor p/\nu \rfloor$. For $1 \le t < m/16e$, there exists a $t\sigma$-change ensemble $(P, \mathcal{Q})$ over p nodes such that $|\mathcal{Q}| = \binom{m}{t}$ and*

$$\chi^2(\langle \mathcal{Q}^{\otimes n}\rangle \| P^{\otimes n}) \le \frac{1}{\binom{m}{t}} \sum_{k=0}^{t} \binom{t}{k}\binom{m-t}{t-k}((1+a_n)^k - 1) \le \exp\left(\frac{t^2}{m} a_n\right) - 1.$$

Further, there exists a $(t\sigma/2, t\sigma)$-packing change ensemble $(P, \widetilde{\mathcal{Q}})$ over p nodes such that

$$|\widetilde{\mathcal{Q}}| \ge \frac{2}{t}\left(\frac{m}{8et}\right)^{t/2}$$

and

$$\forall\, Q \in \widetilde{\mathcal{Q}}, \chi^2(Q^{\otimes n}\|P^{\otimes n}) \le (1+a_n)^t - 1.$$

We note that the proof of the above lemma constructs explicit change ensembles. We will abuse terminology and refer to *the* change ensemble or *the* packing change ensemble of Lemma B.2.3.

The above Lemma requires control on n-fold products of two distributions. However, since the χ^2-divergence is conducive to tensorisation, control for $n = 1$ is usually sufficient. The statement below captures this fact and gives an end to end lower bound on this basis. The statement amounts to collating the various facts described in this section.

Theorem B.2.4. *Let P_0 and Q_0 be as in Lemma B.2.3, and further such that $\chi^2(Q_0\|P_0) \le \kappa$. Then for $1 \le t < m/16e$, where $m = \lfloor p/\nu \rfloor$,*

$$n_{\mathrm{GoF}}(t\sigma, \mathcal{I}_d) \ge \frac{1}{2\log(1+\kappa)} \log\left(1 + \frac{m}{t^2}\right),$$
$$n_{\mathrm{EoF}}(t\sigma, \mathcal{I}_d) \ge \frac{1}{2\log(1+\kappa)} \log\left(\frac{m}{4000t}\right).$$

The 4000 in the above can be improved under mild assumptions, such as if $t \ge 8$, but we do not pursue this further. We conclude this section with proofs of the main claims above.

B.2.1.1 Proof of Lifting Lemma

Proof of Lemma B.2.3. Let G_0, H_0 be the network structures underlying P_0, Q_0, and A_0, B_0 be the weight matrices of G_0, H_0. Recall that these are graphs on ν nodes. Partition $[1:p]$ into $m+1$ pieces $(\pi_1, \pi_2, \dots, \pi_m) = ([1:\nu], [\nu+1:2\nu], \dots [(m-1)\nu+1 : m\nu])$ and $\pi_{m+1} = [m\nu+1:p]$, the last one being possibly empty. We may place a copy of G_0 on each of the first m parts, and leave the final graph disconnected to obtain a graph G with the block diagonal weight matrix $\mathrm{diag}(A_0, A_0, \dots, A_0, 0)$. We let P be the Ising model on G. For any vector $\mathbf{v} \in \{0,1\}^m$ of weight t, let $Q_{\mathbf{v}}$ be the graph which places a copy of B_0 on π_i for all $i : \mathbf{v}_i = 1$, and A_0 as before otherwise. Note the block independence across parts of π induced by this. Concretely, we have

$$P(X = x) = \prod_{i=1}^m P_0(X_{\pi_i} = x_{\pi_i}) \cdot 2^{-|\pi_{m+1}|},$$
$$Q_{\mathbf{v}}(X = x) = P(X = x) \cdot \prod_{i:\mathbf{v}_i = 1} \frac{Q_0(X_{\pi_i} = x_{\pi_i})}{P_0(X_{\pi_i} = x_{\pi_i})}.$$

Now, let $\mathcal{V}_t$ be the t-weighted section of the cube $\{0,1\}^m$, and $\mathcal{V}_t'$ be a maximal $t/2$ packing of $\mathcal{V}_t$.

We let $\mathcal{Q} := \{Q_{\mathbf{v}}, \mathbf{v} \in \mathcal{V}_t\}$ and $\mathcal{Q}' := \{Q_{\mathbf{v}}, \mathbf{v} \in \mathcal{V}_t'\}$. Since (P_0, Q_0) had symmetric difference σ, and since we introduce t differences of this form in $\mathcal{Q}$, $(P, \mathcal{Q})$ forms a $t\sigma$-change ensemble. Further, $\mathcal{Q}'$ inherits the packing structure of $\mathcal{V}_t'$, $(P, \mathcal{Q}')$ forms a $(t\sigma/2, t\sigma)$-packing change ensemble. Next note that $|\mathcal{Q}| = \binom{m}{t}$ trivially. Further, since $|\mathcal{Q}|' = |\mathcal{V}_t|$, it suffices to lower bound the latter to show that $\mathcal{Q}$ is as big as claimed. Since $\mathcal{V}_t'$ is maximal, its cardinality must exceed the $t/2$-covering number of the t-section of the cube. But then, by a volume argument,

$$|\mathcal{V}_t'| \geq \frac{\binom{m}{t}}{\sum_{k=0}^{t/2} \binom{t}{k}\binom{m-t}{k}} \geq \frac{\binom{m}{t}}{(t/2)2^t\binom{m}{t/2}} \geq \frac{2}{t}\left(\frac{m}{t}\right)^t 2^{-t}\left(\frac{2em}{t}\right)^{-t/2} = \frac{2}{t}\left(\frac{m}{8et}\right)^{t/2}$$

where we have used $t \leq m/4$.

Next, note that for any $Q_{\mathbf{v}} \in \mathcal{Q}$, and hence any $Q_{\mathbf{v}} \in \mathcal{Q}'$, we have

$$1 + \chi^2(Q^{\otimes n} \| P^{\otimes n}) = \mathbb{E}_{P^{\otimes n}} \prod_{\mathbf{v}_i = 1} \frac{Q_0^{\otimes n}}{P_0^{\otimes n}}(X_{\pi_i}^n) = \left(1 + \chi^2(Q_0^{\otimes n} \| P^{\otimes n})\right)^t.$$

Finally,

$$\begin{aligned}
1 + \chi^2(\langle \mathcal{Q}^{\otimes n} \rangle \| P^{\otimes n}) &= \frac{1}{|\mathcal{Q}|^2} \sum_{\mathbf{v}, \mathbf{v}' \in \mathcal{V}_t} \mathbb{E}_{P^{\otimes n}} \left[\frac{Q_{\mathbf{v}}^{\otimes n} Q_{\mathbf{v}'}^{\otimes n}}{(P^{\otimes n})^2}(X^n)\right] \\
&= \frac{1}{\binom{m}{t}^2} \sum_{\mathbf{v}, \mathbf{v}' \in \mathcal{V}_t} \prod_{i : \mathbf{v}_i = \mathbf{v}_i' = 1} \mathbb{E}_{P_0^{\otimes n}} \left[\frac{(Q_0^{\otimes n})^2}{(P_0^{\otimes n})^2}(X_{\pi_i}^n)\right] \\
&\leq \frac{1}{\binom{m}{t}^2} \sum_{\mathbf{v}, \mathbf{v}' \in \mathcal{V}_t} (1 + a_n)^{|\{i : \mathbf{v}_i = \mathbf{v}_i' = 1\}|} \\
&= \frac{1}{\binom{m}{t}} \sum_{j=0}^{t} \binom{t}{j}\binom{m-t}{t-j}(1 + a_n)^j.
\end{aligned}$$

Finally, note that the final expression can be written as $\mathbb{E}[(1 + a_n)^{\mathscr{H}}]$ where $\mathscr{H} \sim \mathrm{Hyp}(m, t, t)$. Since hypergeometric random variables are stochastically dominated by the corresponding binomial random variables, we may upper bound the above by the moment generating function of a $\mathrm{Bin}(t, t/m)$ random variable at $(1 + a_n)$ to yield that

$$1 + \chi^2(\langle \mathcal{Q}^{\otimes n} \rangle \| P^{\otimes n}) \leq (1 + (t/m)((1 + a_n) - 1))^t \leq \exp\left(\frac{t^2}{m} a_n\right). \qquad \square$$

B.2.1.2 Proof of Theorem B.2.4

Proof. It is a classical fact that the χ^2-divergence tensorises as

$$\chi^2(Q_0^{\otimes n} \| P_0^{\otimes n}) = (1 + \chi^2(Q_0 \| P_0))^n - 1.$$

The reason for this is that due to independence, $1 + \chi^2(Q_0^{\otimes n} \| P_0^{\otimes n})$ amounts to a product of second moments of relative likelihoods (Q/P) of iid samples.

Thus, since $\chi^2(Q_0 \| P_0) \le \kappa$, we may set $a_n = (1+\kappa)^n - 1$ in Lemma B.2.3. Now, by LeCam's method (Lemma B.2.1), we know that if $R_{\mathrm{GoF}}(t\sigma) < 1/4$ for a given n, then using ensemble from Lemma B.2.3, it must hold that

$$\begin{aligned} \frac{1}{4} &\ge 1 - \sqrt{\frac{1}{2}\log\left(1 + \exp\left(\frac{t^2}{m}a_n\right) - 1\right)} \\ \iff a_n &\ge 2(3/4)^2 \frac{m}{t^2} \\ \implies (1+\kappa)^n - 1 &\ge \frac{m}{t^2} \\ \iff n &\ge \frac{1}{\log(1+\kappa)} \log\left(1 + \frac{m}{t^2}\right) \end{aligned}$$

Thus, the smallest n for which we can test $t\sigma$-changes in $\mathcal{I}_d$ must exceed the above lower bound, giving the stated claim.

The EoF claim follows similarly. Using the packing change ensemble from Lemma B.2.3, and the lower bound Lemma B.2.2, if the risk is at most $1/4$ for some n, then we find that

$$\begin{aligned} \frac{1}{4} &\ge 1 - \frac{1}{|\widetilde{\mathcal{Q}}|} - \sqrt{\frac{(1+a_n)^t - 1}{|\widetilde{\mathcal{Q}}|}} \\ \iff (1+a_n)^t &\ge 1 + |\widetilde{\mathcal{Q}}| \left(\frac{3}{4} - \frac{1}{|\widetilde{\mathcal{Q}}|}\right)^2 \\ \iff (1+\kappa)^{nt} &\ge 1 + |\widetilde{\mathcal{Q}}| \left(\frac{3}{4} - \frac{1}{|\widetilde{\mathcal{Q}}|}\right)^2 \\ \iff n &\ge \frac{1}{t\log(1+\kappa)} \log\left(|\widetilde{\mathcal{Q}}| \left(\frac{3}{4} - \frac{1}{|\widetilde{\mathcal{Q}}|}\right)^2\right) \end{aligned}$$

Now, since $1 \le t \le m/16e$, we observe that

$$|\widetilde{\mathcal{Q}}| \ge \frac{2}{t}\left(\frac{m}{8et}\right)^{t/2} \ge \frac{2}{t} \cdot 2^{t/2} \ge 2.5.$$

Thus, $({}^{3}\!/\!{}_{4} - {}^{1}\!/\!{}_{|\widetilde{\mathcal{Q}}|})^2 \ge 1/9$, and the term in the final log above is at least $\log|\widetilde{\mathcal{Q}}|/9$, which in turn is lower bounded by Lemma B.2.3. Thus continuing the above chain of inequalities, we observe that

$$n \geq \frac{1}{\log(1+\kappa)} \cdot \frac{1}{t}\left(\frac{t}{2}\left(\log\left(\frac{m}{8et}\right) - \frac{(2\log(t/2) + 4\log(3))}{t}\right)\right)$$

Finally, since $\log(x)/(x/2) \leq 1/e$, we may $-2(\log(t/2) + 4\log(3))/t \geq -5$. Folding this -5 into the log gives $8e^6 \leq 4000$ in the denominator. Finally, again, this tells us that the infimum of the n for which the EoF risk is small is at least the above lower bound, yielding the claim. □

B.2.2 Expanded Lower Bound Theorem Statements and Proofs

We give slightly stronger theorem statements than those in the main text, and give the proofs of the claimed bounds. In all cases the proofs involve the use of Lemma B.2.3 - we describe which widgets are used, and what values of σ, t are needed. Then we simply invoke Theorem B.2.4 repeatedly to derive the results.

B.2.3 The case $d \leq s \leq cp$

Proof of Theorem 3.3.1.
High Temperature Bound This is shown by using the Triangle construction of §B.4.1.1. This construction amounts to $\sigma = 1$ and $m = \lfloor p/3 \rfloor$. Thus taking $t = s$, $\mu = \alpha, \lambda = \beta$ and invoking both Proposition B.4.1 and Theorem B.2.4, we find that so long as $p/6 \geq 16es$, the bounds

$$n_{\text{GoF}}(s, \mathcal{I}_d) \geq \frac{1}{C\tanh^2(\alpha)e^{-2\beta}} \log\left(1 + \frac{p}{Cs^2}\right),$$

and similarly

$$n_{\text{EoF}}(s, \mathcal{I}_d) \geq \frac{1}{C\tanh^2(\alpha)e^{-2\beta}} \log\left(\frac{p}{Cs}\right).$$

Low Temperature Bound Let $\beta d \geq \log d$. We show this for even d - odd d follows by reducing d by one. We use the Emmentaler clique versus the full clique of §B.4.2.3 with $\ell = 1$. This corresponds to $\sigma = d/2$ and $m = \lfloor p/d + 1 \rfloor \geq p/2d$. Now take $t = \lceil 2s/d \rceil \leq 4s/d$. Note that the total number of changes is at least s and at most $d/2\lceil 2s/d \rceil \leq 2s$. Notice that $t \leq m$ holds so long as $s \leq p/K$ for some $K \geq 400$. Invoking Proposition B.6.6 in the case of $\mu = \alpha, \lambda = \beta$, and then Theorem B.2.4 with the stated m, σ, t, gives us the bound

$$\begin{aligned} n_{\mathrm{GoF}} &\ge \frac{1}{Cd^2 \min(1, \mu^2 d^4) e^{-2\beta(d-3)}} \log\left(1 + \frac{1}{C}\frac{(p/2d)}{(4s/d)^2}\right) \\ &\ge \frac{e^{2\beta(d-3)}}{C'd^2 \min(1, \mu^2 d^4)} \log\left(1 + \frac{1}{C'}\frac{pd}{s^2}\right), \end{aligned}$$

where the $(d-3)$ in the exponent arises as $(d-1) - 1 - \ell$, and $d-1$ occurs since we may reduce d by 1 to make it even. Similarly

$$n_{\mathrm{EoF}} \ge \frac{e^{2\beta(d-3)}}{C'd^2 \min(1, \mu^2 d^4)} \log\left(1 + \frac{1}{C'}\frac{p}{s}\right).$$

Integrating the bounds. We now note that if $\beta d \le 3 \log d$, then

$$\frac{e^{2\beta(d-3)}}{d^2 \min(1, \mu^2 d^4)} \le \frac{e^{2\beta}}{\tanh^2(\alpha)}.$$

Indeed, in this case, $e^{2\beta(d-3)} \le d^6$, and so the left hand side is at most $\max(d^4, \alpha^{-2})$, which is dominated by the right hand side.

On the other hand even if $\beta d \ge 3 \log d$, we may still use the high temperature bound since this is shown unconditionally. Thus, at least so long as we replace the pd/s^2 in the low temperature bound by p/s^2, we may take the maximum of the expressions in the above bounds to get a concise lower bound - the low temperature term itself only becomes active when $\beta d \le 3 \log d$, in which case it is known to be true. The claim thus follows. □

B.2.4 The case $cp \le s \le cpd^{1-\zeta}$

We first state the commensurate EoF bound -

Theorem B.2.5. *In the setting of Theorem 3.3.2, we further have that*

1. *If $\alpha d^{1-\zeta} \le 1/32$ then $n_{\mathrm{EoF}} \ge C\dfrac{1}{d^{2-2\zeta}\alpha^2} \log\left(1 + C\dfrac{pd^{1-\zeta}}{s}\right)$.*
2. *If $\beta d \ge 4\log(d-4)$ then $n_{\mathrm{EoF}} \ge C\dfrac{e^{2\beta d(1-d^{-\zeta})}}{d^2 \min(1, \alpha^2 d^4)} \log\left(1 + C\dfrac{pd^{1-\zeta}}{s}\right)$.*

Proofs of Thms. 3.3.2 and B.2.5.

High Temperature Bounds Suppose $s = pd^{1-\zeta_0}/K$ for any $\zeta_0 \in (0, 1]$. We invoke the widget of a full $d^{1-\zeta_0}$-clique as Q_0 versus an empty graph as P_0, i.e. the construction of §B.4.1.2. This corresponds to taking $\sigma = d^{2-2\zeta_0}/2 + O(d)$, $m \ge pd^{-(1-\zeta_0)}/2$ and $t = \lfloor 2sd^{-(2-2\zeta_0)} \rfloor$, with the total edit made being at most $2s$. Invoking Proposition B.4.2 with $\mu = \alpha$, and then Theorem B.2.4 gives the bounds on noting that

$$\frac{m}{t^2} \ge C\frac{pd^{-(1-\zeta_0)}}{(sd^{-(2-2\zeta_0)})^2} = C\frac{pd^{3-3\zeta_0}}{s^2},$$
$$\frac{m}{t} \ge C\frac{pd^{-(1-\zeta_0)}}{(sd^{-(2-2\zeta_0)})} = C\frac{pd^{1-\zeta_0}}{s^2}$$

and then finally setting $\zeta_0 \ge \zeta$ to derive the claim.

Low Temperature Bounds Again fix a ζ_0. We invoke the Emmentaler clique v/s full clique widget of B.4.2.3, but this time with $\ell = d^{1-\zeta_0}$. This gives $\sigma \approx d^{2-\zeta_0}/2$, $m = \lfloor p/d \rfloor$ and $t = \lceil 2sd^{-2-\zeta_0} \rceil$. The bound now follows similarly to the above section upon invoking Propositions B.6.6 with $\lambda = \beta, \mu = \alpha$ and then Theorem B.2.4 with the stated m, t, σ. We only track the terms in the log, which are

$$\frac{m}{t^2} \ge C\frac{pd^{-1}}{(sd^{-(2-\zeta_0)})^2} = C\frac{pd^{3-2\zeta_0}}{s^2},$$
$$\frac{m}{t} \ge C\frac{pd^{-1}}{(sd^{-(2-\zeta_0)})} = C\frac{pd^{1-\zeta_0}}{s^2}. \qquad \square$$

B.2.5 Proofs in the setting $s \le d$

The catch in this section is that the Emmentaler clique construction of the proofs above can no longer be employed, since setting even $\ell = 1$ in these induces $\Omega(d)$ changes. We instead turn to the clique with a large hole construction of §B.4.2.2.

Proof of Theorem 3.3.3.

High Temperature Bound This is the same as the high temperature bound of Thm. 3.3.1, and that proof may be repeated.

Low Temperature Bound Suppose $\beta d \ge 3 \log d$. We use the clique with a large hole construction of §B.4.2.2 with the choice of $\ell = \lceil \sqrt{2s} \rceil$. This amounts to $s \le \sigma = s + O(\sqrt{s}) \le 2s$, and $m = \lfloor p/d \rfloor$. We then simply set $t = 1$ in Theorem B.2.4. Now invoking Proposition B.4.8, we find that

$$\begin{aligned} n_{\mathrm{GoF}} &\geq \frac{1}{C\sqrt{s}\sinh^2(\alpha\sqrt{s})e^{-2\beta(d-1-2\sqrt{s})}} \log\left(1 + \frac{p}{Cd}\right) \\ &\geq \frac{e^{2\beta(d-1-2\sqrt{s})}}{Cd^6 \sinh^2(\alpha\sqrt{s})} \log\left(1 + \frac{p}{Cd}\right), \end{aligned}$$

and the same lower bound for n_{EoF} since in this case $m/t^2 = m/t = 1$ (the d^6 is introduced to make the following easy).

Integrating the bounds Similarly to the proof of Thm. 3.3.1, note that for $\beta d \leq 3\log d$, $e^{2\beta d}d^{-6} \leq 1$, allowing us to rewrite the low-temperature bound as the max expression in the theorem statement. Giving up space in the logarithm to $p/s^2 \wedge p/d$ then yields the stated claim for GoF. For EoF, we follow the same procedure, but note that since $s \leq d$, $(p/s \wedge p/d) = p/d$. □

B.3 Appendix to §3.4

B.3.1 Testing Deletions in Forests, and Changes in Trees

B.3.1.1 Proofs of Lower Bounds

Proof of Lower bounds from Theorem 3.4.1. First note that $n \geq 1$ is necessary, since testing/estimation with no samples is impossible. To derive the second term in the converse for GoF and the converse for EoF, we plug in the single-edge widget of §B.4.1.4 with $\mu = \alpha$ into Theorem B.2.4. The widget corresponds to $\nu = 2$ and $\sigma = 1$. Thus, setting $t = s$ and $m = \lfloor p/2 \rfloor \geq p/3$, we obtain both the claimed bounds. □

B.3.1.2 Proof of Upper Bound of Theorem 3.4.1, and of Theorem 3.4.2

We give the proof for $\alpha > 0$. The proof for $\alpha < 0$ follows identically.

We use u as a short hand for a pair (i, j) with $i < j$, and set $Z_u = X_i X_j$. We exploit two key properties of forest structured graphs

1. For any $u = (i, j)$, if nodes i and j are connected via the graph, then $\mathbb{E}[Z_u] = \prod_{v \in \mathrm{path}(u)} \tanh(\theta_v)$, where for $u = (i, j)$ $\mathrm{path}(u)$ is the unique path connecting i and j. If i and j are not connected, then $\mathbb{E}[Z_u] = 0$.

2. For any $u \neq v$, $\mathbb{E}[Z_u Z_v] = \mathbb{E}[Z_u]\mathbb{E}[Z_v]$, that is, the Z_us are pairwise uncorrelated.

The above are standard properties, and are shown by exploiting the fact that conditioning on any node in the forest breaks it into two uncorrelated forests. See, e.g. [BK20] for proofs.

Proof of Upper Bound in Theorem 3.4.1. Recall the statistic

$$\mathscr{T} = \sum_{\ell=1}^{n} \sum_{u \in G(P)} Z_u^\ell / n,$$

where the outer sum is over samples. Suppose $G(P)$ has k edges. Let $\tau := \tanh(\alpha)$. We propose the test

$$\mathscr{T} \underset{\text{Alt}}{\overset{\text{Null}}{\gtrless}} (k - s/2)\tau.$$

Since the sum is over all edges in p, and since all edges have the same weight α, we note that

$$\mathbb{E}_P[\mathscr{T}] = k\tau.$$

Now consider an alternate Q_Δ that deletes some $\Delta \geq s$ of these edges. Since a deletion of an edge in the forest disconnects the nodes at the end of the edges (the path connecting two nodes in a forest is unique, if it exists, and we've just removed that unique path by deleting the edge),

$$\mathbb{E}_{Q_\Delta}[\mathscr{T}] = (k - \Delta)\tau.$$

Next, we consider the variance of the statistic. Due to uncorrelation of Z_us, under any forest structured Ising model we have in the case of $n = 1$

$$\mathrm{Var}[\mathscr{T}] = \sum_{u \in G(P)} (1 - (\mathbb{E}[Z_u])^2,$$

where we have used that $Z_u^2 = (\pm 1)^2 = 1$ always. Using the standard behaviour of variances under averaging of independent samples,

$$\begin{aligned}\mathrm{Var}_{P^{\otimes n}}[\mathscr{T}] &= \sum_{u \in G(P)} \frac{1-\tau^2}{n} = \frac{k(1-\tau^2)}{n},\\ \mathrm{Var}_{Q_\Delta^{\otimes n}}[\mathscr{T}] &= \sum_{u \in G(P) \cap G(Q_\Delta)} \frac{1-\tau^2}{n} + \sum_{u \in G(P) \setminus G(Q_\Delta)} 1/n = \frac{k(1-\tau^2)+\Delta\tau^2}{n}.\end{aligned}$$

Using Tchebycheff's inequality, we then observe that for a given constant $C > 1$, the following hold with probability at least $7/8$:

$$\begin{aligned}\text{Under } P^{\otimes n}: \qquad & \mathscr{T} \geq k\tau - C\sqrt{\frac{k(1-\tau^2)}{n}},\\ \text{Under any } Q_\Delta^{\otimes n}: \quad & \mathscr{T} \leq (k-\Delta)\tau + C\sqrt{\frac{k(1-\tau^2)+\Delta\tau^2}{n}}.\end{aligned}$$

Thus, the test has false alarm and size both at most $1/8$, irrespective of P and Q_Δ, so long as

$$(k-\Delta)\tau + C\sqrt{\frac{k(1-\tau^2)+\Delta\tau^2}{n}} < (k - s/2)\tau < k\tau - C\sqrt{\frac{k(1-\tau^2)}{n}}.$$

Solving out the upper bound on $(k-s/2)\tau$ yields

$$n > 4C^2\frac{k}{s^2}(\tau^{-2}-1),$$

while for the lower bound, since $\Delta \geq s$, the same must hold if

$$(k-\Delta)\tau + C\sqrt{\frac{k(1-\tau^2)+\Delta\tau^2}{n}} < (k-\Delta/2)\tau,$$

which may be rearranged to

$$n > 4C^2\left(\frac{1}{\Delta} + \frac{k}{\Delta^2}(\tau^{-2}-1)\right),$$

which in turn must hold if

$$n > 4C^2\left(1 + \frac{k}{s^2}(\tau^{-2}-1)\right),$$

where the final inequality again utilises $\Delta \geq s$.

Thus, forests with k edges can be tested with risk at most $1/4$ as long as we have at least

$$4C^2 \left(1 + \frac{k}{s^2}(\tau^{-2} - 1)\right) + 1 \leq C' \max\left(1, \frac{k}{s^2}(\tau^{-2} - 1)\right)$$

samples, where $C' \leq 8C^2 + 1$ is a constant. Since forests on p nodes have at most $p - 1$ edges, replacing k by p yields an upper bound on the sample complexity of testing deletions in forests.

Finally, since $\tau = \tanh(\alpha)$, we note that $\tau^{-2} - 1 = \sinh^{-2}(\alpha)$, concluding the proof. □

Some Observations

- While the above proof is for uniform edge weights, this can be relaxed with little change. However, the above proof does strongly rely on the edge weights all having the same sign. If this is not the case, then we may encounter edit the same number of positively and negatively weighted edges, and the statistic $\mathscr{T}$ becomes uninformative.

- The statistic $\mathscr{T}$ similarly loses power in the general setting of testing both additions and deletions in forests. This is because while the variance remains controlled as $k(1 - \tau^2)$, the means under the alternates may not move if the only changes being made are additions.

- On the other hand, if we consider testing only of full trees, i.e. P such that $G(P)$ has the full $(p - 1)$ edges, and further the altered Q are also trees, then something interesting emerges - at least in the setting of uniform weights. Since at least s edges were changed from $G(P)$ to $G(Q)$, and one cannot add an edge to $G(P)$ without introducing a cycle, it must be the case that $G(Q)$ effects at least one edge-deletion for every edge it adds, and so it must make at least $\geq s/2$ deletions. In this case, the statistic discussed above *is* powerful. This, of course, was the point of Theorem

3.4.2 in the main text, which we are now ready to prove

Proof of Upper Bound from Theorem 3.4.2. Assume that $\alpha > 0$. The proof proceeds similarly for $\alpha < 0$. We use the statistic $\mathscr{T}$ from the proof of the upper bound of Thm. 3.4.1 above, and also reuse the notation of τ, Δ and Q_Δ from the above. The claim relies on the above observation that if Δ edges are changed, then at least $\Delta/2 \geq s/2$ edges must be deleted.

In this case, the mean and the variance of $\mathscr{T}$ under P remain unchanged. On the other hand, under Q_Δ, for any edge $u \in G(P)$ that was deleted in $G(Q_\Delta)$, we must have $|\mathbb{E}_{Q_\Delta}[Z_u]| \leq \tau^2$, since the distance between the end points of these edges is now at least 2. Further, since $G(Q)$ is a tree, the variance of the statistic under Q_Δ (for $n = 1$) is

$$\begin{aligned}\mathrm{Var}_{Q_\Delta}[\mathscr{T}] &= \sum_{u \in G(P)} (1 - \mathbb{E}_{Q_\Delta}[Z_u]^2) \\ &\leq (p - 1 - \Delta)(1 - \tau^2) + \Delta \\ &= (p-1)(1-\tau^2) + \Delta\tau^2.\end{aligned}$$

At this point the argument from the earlier proof of Thm. 3.4.1 can be used. The test needs to be updated to declaring for the null only when $\mathscr{T} > (p-1)\tau - s\tau(1 - \tau)/4$. □

We conclude by showing the lower bound in Theorem 3.4.2. This requires a mild departure from the previously discussed lower bounds, in that the lifting trick is not applicable - this fundamentally constructs disconnected graphs, while trees need to be connected. However, pretty much the same approach is used.

Proof of the Lower Bound from Theorem 3.4.2. We use Le Cam's method, as before. The construction is as follows: Let p be odd, and let $m = (p-1)/2$. Take P to be the Ising model with uniform weights α on the graph with the edge set

$$G(P) = \{(p, i) : i \in [1:m]\} \cup \{(i, m+i) : i \in [1:m]\}.$$

This is a 'two-layer star' - one node is singled out as central. Half the remaining nodes are incident on it, and the other half are each incident on one of these 'inner' nodes.

Let $t = \lceil s/2 \rceil$, assumed smaller than m. For each $S \subset [1:m]$ such that $|S| = t$, we define Q_S to be the Ising model with uniform weights α on the following graph

$$G(Q_S) = \{(p,i) : i \in [1:m] \setminus S\} \cup \{(p, m+1) : i \in S\} \cup \{(i, m+i) : i \in [1:m]\}.$$

In words, Q_S detaches node i from node p and attaches node $(m+i)$ to node p for $i \in S$, thus switching some of the inner nodes to being outer and vice versa. Notice that in total, $2|S| = 2t \in \{s, s+1\}$ edges have been changed.

We directly argue that for P as defined above, and $\mathcal{Q} = \{Q_S : S \subset [1:m], |S| = t\}$, it holds that

$$\chi^2\left(\langle \mathcal{Q}^{\otimes n}\rangle, P^{\otimes n}\right) \le \exp\left(\frac{t^2}{m}\left((1 + 2\tanh^2\alpha)^n - 1\right)\right) - 1.$$

This, along with Le Cam's method implies the claim upon noting that $m/t^2 \ge 2(p-1)/(s+1)^2$ which is in turn larger than p/s^2 for $s \ge 4, p \ge 9$.

Let us proceed to show the above claim. By direct computation,

$$1 + \chi^2\left(\langle \mathcal{Q}^{\otimes n}\rangle, P^{\otimes n}\right) = \frac{1}{|\mathcal{Q}|^{\in}} \sum_{S,\tilde{S}} \left(\mathbb{E}_P\left[\frac{Q_S(X)Q_{\tilde{S}}(X)}{P(X)^2}\right]\right)^n.$$

We invoke the following calculation

Lemma B.3.1.

$$\mathbb{E}_P\left[\frac{Q_S(X)Q_{\tilde{S}}(X)}{P(X)^2}\right] \le (1 + 2\tanh^2\alpha)^{|S\cap\tilde{S}|}.$$

Let $\varphi := (1 + 2\tanh^2(\alpha))^n$. Plugging the above result into the expression for the

χ^2-divergence, we find that

$$\begin{aligned}
1 + \chi^2\left(\langle \mathcal{Q}^{\otimes n} \rangle, P^{\otimes n}\right) &\leq \frac{1}{|\mathcal{Q}|^{\in}} \sum_{S, \tilde{S}} \varphi^{|S \cap \tilde{S}|} \\
&= \sum_{k=0}^{t} \frac{\binom{t}{k}\binom{m-t}{t-k}}{\binom{m}{t}} \varphi^{k} \\
&= \mathbb{E}[\varphi^{\mathscr{H}}] \\
&\leq \mathbb{E}[\varphi^{\mathscr{B}}] \\
&= \left(1 + \frac{t}{m}(\varphi - 1)\right)^{t} \leq \exp\left(\frac{t^2}{m}(\varphi - 1)\right),
\end{aligned}$$

where we have used the fact that $\mathcal{Q}$ is parametrised by all subsets of size t of a set of size m, and then proceeded similarly to the proof of the first part in Lemma B.2.3, with $\mathscr{H}$ being a (m, t, t)-hypergeometric random variable, and $\mathscr{B}$ being a $(t, t/m)$-binomial random variable. It remains to show the above Lemma, which is argued below. □

Proof of Lemma B.3.1. Notice that

$$\begin{aligned}
P(x) &= \frac{1}{2^p \cosh^{p-1}(\alpha)} \exp\left(\alpha\left(x_p \sum_{i=1}^{m} x_i + \sum_{i=1}^{m} x_i x_{m+i}\right)\right) \\
Q_S(x) &= \frac{1}{2^p \cosh^{p-1}(\alpha)} \exp\left(\alpha\left(x_p \sum_{i \in S^c} x_i + x_p \sum_{i \in S} x_{m+i} + \sum_{i=1}^{m} x_i x_{m+i}\right)\right)
\end{aligned}$$

Where the partition functions are directly calculated. As a consequence,

$$\begin{aligned}
&2^p \cosh^{p-1}(\alpha)\mathbb{E}_P\left[\frac{Q_S(X)Q_{\tilde{S}}(X)}{P(X)^2}\right] \\
&\quad = 2^p \cosh^{p-1}(\alpha)\sum_x \frac{Q_S(x)Q_{\tilde{S}}(x)}{P(x)} \\
&\quad = \sum_x \exp\left(\alpha\left(x_p \sum_{i\in(S\cup\tilde{S})^c} x_i + \sum_{i\in(S\cup\tilde{S})^c} x_i x_{m+i}\right)\right) \\
&\qquad\quad \times \exp\left(\alpha\left(x_p \sum_{i\in S\triangle\tilde{S}} x_{m+i} + \sum_{i\in S\triangle\tilde{S}} x_i x_{m+i}\right)\right) \\
&\qquad\qquad\quad \times \exp\left(\alpha\left(x_p \sum_{i\in S\cap\tilde{S}} (2x_{m+i} - x_i) + \sum_{i\in S\cap\tilde{S}} x_i x_{m+i}\right)\right).
\end{aligned}$$

Observe that upon fixing a value of x_p, the product above completely decouples into m groups over (x_i, x_{m+i}), which can then be summed separately. Indeed,

$$\begin{aligned}
&2^p \cosh^{p-1}(\alpha)\mathbb{E}_P\left[\frac{Q_S(X)Q_{\tilde{S}}(X)}{P(X)^2}\right] \\
&\quad = \sum_{x_p} \prod_{i\in(S\cup\tilde{S})^c}\left(\sum_{x_i,x_{m+i}} \exp\left(\alpha(x_p x_i + x_i x_{m+i})\right)\right) \\
&\qquad\quad \times \prod_{i\in S\triangle\tilde{S}}\left(\sum_{x_i,x_{m+i}} \exp\left(\alpha(x_p x_{m+i} + x_i x_{m+i}\right)\right) \\
&\qquad\qquad\quad \times \prod_{i\in S\cap\tilde{S}}\left(\sum_{x_i,x_{m+i}} \exp\left(\alpha(x_p(2x_{m+i} - x_i) + x_i x_{m+i}\right)\right).
\end{aligned}$$

There are three types of i - those that lie in neither of $S, \tilde{S}$, those that lie in only one of these, and those that lie in both, which is how the above has been separated. We will explicitly compute the sum over (x_i, x_{m+i}) for each type separately.

1. $i \in (S \cup \tilde{S})^c$:

$$\begin{aligned}&\sum_{x_i, x_{m+i}} \exp\left(\alpha(x_p x_i + x_i x_{m+i})\right)\\ &\quad = e^{\alpha(x_p+1)} + e^{\alpha(x_p-1)} + e^{\alpha(-x_p-1)} + e^{\alpha(-x_p+1)}\\ &\quad = 2e^{\alpha}\cosh(\alpha x_p) + 2e^{-\alpha}\cosh(\alpha x_p)\\ &\quad = 4\cosh^2(\alpha),\end{aligned}$$

   where we have utilised the fact that $x_p \in \pm 1$, and that cosh is an even function.

2. $i \in S \triangle \tilde{S}$: This case is very similar to the above:

$$\begin{aligned}&\sum_{x_i, x_{m+i}} \exp\left(\alpha(x_p x_{m+i} + x_i x_{m+i})\right)\\ &\quad = e^{\alpha(x_p+1)} + e^{\alpha(-x_p-1)} + e^{\alpha(x_p-1)} + e^{\alpha(-x_p+1)}\\ &\quad = 4\cosh^2(\alpha)\end{aligned}$$

3. Finally, for $i \in S \cap \tilde{S}$,

$$\begin{aligned}&\sum_{x_i, x_{m+i}} \exp\left(\alpha(x_p(2x_{m+i} - x_i) + x_i x_{m+i}\right)\\ &\quad = e^{\alpha(x_p+1)} + e^{\alpha(-3x_p-1)} + e^{\alpha(3x_p-1)} + e^{\alpha(-x_p+1)}\\ &\quad = 2(e^{\alpha}\cosh(\alpha) + e^{-\alpha}\cosh(3\alpha)),\end{aligned}$$

Plugging the above calculations in, we find that

$$\begin{aligned}&\mathbb{E}_P\left[\frac{Q_S(X)Q_{\tilde{S}}(X)}{P(X)^2}\right]\\ &\quad = \sum_{x_p} \frac{(4\cosh^2\alpha)^{|(S\cup\tilde{S}^c|+|S\triangle\tilde{S}|}(2(e^{\alpha}\cosh(\alpha) + e^{-\alpha}\cosh(3\alpha)))^{|S\cap\tilde{S}|}}{2^p\cosh^{p-1}(\alpha)}\\ &\quad = 2\cdot\frac{(2\cosh(\alpha))^{2(m-|S\cap\tilde{S}|}(2(e^{\alpha}\cosh(\alpha) + e^{-\alpha}\cosh(3\alpha)))^{|S\cap\tilde{S}|}}{2^p\cosh^{p-1}(\alpha)}\\ &\quad = \left(\frac{e^{\alpha}\cosh(\alpha) + e^{-\alpha}\cosh(3\alpha)}{2\cosh^2(\alpha)}\right)^{|S\cap\tilde{S}|},\end{aligned}$$

where we have used the fact that $(S \cup \tilde{S})^c, S\triangle\tilde{S}, S \cap \tilde{S}$ form a partition of $[1:m]$, and that $2m = p - 1$.

To finish, we observe that

$$\begin{aligned}
&e^x \cosh(x) + e^{-x}\cosh(3x) - 2\cosh^2(x)\\
&\quad = \frac{e^{2x} + e^{-4x} - e^{-2x} - 1}{2}\\
&\quad = e^{-x}\frac{e^{(}3x) + e^{-3x} - (e^x + e^{-x})}{2}\\
&\quad = e^{-x}(\cosh(3x) - \cosh(x))\\
&\quad = e^{-x}(4\cosh^3(x) - 3\cosh(x) - \cosh(x))\\
&\quad = 4e^{-x}\cosh(x)(\cosh^2(x) - 1)\\
&\quad \le 4\sinh^2(x),
\end{aligned}$$

where the final relation uses $x \ge 0$. □

B.3.1.3 Tolerant Testing of Forest Deletions, and of Trees

Proof of Theorem 3.4.3. We repeatedly reuse the notation from the proof of Theorem 3.4.1 above.

For the forest deletion setting, suppose $|G(P)| = k$, and let $\widetilde{P}_{\Delta_0}$ be such that it's network structure is a deletion of most $\Delta_0 \le \varepsilon s$ edges from $G(P)$. It follows from the mean and variance calculations before, that, for any $\Delta \ge s$,

$$\begin{aligned}
\mathbb{E}_{\widetilde{P}_{\Delta_0}^{\otimes n}}[\mathscr{T}] &= (k - \Delta_0)\tau \ge (k - \varepsilon s)\tau,\\
\mathrm{Var}_{\widetilde{P}_{\Delta_0}^{\otimes n}}[\mathscr{T}] &= \frac{k(1-\tau^2) + \Delta_0\tau^2}{n} \le \frac{k(1-\tau^2) + \Delta\tau^2}{n}.
\end{aligned}$$

Consider the test which rejects the null hypothesis when $\mathscr{T} < (k - \frac{1+\varepsilon}{2}s)\tau$. Comparing the above to a Q_Δ as in the proof of Theorem 3.4.1, and proceeding as in it, we find that the risk is appropriately controlled so long as the following relations hold for every $\Delta_0 \le \varepsilon s$, and $\Delta \ge s$, where C is an absolute constant:

$$\begin{aligned}
n &\ge C\frac{k(\tau^{-2} - 1) + \Delta_0}{\left(\frac{1+\varepsilon}{2}s - \Delta_0\right)^2}\\
n &\ge C\frac{k(\tau^{-2} - 1) + \Delta}{\left(\Delta - \frac{1+\varepsilon}{2}s\right)^2}
\end{aligned}$$

The right hand sides of the first and second equations above respectively increase and decrease with Δ_0 and Δ. Thus, setting $\Delta_0 = \varepsilon s$ and $\Delta = s$, and taking the maximum possible $k = p$ tells us that the conditions will be met so long as

$$n \geq 4C\frac{(p-1)\sinh^{-2}(\alpha) + s}{(1-\varepsilon)^2 s^2}$$

For the tree case, the same argument follows but with a small change - in the null case, a change of Δ_0 edges can reduce the mean of $\mathscr{T}$ by $\Delta_0\tau$, but in the alternate, there may exist changes of Δ edges which only drop the mean of $\mathscr{T}$ by $\Delta/2(\tau - \tau^2)$. Thus, we use the test

$$\mathscr{T} \underset{\text{Alt.}}{\overset{\text{Null}}{\gtrless}} (p-1)\tau - \frac{1+2\varepsilon}{4}s\tau + \frac{s}{4}\tau^2.$$

Continuing similarly, and keeping in mind that the variance of $\mathcal{T}$ after Δ changes is at most $(p-1)(1-\tau^2) + \Delta\tau^2$, we find that risk of the above test is controlled so long as for every $\Delta_0 \leq \varepsilon s$, and for every $\Delta \geq s$, the following relations hold

$$\begin{aligned} n &\geq \frac{C}{s^2}\frac{p(\tau^{-2}-1) + \Delta_0}{\left(1 + 2\varepsilon - \tau - 4\Delta_0/s)\right)^2} \\ n &\geq \frac{C}{s^2}\frac{p(\tau^{-2}-1) + \Delta}{\left(2\Delta/s(1-\tau) - (1 + 2\varepsilon - \tau)\right)^2} \end{aligned}$$

It is a matter of straightforward computation that if $\varepsilon \leq \frac{1-\tau}{2}$, then the right hand sides of the first and second inequality above respectively increase and decrease with Δ_0 and Δ. Thus, setting $\Delta_0 = \varepsilon s$ and $\Delta = s$, the above holds if

$$n \geq \frac{C}{(1-2\varepsilon-\tau)^2}\left(\frac{p(\tau^{-2}-1)}{s^2} + \frac{1}{s}\right). \qquad \square$$

B.3.2 Testing Deletions in High-Temperature Ferromagnets

B.3.2.1 Proof of achievability

Proof of the upper bound of Theorem 3.4.4. We follow the strategy laid out in the main text. The proposed test statistic is $\mathscr{T}(\{X^{(i)}\}; P) := \widehat{\mathbb{E}}[\sum_{(i,j)\in G(P)} X_iX_j]$, where the $\{X^{(i)}\}$ are the samples, and $\widehat{E}$ indicates the empirical mean over this data.

Concretely, the test is to threshold $\mathscr{T}$ as

$$\mathscr{T} \underset{\text{Alt.}}{\overset{\text{Null}}{\gtrless}} \mathbb{E}_P[\mathscr{T}] - Cs\alpha,$$

where C the constant left implicit in Lemma B.3.2.

The analysis relies on two facts:

Lemma B.3.2. *Let $P, Q \in \mathcal{H}_d^\eta(\alpha)$, and $G(Q) \subset G(P)$, with $|G(P) \triangle G(Q)| \geq s$. For every $\eta < 1$, there exists a constant $C > 0$ that does not depend on (p, s, α) such that*

$$\mathbb{E}_P[\mathscr{T}] - \mathbb{E}_Q[\mathscr{T}] \geq 2Cs\alpha.$$

Lemma B.3.3. *For any $P, Q \in \mathcal{H}_d^\eta(\alpha)$, which may be equal,*

$$\mathrm{Var}_Q\left[\sum_{(i,j)\in G(P)} X_i X_j\right] \leq C_\eta pd,$$

where C_η may depend on η, but not otherwise on (p, d, s, α).

Applying the variance contraction over n independent samples, we find via a use of Tchebycheff's inequality that the following event have probability at least $1/8$ for the respective hypotheses:

$$\begin{aligned} \textit{Null:} \quad & \mathscr{T} \geq \mathbb{E}_P[\mathscr{T}] - C_\eta\sqrt{\frac{8pd}{n}}, \\ \textit{Alt:} \quad & \mathscr{T} \leq \mathbb{E}_P[\mathscr{T}] - Cs\alpha + C_\eta\sqrt{\frac{8pd}{n}}. \end{aligned}$$

Thus, taking n so large that $Cs\alpha > C_\eta\sqrt{\frac{8pd}{n}}$, the false alarm and missed detection probabilities are both controlled below $1/8$, yielding the claimed result. □

It of course remains to argue the above lemmata. These are both essentially utilisations of existing results.

Proof of Lemma B.3.2. We use the fact that in ferromagnetic models, the correlations between any pair of nodes increases as the weights increase (or contrapositively, if weights are deleted, then correlations must decrease). This is classically shown via (a special case of) Griffith's inequality [Gri69], which claims that for any u, v, i, j, in a ferromagnetic Ising model, $\mathbb{E}[X_u X_v X_i X_j] \geq \mathbb{E}[X_u X_v]\mathbb{E}[X_i X_j]$. This is relevant here due to the fact that

$$\begin{aligned}
&\partial_{\theta_{ij}} \mathbb{E}_{P_\theta}[X_u X_v] \\
&\quad = \partial_{\theta_{ij}} \frac{1}{Z_\theta} \sum_x x_u x_v \exp\left(\sum_{s<t} \theta_{st} X_s X_t\right) \\
&\quad \overset{a}{=} \frac{1}{Z_\theta} \sum_x x_u x_v x_i x_j \exp\left(\sum_{s<t} \theta_{st} X_s X_t\right) \\
&\qquad - \frac{1}{Z_\theta^2} \left(\sum_x x_u x_v \exp\left(\sum_{s<t} \theta_{st} X_s X_t\right)\right) \left(\sum_x x_u x_v \exp\left(\sum_{s<t} \theta_{st} X_s X_t\right)\right) \\
&\quad = \mathbb{E}[X_u X_v X_i X_j] - \mathbb{E}[X_u X_v] \mathbb{E}[X_i X_j] \geq 0.
\end{aligned}$$

Above, equality (a) is a consequence of the quotient rule, and the fact that $Z_\theta = \sum_x \exp\left(\sum_{s<t} \theta_{st} x_s x_t\right)$.

Next, we utilise the following structural lemma, due to Santhanam and Wainwright. While we cite it as a variation on their Lemma 6 below, more accurately this arises via a correction of a subsidiary part of the proof of the same lemma. In particular, we are utilising a corrected version of the unlabelled inequality on Page 4131 that follows the inequality (51), with further specialisation to the high-temperature deletion with a uniform edge weight context.

Lemma B.3.4. (A variation of Lemma 6 of [SW12]) *Let $P \in \mathcal{H}_d^\eta(\alpha)$, and Q be obtained by removing the edge (a, b) from P. Then*

$$\mathbb{E}_P[X_a X_b] - \mathbb{E}_Q[X_a X_b] \geq \frac{\alpha}{400}.$$

With this in hand, we develop our result by arguing over each deleted edge in a sequence. For a given P and Q, such that Q occurs by deleting $\Delta \geq s$ edges from P, take a chain of laws $P = Q_0, Q_1, Q_2, \ldots, Q_\Delta = Q$, where each Q_{t+1} is obtained by deleting one edge from Q_t. Let (i_{t+1}, j_{t+1}) be the edge deleted in going from Q_t to Q_{t+1} Since each model is ferromagnetic, and each Q_{t+1} deletes an edge from Q_t, we find that

$$\begin{aligned}
&\mathbb{E}_{Q_t}\left[\sum_{(i,j)\in G(P)} X_iX_j\right] - \mathbb{E}_{Q_{t+1}}\left[\sum_{(i,j)\in G(P)} X_iX_j\right]\\
&\geq \mathbb{E}_{Q_t}\left[X_{i_{t+1}}X_{j_{t+1}}\right] - \mathbb{E}_{Q_{t+1}}\left[X_{i_{t+1}}X_{j_{t+1}}\right]\\
&\geq \frac{\alpha}{400}.
\end{aligned}$$

Summing up the left hand side over $t = 0$ to $\Delta - 1$ leads to a telescoping sum, while $\Delta \geq s$ copies of the right hand side get added, directly leading to our conclusion

$$\begin{aligned}
\mathbb{E}_P\left[\sum_{(i,j)\in G(P)} X_iX_j\right] - \mathbb{E}_Q\left[\sum_{(i,j)\in G(P)} X_iX_j\right]&\\
&= \mathbb{E}_{Q_0}\left[\sum_{(i,j)\in G(P)} X_iX_j\right] - \mathbb{E}_{Q_\Delta}\left[\sum_{(i,j)\in G(P)} X_iX_j\right]\\
&= \sum_{t=0}^{\Delta-1}\mathbb{E}_{Q_t}\left[\sum_{(i,j)\in G(P)} X_iX_j\right] - \mathbb{E}_{Q_{t+1}}\left[\sum_{(i,j)\in G(P)} X_iX_j\right]\\
&\geq \sum_{t=0}^{\Delta-1}\frac{\alpha}{400} = \Delta\frac{\alpha}{400} \geq s\frac{\alpha}{400}.
\end{aligned}$$

□

To complete the proof, we prove the key lemma used in the above argument.

Proof of Lemma B.3.4. We note that this proof assumes familiarity with the proof of Lemma 6 of [SW12]. The main reason is that the proof really consists of fixing an equation in the proof of this result, and then utilising the ferromagnetic properties a little. As a result, there is no neat way to make this proof self contained (reproducing the proof of the aforementioned lemma is out of the question, since this is a long and technical argument in the original paper). With this warning out of the way, let us embark.

Let ∂a and ∂b be the neighbours of, respectively, a and b in $G(P)$ (which, since $G(Q)$ only deletes (a, b) from $G(P)$, contain all the neighbours of a and b in $G(Q)$ as well).

Before proceeding, we must first point out a (small) error in the proof of Lemma 6 in [SW12]. The clearest way to see this error is to note the inequality following equation (51) in the text, which claims that if $(a, b) \in G(P)\triangle G(Q)$, then some quantity ($J$ in the paper) known to be positive is upper bounded by

$$J \leq \sum_{u \in \partial a \setminus \{b\}} (\{\mathbb{E}_P - \mathbb{E}_Q\}[X_u X_a])(\theta^P_{ua} - \theta^Q_{ua}) + \sum_{v \in \partial b \setminus \{a\}} (\{\mathbb{E}_P - \mathbb{E}_Q\}[X_v X_b])(\theta^P_{vb} - \theta^Q_{vb}).$$

Note that we have specialised the above to the case where $G(Q) \subset G(P)$. Now, observe than when the only change made is in the edge (a, b), then the above upper bound is 0. Indeed, $\theta^P_{ua} = \theta^Q_{ua}$ for every $u \in \partial a \setminus \{b\}$, since none of these edges have been altered, making the first sum 0, and similarly the second, contradicting the claim that the sum is bigger than J (which is positive). The error actually lies a few lines up, in the decomposition for the term $\Delta(\theta, \theta')$, which along with the claimed terms, should also include the term $(\{\mathbb{E}_P - \mathbb{E}_Q\}[X_a X_b])(\theta^P_{ab} - \theta^Q_{ab})$, which is missing from the text of [SW12]. This term is present since the $P_{\theta[x_C]}$ and $P_{\theta'[x_C]}$ are, of course, laws on X_a and X_b, and thus have $\theta^P_{ab} x_a x_b$ and $\theta^Q_{ab} x_a x_b$ in the Ising potentials.[1] Putting this term back in, the correct equation is that

$$\begin{aligned}\kappa \leq (\{\mathbb{E}_P - \mathbb{E}_Q\}[X_a X_b])(\theta^P_{ab} - \theta^Q_{ab}) + \sum_{u \in \partial a \setminus \{b\}} (\{\mathbb{E}_P - \mathbb{E}_Q\}[X_u X_a])(\theta^P_{ua} - \theta^Q_{ua}) \\ + \sum_{v \in \partial b \setminus \{a\}} (\{\mathbb{E}_P - \mathbb{E}_Q\}[X_v X_b])(\theta^P_{vb} - \theta^Q_{vb}),\end{aligned}$$

where κ is the lower bound on J, that is (specialised to our case of uniform weights),

$$\kappa = \frac{\sinh^2(\alpha/4)}{1 + 3\exp(\alpha d)}.$$

We note that the conclusion of Lemma 6 of [SW12] is not affected by the above error[2].

With this out of the way, we may now argue our point. In our case, we know that since only the edge (a, b) has been altered, the second and third terms in the updated sum are 0. Further, we know that $\theta^P_{ab} = \alpha \geq 0$, and $\theta^Q_{ab} = 0$. Thus, we conclude that

$$\mathbb{E}_P[X_a X_b] - \mathbb{E}_Q[X_a X_b] \geq \frac{\kappa}{\alpha} \geq \frac{\sinh^2 \alpha/4}{\alpha(1 + 3\exp(2\alpha d))}.$$

[1] note however that exactly one of θ^P_{ab} and θ^Q_{ab} is zero, since (a, b) lies in one but not the other graph.

[2] The expression $2\alpha d \max_{u \in \{a,b\}, v \in V} |\mu_{uv} - \mu'_{uv}|$ already accounts for the extra term we add, since it allows us to take $u = a, v = b$.

Finally, we use our high temperature condition. Firstly, note that $\alpha d \le \eta < 1$, and thus $(1 + 3\exp(2\alpha d)) \le 1 + 3e^2 \le 24$. Next, since $\sinh(x) \ge x$, $\sinh^2(\alpha/4) \ge \alpha^2/16$. Putting these together, we find that

$$\mathbb{E}_P[X_a X_b] - \mathbb{E}_Q[X_a X_b] \ge \frac{\alpha^2/16}{\alpha \cdot 24} = \frac{\alpha}{384} \ge \frac{\alpha}{400} \qquad \square$$

Proof of Lemma B.3.3. We directly utilise the concentration result [AKPS19, Ex. 2.5], which shows that for bilinear forms $f(X) = \langle A, XX^\top \rangle$, where the inner product is the Frobenius dot product, and for a high temperature Ising model P, there exists a C_η depending only on η such that[3]

$$P(|f - \mathbb{E}[f]| \ge t) \le 2\exp\left(-\frac{t}{C_\eta \|A\|_F}\right).$$

Via the standard integral representation $\mathbb{E}[(f - \mathbb{E}[f])^2] = \int_0^\infty P(|f - \mathbb{E}[f]|^2 \ge r)\mathrm{d}r$ and the above upper bound, we directly obtain that the variance of any f such as the above is bounded by $3\|A\|_F^2 C_\eta^2$.

Now, out statistic is a bilinear function of the above form. Indeed,

$$\sum_{(i,j) \in G(P)} X_i X_j = \langle G(P)/2, XX^\top \rangle,$$

where we treat $G(P)$ as it's adjacency matrix, and thus we immediately obtain that the variance is bounded by $1.5 C_\eta^2 \|G(P)\|_F^2$. Notice that $\|G(P)\|_F^2$ is merely twice the number of edges in $G(P)$, and since this has degree at most d, this number is at most $2pd$. The claim follows. $\square$

B.3.2.2 Proof of Lower Bounds

The lower bounds are argued using Thm. B.2.4, with the widget(s) that consist of comparing a full clique to an empty graph, which of course satisfy the constraint that the alternate models are derived by deleting edges from the null graph. Concretely, we use the bound of Proposition B.4.3, to show the following result

Proposition B.3.5. *Suppose $s \le pd/K$ for large enough K and $\alpha d \le \eta \le 1/32$.*

[3] Instead of the Frobenius norm $\|A\|_F$, the bound of [AKPS19] features the Hilbert-Schmidt norm of A. These are the same thing for finite dimensional operators.

Then there exists a C independent of all parameters such that

$$n_{\mathrm{GoF,del}}(s, \mathcal{H}_d^\eta(\alpha)) \ge \max_{s/Kp \le k \le d} \frac{1}{Ck^2\alpha^2} \log\left(1 + \frac{pk^3}{Cs^2}\right),$$
$$n_{\mathrm{GoF,del}}(s, \mathcal{H}_d^\eta(\alpha)) \ge \max_{s/Kp \le k \le d} \frac{1}{Ck^2\alpha^2} \log\left(1 + \frac{pk}{Cs}\right),$$

where the maximisation is over integers $k \ge 2$ in the stated ranges. In particular, the bounds in the main text correspond to taking $k = d$.

Proof. The proof relies on the fact that if $\alpha d \le 1/32$, then $\alpha k \le 1/32$ for any $k \le d$ as well, which allows us to utilise Prop. B.4.3 for each k. For each valid choice of k, we take P_0 to be the Ising model on the complete graph on k nodes with uniform edge weight α, and Q_0 to be the Ising model on the empty graph on k nodes. The relevant quantities are $\sigma = \binom{k}{2}$, $m = \lfloor p/k \rfloor$, and $t = \lceil s/\binom{k}{2} \rceil$, with the total number of changes lying between s and $2s$. Repeated use of Thm.B.2.4 concludes the argument. □

B.3.3 Simulation Details

Details about the generation of Figure 3·3 are as follows:

- **Sampling from Ising Models** Samples from Ising models were generated by running Glauber dynamics for 1600 steps. This number is chosen to be four times the ‘autocorrelation time’, which is the time at which the autocorrelation of the test statistic $\langle XX', G\rangle/2$ drops to near 0, and serves as a proxy for the mixing time of the dynamics (at least for the relevant statistics). Note that raw samples were outputted from the dynamics (i.e., we did not take ergodic averages).

- **Constructing Ps and Qs** Throughout, P was the Ising model on a complete binary tree on 127 nodes. For each value of s and each experiment, s random edges from this tree were deleted.

- **Experiment Structure** For each $s \in \{3, 6, \dots, 60\}$ and $n \in \{20, 40, \dots, 480\}$, we carried out a simulation of the GoF testing risk of our statistic for s deletions using n samples. We refer to each of these as an experiment. Each experiment was carried out by running 100 independent tests (on independent data), which each consisted of two parts - first we generated samples from P, and declared a false alarm if $\mathscr{T}$ fell below $(p-1-s/2)\tanh(\alpha)$ for this. Next, we generated a Q by deleting s edges, and then generated samples from Q, and finally declared a missed detection if $\mathscr{T}$ was above the same threshold. Risks were computed by adding up the total number of false alarm and missed detection events in these 100 runs, and dividing them by 100.

- **Structure of Figure 3·3** Each box in the figure corresponds to a simulation for s changes and n nodes, where (s, n) are the coordinates of the upper right corner of the box. The boxes are coloured according to their empirical risk - if this was greater than 0.35, then the box was coloured black; if smaller than 0.15, then coloured white, while if it was between these values, the box was coloured orange.

Additionally, we note that structure learning performs very poorly for this setup. This is best illustrated by the Figure B·1, which shows the number of edge-errors (i.e. $|G(P)\triangle\hat{G}|$) versus the sample size when the Chow-Liu algorithm was run on data generated by the null model (i.e., the full binary tree). The Chow-Liu algorithm was run by computing the covariance matrix, and computing the weighted maximum spanning tree for it via the library methods in MATLAB. The number of errors is again averaged over 100 trials. This demonstrates that the naïve scheme of recovering the graph and testing against it is infeasible for $s \leq 60$ if $n \leq 1500$, empirically demonstrating the separation between structure learning and testing.

B.4 Widgets

As discussed in the previous section, we will utilise Lemma 3.3.4, in order to do which we need to provide specific instances of (P_0, Q_0) that are close in χ^2-divergence. We will abuse terminology and call this pair an ensemble. This section lists a few such

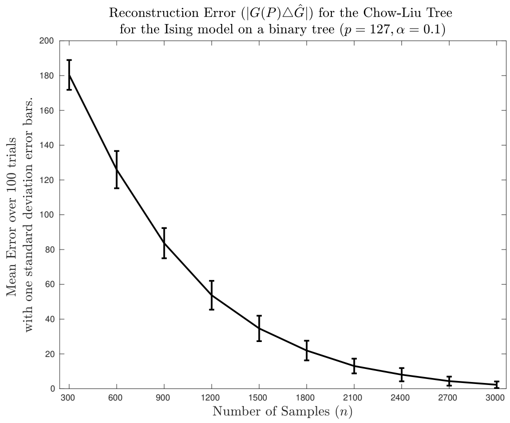


Figure B·1: Reconstruction Error of the Chow-Liu Tree for the Ising model on a complete Binary Tree with $p = 127, \alpha = 0.1$.

pairs of graphical models, along with the χ^2-divergence control we offer for the same, proofs for which are left to §B.6. Throughout, we will use λ and μ as weights of edges, with $\lambda \geq |\mu| > 0$. I the proofs of the theorems, we will generally set $\lambda = \beta$ and $\mu = \alpha$, but retaining these labels aids in the proofs of χ^2-divergence control offered for these widgets.

B.4.1 High-Temperature Obstructions

The following graphs are used to construct obstructions in high temperature regimes. The first is the triangle graph, as described in §3.3.1. The second is a full clique in high temperatures. The latter section is derived from the bounds of [CNL18].

B.4.1.1 The Triangle

We start simple. Let P_{Triangle} be the Ising model on 3 nodes with edges $(1, 2)$ and $(2, 3)$, each with weight λ, and Q_{Triangle} be the same with the edge $(1, 3)$ of weight μ appended (see Figure B·2). The bound below follows from an explicit calculation, which is tractable in this small case.

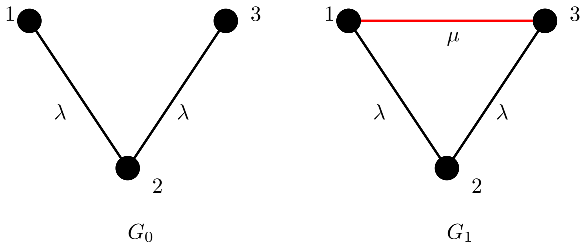


Figure B·2: Ensemble used for Proposition B.4.1

Proposition B.4.1. *For $\lambda \geq |\mu| > 0$,*

$$\chi^2(Q_{\text{Triangle}} \| P_{\text{Triangle}}) \leq 8e^{-2\lambda} \tanh^2 \mu.$$

B.4.1.2 Full Clique versus Empty Graph

[CNL18] shows the remarkable fact that high-temperature cliques are difficult to separate from the empty graph. We present this result below.

Proposition B.4.2. *Let P be the Ising model on the empty graph with k nodes, and let Q be the Ising model on the k-clique, with uniform edge weights μ. If $32\mu k \leq 1$, then*

$$\chi^2(Q\|P) \leq 3k^2\mu^2.$$

In the notation of [CNL18], this is the bound at the bottom of page 22, instantiated with $G = G'$ and the $\mathcal{R}, \mathcal{B}, \Gamma$ values as determined in the proof of Example 2.7.

We will also utilise the following reversed χ^2-divergence bound. This is not formally shown in [CNL18], and thus, we include a proof of the same, using the techniques of the cited paper, in §B.6.2.5.

Proposition B.4.3. *Let P be the Ising model on a clique on m nodes with uniform edge weights μ, and let Q be the Ising model on the empty graph on m nodes. If $32\mu m \leq 1$, then*

$$\chi^2(Q\|P) \leq 8(\mu m)^2.$$

B.4.1.3 Fan Graph

This widget is not required for the main text, although it may serve as a more involved construction to show the bounds of Thms. 3.3.1 and 3.3.3. Its main use is in Appendix B.5.2, where it is used to show an obstruction to testing of maximum degree in a graph.

Generalising the triangle of the previous section, we may hang many triangles from a single vertex, getting a graph that resembles an axial fan with many blades. Using such a graph, we may demonstrate high-temperature obstructions to determining the maximum degree of a graph.

Concretely, for a natural B we define a fan with B blades to be the graph on $2B+1$ nodes where, nodes $[1:2B]$ are each connected to the central node $2B+1$, and further, for $i \in [1:B]$, nodes $2i$ and $2i-1$ are connected. We call the edges incident on the central node $(B+1)$ axial, and the remaining edges peripheral.

Treating ℓ as a parameter, the Ising models $P_{\ell,\text{Fan}}$ and $Q_{\ell,\text{Fan}}$ are determined as followed:

- $Q_{\ell,\text{Fan}}$ places a weight λ on each peripheral edge, and a weight of μ on each axial edge.
- $P_{\ell,\text{Fan}}$ 'breaks in half' ℓ of the blades in the graph - concretely, for $i \in [1:\ell]$, we delete the edges $\{2i-1, 2B+1\}$.

Viewing P as the null graph, note that in Q we have added an excess of ℓ edges, and increased the degree of the central node from $2B-\ell$ to $2B$. The fan graph serves as a high-temperature obstruction to determining the maximum degree of the graph underlying an Ising model via the following claim.

Proposition B.4.4. *For $\ell \le B$, if $\lambda\mu \ge 0$, then*

$$\chi^2(Q_{\ell,\text{Fan}} \| P_{\ell,\text{Fan}}) \le \left(1 + 16e^{-2\lambda} \tanh^2 \mu\right)^\ell - 1.$$

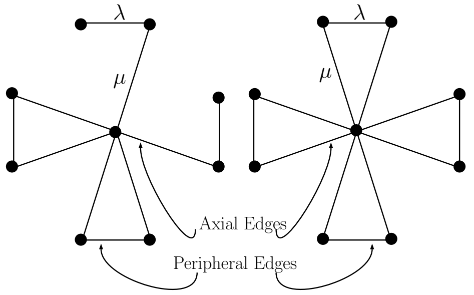


Figure B·3: The Fan graphs for $P_{\ell,\text{Fan}}$ (left) and $Q_{\ell,\text{Fan}}$ (right) in the setting $B = 4, \ell = 2$.

B.4.1.4 Single Edge

This construction is possibly the simplest, and is used to show the lower bound in Thm. 3.4.1. We consider the two possible Ising models on two nodes - P is the one with an edge, of weight μ, while Q has no edges. The characterisation is trivial, and we omit the proof.

Proposition B.4.5. $\chi^2(Q\|P) = \sinh^2(\mu)$.

B.4.2 Low-Temperature Obstructions via Clique-Based Graphs

The computations in this and the subsequent cases are rather more complicated that in the previous case, and will intimately rely on a 'low temprature' assumption. The basic unit is that of a clique on some $d + 1 \gg 1$ nodes, in the setting of temperature $\lambda d \geq \log d$.

The intuition behind these is rather simple - Ising models on cliques tend to 'freeze' in low temprature regimes, i.e. the distribution concentrates to the states $\pm(1, 1, \dots, 1)$ with probability $1-\exp(-\Omega(\beta d))$ for $\beta d \gg 1$. This effect is fairly robust, and dropping or adding even a large number of edges does not alter it significantly. Thus, one has to collect an exponential in βd number of samples merely to obtain some diversity in the samples, which will be necessary to distinguish any of these variations of a clique from the full thing.

While generic arguments can be offered for each of the settings below on the basis of the above intuition, these tend to be lossy in how they handle the effect of very low edge weights. To counteract this, we individually analyse each setting, and while the arguments have structural similarities, the particulars vary.

It is worth noting that our bounds rely on below diverge from the classical literature in the low temperature condition we impose. We generally demand conditions like $\beta d \geq \log d$, while most other lower bounds demand that $\beta d \geq 1$. This extra room allows us to tighten the exponents in the sample complexity bounds as opposed to previous work, but has the obvious disadvantage of reduced applicability. We note, however, that in most settings, this only yields a lost factor of d in the resulting bounds, and frequently not even that. Functionally, thus, there is little to no loss in the use of this stronger low-temperature condition.[4] A similar notion of low temperature has appeared in e.g. [Bez+19].

B.4.2.1 Clique with a deleted edge

This calculation is the simplest demonstration of our bounding technique, and all following settings are analysed in a similar way. While it is superseded by the section immediately following it, the bound is thus important for the reasons of comprehension if nothing else.

We consider graphs on $d+1$ nodes, and let P_{Clique} be the Ising model on the complete graph on $d+1$ nodes, with edge $(1,2)$ of weight μ, and every other edge of weight λ. Q_{Clique} is formed by deleting the edge $(1,2)$ in P_{Clique} Note that such underlying constructions feature in nearly every lower bound on structural inference on degree bounded Ising models.

With the exposition out of the way, we state the bound below.

[4] This effect is linked to the concentration of the Ising model on the clique we mentioned before. Notice that the probability of a uniform state is as $1 - \exp\left(-\Omega(\beta d)\right)$. For this to be appreciable, i.e., at least polynomially close to 1, a condition like $\beta d = \Omega(\log d)$ is in fact necessary.

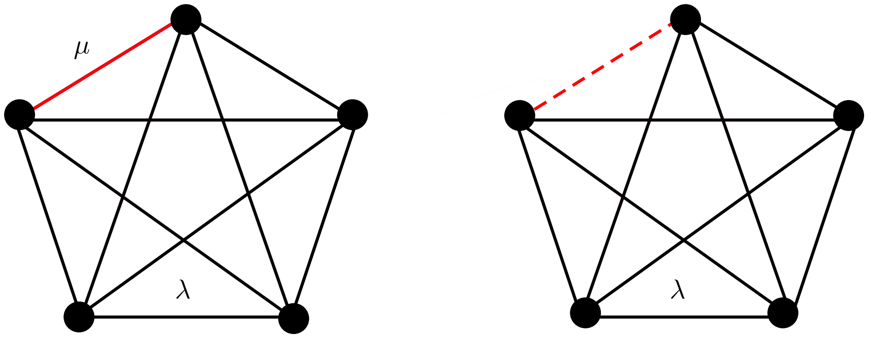


Figure B·4: The clique with uniform weight λ barring one edge, and the same edge deleted. Here $d = 4$.

Proposition B.4.6. *Suppose* $\lambda d > \log d$. *Then*

$$\chi^2(Q_{\text{Clique}} \| P_{\text{Clique}}) \le 16 e^{-2\lambda(d-1)} \sinh^2 \mu.$$

B.4.2.2 The clique with a large hole

To allow for a greater number of changes, we modify the previous construction by removing a large subclique from the K_{d+1} used above, instead of just one edge. More formally, for some $\ell < d/8$, let K_ℓ be the complete graph on nodes $[1 : \ell]$. We set $P_{\ell,\text{Clique}}$ to the the Ising model on K_{d+1} such that the edges in K_ℓ have weight μ, and all other edges have weight λ, while $Q_{\ell,\text{Clique}}$ instead deletes the edges in K_ℓ. Note that as a conseuquence, we have effected a deletion of $\sim \ell^2/2$ edges from the original model.

Proposition B.4.7. *If* $\ell + 1 \le d/8$, $\lambda \ge |\mu|$ *and* $\lambda d > 3 \log d$, *then*

$$\chi^2(Q_{\ell,\text{Clique}} \| P_{\ell,\text{Clique}}) \le 32 \ell e^{-2\beta(d+1-\ell)} \sinh^2(\mu(\ell - 1)).$$

Note that the bound of the previous subsection (up to some factors) can be recovered by setting $\ell = 2$ in the above.

Control on the χ^2-divergence with P and Q exchanged is also useful.

Proposition B.4.8. *If $\ell + 1 \leq d/12$, $\lambda \geq |\mu|$ and $\lambda d > 3 \log d$, then*

$$\chi^2(P_{\ell,\text{Clique}} \| Q_{\ell,\text{Clique}}) \leq 64\ell e^{-2\beta(d+1-\ell)} \sinh^2(2\mu(\ell - 1)).$$

B.4.2.3 Emmentaler Clique

As a development of the Clique with a large hole, we may in fact put in many large holes, leading to a pockmarked clique reminiscent of a Swiss cheese. Concretely, let ℓ be a number such that $B := d/(\ell + 1)$ is an integer. We define a graph on d nodes in the following way: Divide the nodes into B groups of equal size, $V_1, \ldots, V_B$. Form the complete graph on d nodes, and then delete the $\ell + 1$-sublique on V_i for each i. Note that equivalently, the graph above is the complete symmetric B-partite graph on d nodes. The graph effects a deletion of $\sim d\ell/2$ edges from a clique.

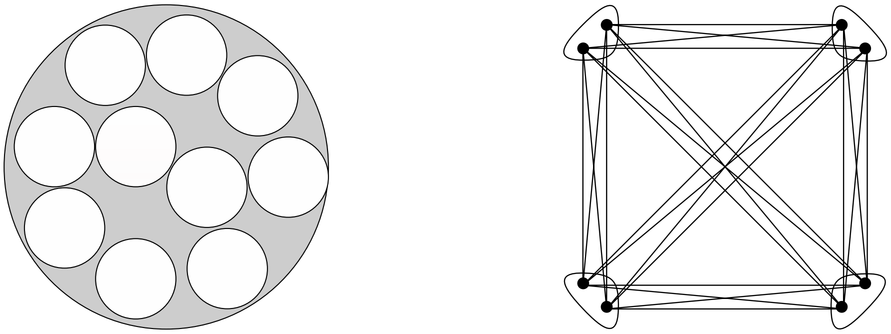


Figure B·5: Two views of the Emmentaler cliques. The left represents the base clique as the large grey circle, while the uncoloured circles within represent the groups V_i with no edges within (this should be viewed as $\ell \gg 1, B = 10$). This view is inspiration for the name. On the right, we represent the Emmentaler as the graph $K_{\ell+1,\ell+1,\ldots,\ell+1}$ - here $d = 8$ and $\ell = 1$ is shown.

The key property of the Emmentaler is that it still freezes at a exponential rate, and it has sufficient 'space' in it to accommodate significantly more edges. In particular, the graph is regular and the degrees of each node are uniformly $d - \ell - 1$. We use this in two ways:

Emmentaler with one extra node We show that determining the degree of a node connected to many of the nodes of an Emmentaler is hard. Concretely, we construct the following two graphs on $d+1$ nodes:

Construct an Emmentaler Clique on the first d nodes. Next, connect the node $d+1$ to each node in $\bigcup_{i=1}^{B-1} V_i$. Notice that node $d+1$ is not connected to one of the parts of the Emmentaler. We choose P_ℓ to be the Ising model with uniform weight λ on the this graph. For Q_ℓ, we additionally add edges between node $d+1$ and each node in V_B with weight μ. The following result holds.

Proposition B.4.9. *If* $2 \le \ell + 1 \le d/4$ *and* $\lambda(d-4) \ge 3 \log d$, *and* $|\mu| \le \lambda$, *then*

$$\chi^2(Q_\ell \| P_\ell) \le 32 d e^{-2\lambda(d-1-\ell)}.$$

Notice that the above proposition does not show a μ dependence. This is due to inefficiencies in our proof technique. We strongly conjecture that a bound of the form $(1 + Cd \tanh^2(\mu(\ell+1)) e^{-2\lambda(d-\ell-1)})^n$ holds.

Emmentaler v/s Full Clique We show that it is difficult to distinguish between an Emmentaler and a full clique. Concretely, we let P_ℓ be an Emmentaler as above, and in Q_ℓ, we add back the deleted subcliques to each V_i, but with weight μ.

Proposition B.4.10. *If* $\ell + 1 \le d/4$ *and* $\lambda(d-4) \ge 3 \log d$, *then*

$$\chi^2(Q_\ell \| P_\ell) \le d^2 \min(1, \mu^2 d^4) e^{-2\lambda(d-1-\ell)}.$$

B.5 Miscellaneous

B.5.1 Using statistical formulations to test structural changes

The main text makes the case that statistical formulations of GoF do not give us the whole story when one is interested in structural changes. Concretely, though, this only directly affects the lower bounds. On the other hand, when we restrict alternate hypotheses in the GoF problem to make a lot of changes, then one may expect that tests under statistical formulations are powerful.

Intuitively, this expectation is rendered plausible by the fact that the notion of being close to a given model is similar under the statistical and the structural formulations - equality under one is also equality under the second, at least in the setting of unique network structures, and mere continuity suggests that, at least locally, setting some value of $s(P,\varepsilon)$ or $\varepsilon(s,P)$ should allow one to translate tests from the statistical to the structural notions of changes and vice versa.[5] However, this strategy does not work too well, at least with our current understanding of Ising models. More concretely - utilising statistical tests for structural testing in a sample efficient way requires a *local* understanding of the distortion of the edge-Hamming distance of the graph under the map $(\theta,\theta') \mapsto \mathrm{SKL}(\theta\|\theta')$, which is not available as of now. Global constraints on the same are available, and are unhappily both rather pessimistic, and essentially tight. This means that using the methods developed for testing for statistical divergences in the setting of structural identity testing is problematic.

Some details - the best available results that translate edge-differences to symmetrised KL divergence is via Lemma 4 of [SW12]. The Bhattacharya coefficient of two distributions is $\mathrm{BC}(P,Q) := \sum_x \sqrt{P(x)Q(x)}$. The cited lemma argues that under s changes,

$$\mathrm{BC} \le \exp\left(-Cs\sinh^2(\alpha)e^{-2\beta d}/d\right).$$

Let $-\varphi$ denote the exponent in the above, for conciseness. Since $-2\log\mathrm{BC} \le \mathrm{KL}$, this induces $D_{\mathrm{SKL}} \gtrsim \varphi$, and similarly, since $1-\mathrm{BC} \le \mathrm{TV}$, this tells us also that $\mathrm{TV} \ge 1-\exp\left(-\varphi\right).$ Since $1-e^{-z} \le z$, this means that the best lower bound we can

[5] It should be noted that this analogy is flawed - while the notions of being close are indeed similar, the notion of being far from a model is significantly different under the two formulations. The main text mentions an example illustrating this - if a small group of disconnected nodes is bunched into a clique, a large statistical change is induced due to the marked difference in the marginal law of this group, but the structural change is tiny. Of course, being close and far are ultimately related concepts, and some shadow of this effect must be cast on the closeness argument we have just presented.

possibly derive this way is $\mathrm{TV} \ge \varphi$.

Now, the best known upper bounds for statistical testing under SKL is $(\beta pd/\varepsilon)^2$ up to log factors [DDK16], and under TV for ferromagnets this may be improved to $(pd/\varepsilon)^2$ [Bez+19]. Plugging in the values of ε implicit in the above, the first of these then requires about

$$\left(\frac{\beta pd}{\varphi}\right)^2 \sim \frac{e^{4\beta d}}{\alpha^4}\left(\frac{\beta pd^2}{s}\right)^2,$$

which is worse than the testing by first recovering the underlying network. Similarly, under TV, a similar number is required, but without an extra β-factor, which has little effect in light of terms like $e^{\beta d}$ showing up. So, naïvely using this structural characterisation does not give promising results.

Further, unfortunately, the characterisation of BC, and indeed of KL and TV divergences offered through this is essentially tight. This essentially follows from our results providing control on the χ^2-divergences in various construction, and the control this imposes on KL, TV via the monotonicity of Rényi divergences and Pinsker's inequality. It may be the case that in some special cases, tight bounds for structural testing may be derived via the statistical testing approach above. We have not explored this possibility in detail.

B.5.2 Lower Bounds on Property Testing

In passing, we mention that our constructions improve upon lower bounds for some of the property tests studied in [NL19]. For instance, the triangle construction provides an obstruction to cycle testing that does not require explicit control on α as in [NL19]. Similarly, the Clique with a hole, and the Emmentaler clique with an extra node constructions may serve as obstructions to testing the size of the largest clique, and to testing the value of the maximum degree of the network structures in low temperatures. In high temperatures, the Fan graph construction shows that testing maximum degree is hard. In each case this either improves upon the lower bounds of [NL19] by either improving the exponent from $\beta d/4$ to $2\beta d(1 - o_d(1))$, or by removing an explicit high-temperature condition that is enforced in the lower bound.

B.6 Proofs of Widget Bounds

An Observation For Ising models P, Q,

$$1 + \chi^2(Q\|P) = \sum_x \frac{Q(x)^2}{P(x)} = \sum_x \frac{Z_P}{Z_Q^2} \exp\left(x^T 2\theta_Q x - x^T \theta_P x\right) = \frac{Z_P Z_{2Q-P}}{Z_Q^2},$$

where $Z_{2Q-P} := \sum_x \exp\left(x^T(2\theta_Q - \theta_P)x\right)$ is yet another partition function. We will repeatedly use this form of the χ^2-divergence, without further comment, in the following.

B.6.1 Star-Based Widgets

B.6.1.1 Triangle

Proof of Proposition B.4.1. Let $P = P_{\text{Triangle}}, Q = Q_{\text{Triangle}}$. Note that

$$\begin{aligned} P(x) &= \frac{1}{Z_P} e^{\lambda x_2(x_1+x_3)} \\ Q(x) &= \frac{1}{Z_Q(\mu)} e^{\lambda x_2(x_1+x_3)} e^{\mu x_1 x_3} \end{aligned}$$

Where the partition functions may simply be computed to obtain the expressions below:

$$\begin{aligned} Z_P &= 2^3 \cosh^2 \lambda = 4(\cosh 2\lambda + 1) \\ Z_Q(\mu) &= 4(e^\mu \cosh 2\lambda + e^{-\mu}). \end{aligned}$$

Further, we have that

$$W := \mathbb{E}_P[(Q/P)^2] = \left(\frac{Z_P}{Z_Q(\mu)}\right)^2 \cdot \frac{1}{Z_P} \cdot \sum e^{\lambda x_2(x_1+x_3)} e^{2\mu x_1 x_3} = \frac{Z_P Z_Q(2\mu)}{Z_Q(\mu)^2}.$$

Inserting the previous computed values of these partition functions, we have

$$\begin{aligned}
W &= \frac{(\cosh 2\lambda + 1)(e^{2\mu}\cosh 2\lambda + e^{-2\mu})}{(e^{\mu}\cosh 2\lambda + e^{-\mu})^2} \\
&= \frac{e^{2\mu}\cosh^2 2\lambda + e^{-2\mu} + \cosh 2\lambda(e^{2\mu} + e^{-2\mu})}{(e^{\mu}\cosh 2\lambda + e^{-\mu})^2} \\
&= 1 + \frac{\cosh 2\lambda(e^{\mu} - e^{-\mu})^2}{(e^{\mu}\cosh 2\lambda + e^{-\mu})^2} \\
&\leq 1 + \frac{(e^{\mu} - e^{-\mu})^2}{e^{2\mu}\cosh 2\lambda} \\
&\leq 1 + \frac{4\sinh^2 \mu}{\cosh^2 \mu \cosh 2\lambda} \\
&\leq 1 + 8e^{-2\lambda}\tanh^2 \mu
\end{aligned}$$

where the second and third inequalities both use that $e^x \geq \cosh x \geq e^x/2$, for $x \geq 0$. □

B.6.1.2 Fan with deletions

In keeping with the rest of the text, these proofs will set $2B = d$. Note that the value of B does not enter the resulting bounds.

Proof of Proposition B.4.4. Let

$$P_{\ell,\eta,\mu,\lambda}(x) := \frac{1}{Z(\ell,\eta,\mu,\lambda)} \exp\left(\lambda x_{d+1}(\sum_{i=1}^{d/2} x_{2i}) + \mu x_{d+1}(\sum_{i=\ell+1}^{d/2} x_{2i-1})\right) \cdot \exp\left(\eta x_{d+1}(\sum_{i=1}^{\ell} x_{2i-1}) + \lambda(\sum_{i=1}^{d/2} x_{2i}x_{2i-1})\right).$$

Then $P_{\ell,\mathrm{Fan}} = P_{\ell,0,\mu,\lambda}$, $Q_{\ell,\mathrm{Fan}} = P_{\ell,\mu,\mu,\lambda}$. Further, $Z_{2Q-P} = Z(\ell, 2\mu, \mu, \lambda)$.

Here again the partition function is simple to compute. In essence, the groups (x_{2i-1}, x_{2i}) across i are independent given x_{d+1}, and the expressions, unsurprisingly, are invariant to value of x_{d+1}.

Unfortunately the calculations get a little messy. If one is not interested in the results on property testing in §B.5.2, then the following may be safely skipped. We do note that the steps below are elementary, it is just the form of the expressions that is long.

$$
\begin{aligned}
&Z(\ell,\eta,\mu,\lambda)\\
&= \sum_{x_{d+1}} \sum_{x_{1:d}} \exp\left(\lambda x_{d+1}(\sum_{i=1}^{d/2} x_{2i}) + \mu x_{d+1}(\sum_{i=\ell+1}^{d/2} x_{2i-1}) + \eta x_{d+1}(\sum_{i=1}^{\ell} x_{2i-1}) + \lambda(\sum_{i=1}^{d/2} x_{2i}x_{2i-1})\right)\\
&= \sum_{x_{d+1}} \prod_{i=1}^{\ell} \sum_{x_{2i-1},x_{2i}} e^{x_{d+1}(\eta x_{2i-1}+\lambda x_{2i})+\lambda x_{2i}x_{2i-1}} \cdot \prod_{i=\ell+1}^{d/2} \sum_{x_{2i-1},x_{2i}} e^{x_{d+1}(\mu x_{2i-1}+\lambda x_{2i})+\lambda x_{2i}x_{2i-1}}\\
&= \sum_{x_{d+1}} \left(2e^{\lambda}\cosh((\lambda+\eta)x_{d+1}) + 2e^{-\lambda}\cosh((\lambda-\eta)x_{d+1})\right)^{\ell}\\
&\qquad\qquad \cdot \left(2e^{\lambda}\cosh((\lambda+\mu)x_{d+1}) + 2e^{-\lambda}\cosh((\lambda-\mu)x_{d+1})\right)^{d/2-\ell}\\
&= 2^{d+1}\left(e^{\lambda}\cosh(\lambda+\eta) + e^{-\lambda}\cosh(\lambda-\eta)\right)^{\ell}\left(e^{\lambda}\cosh(\lambda+\mu) + e^{-\lambda}\cosh(\lambda-\mu)\right)^{d/2-\ell}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
1+\chi^2(Q\|P) &= \frac{Z(\ell,0,\mu,\lambda)Z(\ell,2\mu,\mu,\lambda)}{Z(\ell,\mu,\mu,\lambda)^2}\\
&= \left(\frac{\left(e^{\lambda}\cosh(\lambda)+e^{-\lambda}\cosh(\lambda)\right)\left(e^{\lambda}\cosh(\lambda+2\mu)+e^{-\lambda}\cosh(\lambda-2\mu)\right)}{\left(e^{\lambda}\cosh(\lambda+\mu)+e^{-\lambda}\cosh(\lambda-\mu)\right)^2}\right)^{\ell}\\
&=: U^{\ell}.
\end{aligned}
$$

We proceed to estimate U.

$$
\begin{aligned}
U &= \frac{\left(e^{\lambda}\cosh(\lambda)+e^{-\lambda}\cosh(\lambda)\right)\left(e^{\lambda}\cosh(\lambda+2\mu)+e^{-\lambda}\cosh(\lambda-2\mu)\right)}{\left(e^{\lambda}\cosh(\lambda+\mu)+e^{-\lambda}\cosh(\lambda-\mu)\right)^2}\\
&= \frac{e^{2\lambda}\cosh\lambda\cosh(\lambda+2\mu)+e^{-2\lambda}\cosh\lambda\cosh(\lambda-2\mu)}{e^{2\lambda}\cosh^2(\lambda+\mu)+e^{-2\lambda}\cosh^2(\lambda-\mu)+2\cosh(\lambda+\mu)\cosh(\lambda-\mu)}\\
&\qquad + \frac{\cosh(\lambda)\cosh(\lambda+2\mu)+\cosh(\lambda)\cosh(\lambda-2\mu)}{e^{2\lambda}\cosh^2(\lambda+\mu)+e^{-2\lambda}\cosh^2(\lambda-\mu)+2\cosh(\lambda+\mu)\cosh(\lambda-\mu)}
\end{aligned}
$$

By eliminating one factor of the denominator from the numerator above, we obtain the sequence of relations that follows below.

$$\begin{aligned} U &\overset{(a)}{=} 1 + \frac{(e^{2\lambda} + e^{-2\lambda}) \sinh^2 \mu + \sinh(\mu) \left(\sinh(2\lambda + \mu) - \sinh(2\lambda - \mu)\right)}{e^{2\lambda} \cosh^2(\lambda + \mu) + e^{-2\lambda} \cosh^2(\lambda - \mu) + 2 \cosh(\lambda + \mu) \cosh(\lambda - \mu)} \\ &\overset{(b)}{=} 1 + \frac{2 \cosh(2\lambda) \sinh^2 \mu + 2 \cosh(2\lambda) \sinh^2 \mu}{\left(e^{\lambda} \cosh(\lambda + \mu) + e^{-\lambda} \cosh(\lambda - \mu)\right)^2} \\ &= 1 + \frac{4 \sinh^2(\mu) \cosh(2\lambda)}{e^{2\lambda} \cosh^2(\lambda + \mu) + e^{-2\lambda} \cosh^2(\lambda - \mu) + 2 \cosh(\lambda + \mu) \cosh(\lambda - \mu)} \\ &\overset{(c)}{\leq} 1 + 4 \frac{\sinh^2 \mu}{\cosh^2(\lambda + \mu)} \leq 1 + 4 \frac{\sinh^2 \mu}{\cosh^2 \lambda \cosh^2 \mu} \\ &\leq 1 + 16 e^{-2\lambda} \tanh^2 \mu, \end{aligned}$$

where (a) follows by the identities

$$\begin{aligned} \cosh(u) \cosh(u + 2v) - \cosh^2(u + v) &= \sinh^2 v \\ \cosh(u) \cosh(u + 2v) - \cosh(u + v) \cosh(u - v) &= \sinh(v) \sinh(2u + v), \end{aligned}$$

(b) uses

$$\sinh(2u + v) - \sinh(2u - v) = 2 \cosh(2u) \sinh u,$$

and (c) follows by dropping all terms but the first in the denominator, and observing that $e^{2\lambda} \geq \cosh(2\lambda)$. Finally, the inequality $\cosh(\lambda + \mu) \geq \cosh \lambda \cosh \mu$ holds because $\lambda, \mu \geq 0$. □

B.6.2 Clique-based Widgets

The method for showing the bounds is developed in the case of the Clique with a single edge deleted. While there are variations in the proofs of the following two cases, the basic recipe remains the same.

We begin with a technical lemma that is repeatedly used in the following.

Lemma B.6.1. *Let $\tau : [a, b] \to \mathbb{R}$ be a function differentiable on (a, b) such that τ' is strictly concave. If $\tau(a) < 0$ and $\tau(b) > 0$, then τ has exactly one root in (a, b)*

Proof. Since τ' is concave, it can have at most two roots in (a, b). Indeed, if there were three roots $a < x_1 < x_2 < x_3 < b$, then $\exists t \in (0, 1) : x_2 = t x_1 + (1 - t) x_3$, and

$0 = f(x_2) = tf(x_1) + (1-t)f(x_3)$ violates strict concavity. Further, between its roots, τ' must be positive, again by concavity.

Thus, we can break $[a,b]$ into three intervals (I_1, I_2, I_3), some of them possibly trivial[6], of the from $([a, x_1), [x_1, x_2], (x_2, b])$, such that τ is monotone decreasing on the interiors of I_1, I_3 and monotone increasing on the interior of I_2.

Note that τ has at least one root by the intermediate value theorem. We now argue that it cannot have more than one. Since τ is falling on I_1, it follows that $\sup_{x \in I_1} \tau(x) = \tau(a) < 0$, and there is no root in I_1. Similarly, since τ is falling on I_3, $\tau(b) = \inf_{x \in I_3} \tau(x) > 0$, and there is no root in I_3. This leaves I_2, and since τ is monotone on I_2, it has at most one root on the same. □

B.6.2.1 Clique with a single edge deleted

Proof of Proposition B.4.6. Let $P = P_{\text{Clique}}$ and $Q = Q_{\text{Clique}}$ as defined in the main text. For given λ, η, let

$$P_{\lambda,\eta}(x) := \frac{1}{Z(\lambda,\eta)} e^{\frac{\lambda}{2}\left((\sum x_i)^2 - (d+1)\right)} e^{-\eta x_1 x_2}$$

Note that $P = P_{\lambda,\lambda-\mu}$, and $Q = P_{\lambda,\lambda}$. Further,

$$W := \mathbb{E}_P[(Q/P)^2] = \frac{Z(\lambda,\lambda-\mu)Z(\lambda,\lambda+\mu)}{Z(\lambda,\lambda)^2}.$$

We begin by writing Z in a convenient form, derived by breaking the configurations into bins depending on the number of x_is that take the value -1:

$$Z(\lambda,\eta) = \sum_{j=0}^{d-1} \binom{d-1}{j} \Big\{ e^{-\eta} \left(e^{\frac{\lambda}{2}(d+1-2j)^2 - (d+1)} + e^{\frac{\lambda}{2}(d-3-2j)^2 - (d+1)} \right) \\ + 2e^{\eta} e^{\frac{\lambda}{2}((d-1-2j)^2 - (d+1))} \Big\}.$$

Notice above that since $(d-3-2(d-1-j))^2 = (d+1-2j)^2$, and $\binom{d-1}{j} = \binom{d-1}{d-1-j}$, it follows that the sums over the first two terms above are identical. Thus,

[6] i.e. of cardinality 0 or 1. More precise characterisation can be obtained by casework on the number of roots of τ'.

$$\begin{aligned}
Z(\lambda,\eta) &= 2\sum \binom{d-1}{j} e^{-\eta} e^{\frac{\lambda}{2}(d+1-2j)^2-(d+1)} + 2\sum e^{\eta} e^{\frac{\lambda}{2}((d-1-2j)^2-(d+1))} \\
\iff \underbrace{\frac{Z(\lambda,\eta)}{2e^{\lambda/2(d^2-d)}}}_{=:\widetilde{Z}(\lambda,\eta)} &= e^{\lambda d-\eta} \underbrace{\sum \binom{d-1}{j} e^{-2\lambda j(d+1-j)}}_{=:S_1(\lambda)} + e^{-(\lambda d-\eta)} \underbrace{\sum \binom{d-1}{j} e^{-2\lambda j(d-1-j)}}_{=:S_2(\lambda)} \\
\iff \widetilde{Z}(\lambda,\eta) &= e^{\lambda d-\eta} S_1(\lambda) + e^{-\lambda d+\eta} S_2(\lambda).
\end{aligned}$$

Since the term appears often, we set $d' = d-1$. As a consequence of the above, we have

$$\begin{aligned}
W &= \frac{Z(\lambda,\lambda-\mu)Z(\lambda,\lambda+\mu)}{Z(\lambda,\lambda)^2} = \frac{\widetilde{Z}(\lambda,\lambda-\mu)\widetilde{Z}(\lambda,\lambda+\mu)}{\widetilde{Z}(\lambda,\lambda)^2} \\
&= \frac{(e^{\lambda d'+\mu} S_1(\lambda) + e^{-\lambda d'-\mu} S_2(\lambda))(e^{\lambda d'-\mu} S_1(\lambda) + e^{-\lambda d'+\mu} S_2(\lambda))}{(e^{\lambda d'} S_1(\lambda) + e^{-\lambda d'} S_2(\lambda))^2} \\
&= 1 + 4\sinh^2 \mu \frac{S_1 S_2}{(e^{\lambda d'} S_1 + e^{-\lambda d'} S_2)^2} \\
&\le 1 + 4\sinh^2 \mu \frac{e^{-2\lambda d'} S_2(\lambda)}{S_1(\lambda)}.
\end{aligned}$$

The bounds are now forthcoming by controlling S_1, S_2 as in the following

Lemma B.6.2. *If $d \ge 5$ and $\lambda(d-4) \ge \log(d)$, then*

$$\begin{aligned}
&S_1(\lambda) \ge 1 \\
&S_2(\lambda) \le 2 + 3de^{-2\lambda(d-2)} \le 2 + 3/d.
\end{aligned}$$

The bound follows directly from the control offered above. □

This proof describes closely the structure of the forthcoming proofs

- Begin by introducing one free parameter, η varying which yields Ising models that interpolate between P and Q.

- Express the χ^2 divergence as a ratio of partition functions.

- Exploit the symetries of the mean field Ising model to more conveniently write these partition functions.

- Control the terms arising via a 'ratio trick' as in the proof of Lemma B.6.2. At time this is used more than once, or a more direct form of this trick is used instead.

We conclude by showing Lemma B.6.2.

Proof of Lemma B.6.2. $S_1 \geq 1$ follows trivially, since all terms in the sum are non-negative and the first term is $\binom{d-1}{0}e^0 = 1$.

Concentrating on S_2, let $T_j := \binom{d-1}{j}e^{-2\lambda j(d-1-j)}$. Note that $S_2 = \sum T_j$, and that $T_j = T_{d-1-j}$ for every j. Further, for $j \in [0 : d-2]$,

$$\frac{T_{j+1}}{T_j} = \frac{d-1-j}{j+1}e^{-2\lambda(d-2-2j)}.$$

Treating j as a real number in $[0, d-2]$, define

$$\tau(j) = \log(d-1-j) - \log(j+1) - 2\lambda(d-2-2j).$$

We have

$$\begin{aligned}
\tau'(j) &= -\frac{1}{d-1-j} - \frac{1}{j+1} + 4\lambda \\
\tau''(j) &= -\frac{1}{(d-1-j)^2} + \frac{1}{(j+1)^2} \\
\tau'''(j) &= -\frac{2}{(d-1-j)^3} - \frac{2}{(j+1)^3} < 0.
\end{aligned}$$

We may now note that τ' is a strictly concave function on the relevant domain. Further, note that since $\log(d-1) \leq 2\lambda(d-2)$ follows from our conditions, $\tau(0) < 0$, and similarly, $\tau(d-2) > 0$. By Lemma B.6.2, τ has exactly one root in $[0, d-2]$ - in particular, this lies at $j = d/2 - 1$. But since $T_{j+1}/T_j = e^{\tau(j)}$, we obtain that for $j \leq d/2 - 1, T_{j+1} \leq T_j$, and for $j \geq d/2 - 1, T_{j+1} \geq T_j$.

Since Ts are decreasing until $d/2 - 1$ and increasing after $d/2$, it follows that for all $j \in [2 : d-3]$, $T_j \leq \max(T_2, T_{d-3}) = T_2$. Now, under the conditions of the theorem,

$$\begin{aligned}\frac{T_2}{T_1} &= \exp\left(\tau(1)\right) = \exp\left(\log(d-2) - \log 2 - 2\lambda(d-4)\right)\\ &\leq \exp\left(\log(d-2) - \log 2 - 2\log(d)\right) \leq 1/d,\end{aligned}$$

where we have used the assumption $\lambda(d-4) \geq \log d$. Thus,

$$\begin{aligned}S_2 &= T_0 + T_1 + \sum_{j=2}^{d-3} T_j + T_{d-2} + T_{d-1}\\ &\leq 1 + T_1 + \frac{d-4}{d} T_1 + T_1 + 1\\ &\leq 2 + 3d\exp\left(-2\lambda(d-2)\right) \leq 2 + 3/d. \qquad \square\end{aligned}$$

We call this method of estimating sums such as S_2 the *ratio trick*, since they control the values of the sums by controlling the ratios of subsequent terms.

B.6.2.2 Clique with Large Hole

The computations of this section are in essence a deepening of the previous section, and we will frequently make references to the same.

Proof of Proposition B.4.7. Once again condensing notation, let $P := P_{\ell,\mathrm{Clique}}, Q := Q_{\ell,\mathrm{Clique}}$.

Further, let

$$P_{\ell,\lambda,\eta}(x) := \frac{1}{Z_\ell(\lambda,\eta)} e^{\frac{\lambda}{2}\left(\sum_{1\leq i\leq d+1} x_i\right)^2 - (d+1)} e^{-\frac{\eta}{2}\left(\sum_{1\leq i\leq \ell} x_i\right)^2 - \ell}$$

Again, $P = P_{\ell,\lambda,\lambda-\mu}, Q = P_{\ell,\lambda,\lambda}$ holds. Z_ℓ is the central object for this section, and has the following expression. This is derived by tracking the number of negative x_is in both the bulk of the clique and the single 'hole' separately.

$$\begin{aligned}Z_\ell(\lambda,\eta) &:= \sum_{\{\pm 1\}^{d+1}} e^{\frac{\lambda}{2}\left(\sum_{1\leq i\leq d+1} x_i\right)^2 - (d+1)} e^{-\frac{\eta}{2}\left(\sum_{1\leq i\leq \ell} x_i\right)^2 - \ell}\\ &= \sum_{i,j} \binom{\ell}{i}\binom{d+1-\ell}{j} e^{\frac{\lambda}{2}(d+1-2i-2j)^2 - (d+1)} e^{\frac{-\eta}{2}(\ell-2i)^2 - \ell}\end{aligned}$$

We normalise Z_ℓ by $e^{\lambda/2((d+1)^2-(d+1))}e^{-\eta/2(\ell^2-\ell)}$, and put a $\sim$ over the normalised version[7] to get

$$\begin{aligned}\widetilde{Z}_\ell(\lambda,\eta) &:= \sum_{i,j}\binom{\ell}{i}\binom{d+1-\ell}{j}e^{-2\lambda j(d+1-2i-j)}e^{2\eta i(\ell-i)}e^{-2\lambda i(d+1-i)}\\ &=: \sum_{i=0}^{\ell}\binom{\ell}{i}e^{2\eta i(\ell-i)}e^{-2\lambda i(d+1-i)}S_i(\lambda)\end{aligned}$$

where

$$S_i(\lambda) := \sum_j \binom{d+1-\ell}{j}e^{-2\lambda j(d+1-2i-j)}.$$

Notice that $S_i \geq 0$ for every i.

As before, we are interested in controlling

$$W := \frac{Z_\ell(\lambda,\lambda-\mu)Z_\ell(\lambda,\lambda+\mu)}{Z_\ell(\lambda,\lambda)^2} = \frac{\widetilde{Z}_\ell(\lambda,\lambda-\mu)\widetilde{Z}_\ell(\lambda,\lambda+\mu)}{\widetilde{Z}_\ell(\lambda,\lambda)^2}.$$

To this end, note first that $2\lambda i(\ell-i) - 2\lambda i(d+1-i) = -2\lambda(d+1-\ell)$, and so, for instance,

$$\widetilde{Z}_\ell(\lambda,\lambda+\mu) = \sum_i \binom{\ell}{i}e^{2\mu i(\ell-i)}e^{-2\lambda i(d+1-\ell)}S_i(\lambda).$$

Collecting like terms in expressions of the above form, we obtain that

$$\frac{\widetilde{Z}_\ell(\lambda,\lambda-\mu)}{\widetilde{Z}_\ell(\lambda,\lambda)} = 1 + \frac{\sum_{i=1}^{\ell-1}\binom{\ell}{i}\left(e^{-2\mu i(\ell-i)}-1\right)e^{-2\lambda i(d+1-\ell)}S_i(\lambda)}{\widetilde{Z}_\ell(\lambda,\lambda)}$$

and

$$\frac{\widetilde{Z}_\ell(\lambda,\lambda+\mu)}{\widetilde{Z}_\ell(\lambda,\lambda)} = 1 + \frac{\sum_{i=1}^{\ell-1}\binom{\ell}{i}\left(e^{2\mu i(\ell-i)}-1\right)e^{-2\lambda i(d+1-\ell)}S_i(\lambda)}{\widetilde{Z}_\ell(\lambda,\lambda)},$$

where the terms involving $i = 0$ and $i = \ell$ in the numerator drop out because

[7] Unlike in §B.6.2.1, we include the factor due to η in the normalisation. This does not affect the further calculations since these factors cancel in the expression for W below. More importantly, the normalisation includes a factor of $e^{\lambda/2((d+1)^2-(d+1))}$ instead of $e^{\lambda/2(d^2-d)}$. While the latter lent the formulae in the $\ell = 2$ case of the previous section a pleasant symmetry, the former yields more convenient expressions when dealing with ℓ abstractly. Due to this, the terms are further reduced by a common factor of $e^{\lambda d}$. We highlight this here because of the cosmetic differences arising from these changes—for instance, the leading term in $\widetilde{Z}_\ell$ is just S_1 instead of $e^{\lambda d-\eta}S_1$ as in the §B.6.2.1—which may irk the careful reader at first glance.

$e^{2\mu i(\ell-i)} = 1$ in these cases.

Now, if $\mu \geq 0$ the second terms in the above two expressions are respectively negative and positive, while if $\mu < 0$, they are respectively positive and negative. It is a triviality that for $A < 0 < B, (1+A)(1+B) \leq 1 + A + B$. We thus have the upper bound

$$\begin{aligned} W &\leq 1 + \frac{\sum_{i=1}^{\ell-1} \binom{\ell}{i} 2\left(\cosh 2\mu i(\ell-i) - 1\right) e^{-2\lambda i(d+1-\ell)} S_i(\lambda)}{\widetilde{Z}_\ell(\lambda,\lambda)} \\ &= 1 + 4\frac{\sum_{i=1}^{\ell-1} \binom{\ell}{i} \sinh^2(\mu i(\ell-i)) e^{-2\lambda i(d+1-\ell)} S_i(\lambda)}{\widetilde{Z}_\ell(\lambda,\lambda)} \end{aligned} \tag{B.1}$$

While we will provide full proofs in the sequel, it may help to see where we are going first. Roughly, we argue via the ratio trick in the proof of Lemma B.6.2 in the previous section, that S_i is bounded by $2(1 + e^{-2\lambda(\ell-2i)(d+1-\ell)})$, under conditions such as $\lambda(d+1-2\ell) \geq \log d + 1 - 2\ell$. Plugging in this upper bound, and noting that after multiplication with $e^{-2\lambda i(d+1-\ell)}$ we have a sum that is completely symmetric under $i \mapsto \ell - i$, we can bound W as

$$W \leq 1 + 16\frac{\sum_{i=1}^{\ell-1} \binom{\ell}{i} \sinh^2(\mu i(\ell-i)) e^{-2\lambda i(d+1-\ell)}}{\widetilde{Z}_\ell(\lambda,\lambda)}.$$

We then show that under the conditions of the proposition, the first term in the above sum dominates all the remaining terms, in the process utilising the condition $|\mu| \leq \lambda$. Finally, using the trivial bound $\widetilde{Z}_\ell(\lambda,\lambda) \geq 1$, we get the claied upper bound.

Let us then proceed. The control on the S_is is offered below.

Lemma B.6.3. *If $\lambda(d+1-2\ell) \geq \log(d+1-2\ell)$ and $d \geq 4\ell$, then for every $i \in [1:\ell-1]$,*

$$S_i(\lambda) \leq 2 + 2e^{-2\lambda(\ell-2i)(d+1-\ell)}.$$

Incorporating the above lemma into (B.1), we have

$$\begin{aligned}
W &\le 1 + 8\frac{\sum_{i=1}^{\ell-1}\binom{\ell}{i}\sinh^2(\mu i(\ell-i))e^{-2\lambda i(d+1-\ell)}\left(1+e^{-2\lambda(\ell-2i)(d+1-\ell)}\right)}{\widetilde{Z}_\ell(\lambda,\lambda)} \\
&\le 1 + 8\frac{\sum_{i=1}^{\ell-1}\binom{\ell}{i}\sinh^2(\mu i(\ell-i))\left(e^{-2\lambda i(d+1-\ell)}+e^{-2\lambda(\ell-i)(d+1-\ell)}\right)}{\widetilde{Z}_\ell(\lambda,\lambda)} \\
&\overset{(a)}{=} 1 + 16\frac{\sum_{i=1}^{\ell-1}\binom{\ell}{i}\sinh^2(\mu i(\ell-i))e^{-2\lambda i(d+1-\ell)}}{\widetilde{Z}_\ell(\lambda,\lambda)} \\
&= 1 + \frac{16}{\widetilde{Z}_\ell(\lambda,\lambda)}\left(\sinh^2(\mu(\ell-1))e^{-2\lambda(d+1-\ell)} + \sum_{i=2}^{\ell-1}\binom{\ell}{i}\sinh^2(\mu i(\ell-i))e^{-2\lambda i(d+1-\ell)}\right) \\
&\overset{(b)}{\le} 1 + \frac{16}{\widetilde{Z}_\ell(\lambda,\lambda)}\left(\sinh^2(\mu(\ell-1))e^{-2\lambda(d+1-\ell)} + \sum_{i=2}^{\ell-1}\binom{\ell}{i}e^{2|\mu|i\ell-2\lambda i(d+1-\ell)}\right) \\
&\overset{(c)}{\le} 1 + \frac{16}{\widetilde{Z}_\ell(\lambda,\lambda)}\left(\sinh^2(\mu(\ell-1))e^{-2\lambda(d+1-\ell)} + \sum_{i=2}^{\ell-1}\binom{\ell}{i}e^{-2\lambda i(d+1-2\ell)}\right) \qquad \text{(B.2)}
\end{aligned}$$

where the equality (a) follows since each term in the sum is invariant under the map $i \mapsto \ell - i$, (b) follows since $\sinh x \le e^x$, and (c) used $\lambda \ge |\mu|$. .

For $i \in [2:\ell]$, let V_i denote the term corresponding to i in the summation above, and let $V_1 = \sinh^2(\mu(\ell-1)e^{-2\lambda(d+1-\ell)}$. We will argue that V_1 dominates V_i for every i by using a weakened ratio trick.

Note that

$$V_1 \ge e^{-2\lambda(d+1-\ell)-2|\mu|(\ell-1)} \ge e^{-2\lambda d}.$$

Further,

$$\frac{V_i}{V_1} \le \exp\left(i\log\ell + 2\lambda d - 2\lambda i(d+1-2\ell)\right).$$

This is smaller than $1/\ell$ so long as for every i,

$$i(2\lambda(d+1-2\ell) - \log\ell) > 2\lambda d + \log(\ell),$$

which hold if the following conditions are true:

$$\begin{aligned}
2\lambda(d+1-2\ell) &> \log\ell \\
4\lambda(d+1-2\ell) &> 3\log\ell + 2\lambda d.
\end{aligned}$$

The above hold if $\lambda(d+2-4\ell) \ge 3/2\log\ell$, which is true under the conditions of the proposition since $\ell < d/8$, and since $\lambda(d+2-4\ell) \geq \lambda d/2 \geq 3/2 \log d$.

Finally, it remains to show that $\widetilde{Z}_\ell(\lambda, \lambda)$ is non-trivially large. But note that $\widetilde{Z}_\ell(\lambda, \lambda) \geq S_0(\lambda) \geq 1$.

Thus, we have shown that

$$W \leq 1 + 32\ell \sinh^2(\mu(\ell-1))e^{-2\lambda(d+1-\ell)}.$$

□

Proof of Lemma B.6.3. For $j \in [0 : d+1-\ell]$, let

$$T_j := \binom{d+1-\ell}{j} e^{-2\lambda j(d+1-2i-j)}.$$

Recall that $S_i = \sum T_j$. We will use the ratio trick again. To this end, observe that

$$\frac{T_{j+1}}{T_j} = \frac{d+1-\ell-j}{j+1} \exp\left(-2\lambda(d-2i-2j)\right. .$$

Again treating j as a real number in $[0 : d-\ell]$, let

$$\tau(j) := \log(d+1-\ell-j) - \log(1+j) - 2\lambda(d-2i-2j).$$

By considerations similar to the previous section, τ is strictly concave, and by Lemma B.6.2, τ has exactly one root so long as $\tau(0) < 0$ and $\tau(d-\ell) > 0$. In this setting these conditions translate to

$$\begin{aligned}
&\log(d+1-\ell) < 2\lambda(d-2i) \\
&\log(d+1-\ell) < -2\lambda(d-2i-2(d-\ell)) = 2\lambda(d-2(\ell-i)).
\end{aligned}$$

The above hold for every i so long as $\log(d+1-\ell) < 2\lambda(d+2-2\ell)$.

Since τ has a single root and is initially negative, we again find that for all $j \in [2 : d-1-\ell]$, $T_j \leq \max(T_2, T_{d-1-\ell})$. Further,

$$\begin{aligned}
\frac{T_2}{T_1} &= \frac{d-\ell}{2} \exp\left(-2\lambda(d-2-2i)\right) \leq \frac{d-\ell}{2} \exp\left(-2\lambda(d-2\ell)\right) \leq \frac{1}{d-\ell} \\
\frac{T_{d-\ell-1}}{T_{d-\ell}} &= \frac{d-\ell}{2} \exp\left(-2\lambda(d-2(\ell-i)\right) \leq \frac{1}{d-\ell}.
\end{aligned}$$

Further,

$$\max\left(\frac{T_1}{T_0}, \frac{T_{d-\ell}}{T_{d+1-\ell}}\right) \le (d+1-\ell)e^{-2\lambda(d-2\ell)} \le 1/2.$$

Thus,

$$\begin{aligned} S_1 &\le T_0 + T_{d+1-\ell} + (1 + (d-\ell-2)/(d-\ell)) \max(T_1, T_{d-\ell}) \\ &\le T_0 + T_{d+1-\ell} + 2\max(T_1, T_{d-\ell}) \\ &\le 2(T_0 + T_{d+1-\ell}). \end{aligned}$$

Now notice that

$$\begin{aligned} T_0 &= 1 \\ T_{d-\ell+1} &= \exp\left(-2\lambda(d+1-\ell)(d+1-2i-d-1+\ell)\right) \\ &= \exp\left(-2\lambda(\ell-2i)(d+1-\ell)\right), \end{aligned}$$

and thus the claim follows. □

We now prove the reverse direction, i.e. control on $\chi^2(P\|Q)$. This is essentially a small variation on the previous setting.

Proof of Proposition B.4.8. Referring to the previous proof, we instead need to control

$$W' = \frac{\widetilde{Z}_\ell(\lambda,\lambda)\widetilde{Z}_\ell(\lambda,\lambda+2\mu)}{\widetilde{Z}_\ell(\lambda,\lambda+\mu)^2}.$$

Proceeding in the same way, we may conntrol

$$W' \le 1 + \frac{\sum_{i=1}^{\ell-1}\binom{\ell}{i}\left(\cosh(4\mu i(\ell-i)) - 2\cosh(2\mu(i(\ell-i)) + 1\right)e^{-2\lambda i(d+1-\ell)}S_i(\lambda)}{\widetilde{Z}_\ell(\lambda,\lambda+\mu)}$$

For succinctness, let $f(x) := \cosh(4\mu x) - 2\cosh(2\mu x) + 1$. Note that $1 \le f(x) \le$

$e^{4|\mu|x}$. Since the S_i are identical to the previous case, Lemma B.6.3 applies, and

$$\begin{aligned} W' &\leq 1 + 8\frac{\sum_{i=1}^{\ell-1}\binom{\ell}{i} f(i(\ell-i))e^{-2\lambda i(d+1-\ell)}\left(1+e^{-2\lambda(\ell-2i)(d+1-\ell)}\right)}{\widetilde{Z}_\ell(\lambda,\lambda+\mu)} \\ &\leq 1 + 16\frac{\sum_{i=1}^{\ell-1}\binom{\ell}{i} f(i(\ell-i))e^{-2\lambda i(d+1-\ell)}}{\widetilde{Z}_\ell(\lambda,\lambda+\mu)} \\ &\leq 1 + \frac{16}{\widetilde{Z}_\ell(\lambda,\lambda+\mu)}\left(f(\ell-1)e^{-2\lambda(d+1-\ell)} + \sum_{i=2}^{\ell-1}\binom{\ell}{i} e^{4|\mu|i\ell-2\lambda i(d+1-\ell)}\right) \\ &\leq 1 + \frac{16}{\widetilde{Z}_\ell(\lambda,\lambda+\mu)}\left(f(\ell-1)e^{-2\lambda(d+1-\ell)} + \sum_{i=2}^{\ell-1}\binom{\ell}{i} e^{-2\lambda i(d+1-3\ell)}\right) \end{aligned}$$

Notice the distinction that the exponent in the second sum contains a -3ℓ instead of a -2ℓ. Using $f(x) \geq 1$, the same control on the relative values of S_i and the summation holds as long as

$$4\lambda(d+1-3\ell) > 3\log\ell + 2\lambda d.$$

This translates to demanding that $2\lambda(d-6\ell) > 3/2\lambda d$, which holds for $\ell \leq d/12$. Finally, $\widetilde{Z}_\ell(\lambda,\lambda+\mu) \geq 1$ as well, and thus,

$$W' \leq 1 + 32\ell e^{-2\lambda(d+1-\ell)}\left(\cosh(4\mu(\ell-1)) - 2\cosh(2\mu(\ell-1)) + 1\right).$$

Finally, we note that for any x,

$$\begin{aligned} \cosh(4x) - 2\cosh(2x) + 1 &= \sinh^2(2x) + (\cosh(2x)-1)^2 \\ &= 4\sinh^2 x\cosh^2 x + 4\sinh^4 x \\ &= 4\sinh^2 x\cosh^2 x(1+\tanh^2 x) \\ &\leq 2\sinh^2(2x). \end{aligned}$$

□

B.6.2.3 Emmentaler Cliques

Proof of Proposition B.4.9. Recall the setup - $d+1$ nodes are divided into $B = d/(\ell+1)$ groups of $\ell+1$ nodes each, denoted $V_1, \ldots, V_B$, and the final node $d+1$ is kept separate. Recall that for a set S, $x_S := \sum_{u\in S} x_u$. Define

$$P_{\ell,\lambda,\eta} = \frac{1}{Z_\ell(\lambda,\eta)} \exp\left(\lambda/2\left(\sum_{i=1}^{B} x_{V_i}\right)^2 - \lambda/2\sum_{i=1}^{B}(x_{V_i}^2) + \lambda x_v \sum_{i=2}^{B} x_{V_i} + \eta x_v x_{V_1}\right).$$

Then $P = P_{\ell,\text{Emmentaler}} = P_{\ell,\lambda,0}, Q = Q_{\ell,\text{Emmentaler}} = P_{\ell,\lambda,\mu}$ and $Z_{2Q-P} = Z_\ell(\lambda, 2\mu)$ Marginalising over x_v, we get

$$\begin{aligned} &Z_\ell(\lambda,\eta) \\ &\quad = 2\sum_x \exp\left(\lambda/2\left(\sum_{i=1}^{B} x_{V_i}\right)^2 - \lambda/2\sum_{i=1}^{B}(x_{V_i}^2)\right)\cosh\left(\lambda\sum_{i=2}^{B} x_{V_i} + \eta x_{V_1}\right) \\ &\quad \le 2\cosh(\lambda(d-\ell-1)+\eta(\ell+1))\sum_x \exp\left(\lambda/2\left(\sum_{i=1}^{B} x_{V_i}\right)^2 - \lambda/2\sum_{i=1}^{B}(x_{V_i}^2)\right), \end{aligned}$$

while dropping all terms for which $|\sum_i x_{V_i}| < d$, we get

$$\begin{aligned} Z_\ell(\lambda,\eta) &\ge 4\cosh(\lambda(d-\ell-1)+\eta(\ell+1))e^{\lambda/2(B^2-B)(\ell+1)^2} \\ &= 4\cosh(\lambda(d'-\ell-1)+\mu(\ell+1))e^{\lambda/2(d^2-d(\ell+1))}. \end{aligned}$$

To control Z_ℓ from above, it is necessary to control the partition function of the Emmentaler graph on d nodes (i.e., with only the groups $V_1, \dots V_B$, and without the extra node from above. We set this equal to $Y_\ell(\lambda)$. Then, similarly tracking configurations by the number of negative x_is in each part,

$$\begin{aligned} Y_\ell :=& \sum_x \exp\left(\lambda/2\left(\sum_{i=1}^{B} x_{V_i}\right)^2 - \lambda/2\sum_{i=1}^{B}(x_{V_i}^2)\right). \\ =& \sum_{j_1,\dots,j_B}\prod\binom{\ell+1}{j_i}\cdot\exp\left(\lambda/2\left((d-2\sum j_i)^2 - \sum(\ell+1-2j_i)^2\right)\right) \\ =& e^{\lambda/2(d^2-d(\ell+1))}\sum_{j_1,\dots,j_B}\prod\binom{\ell+1}{j_i}\cdot\exp\left(-2\lambda\left((d-\ell-1)(\sum j_i) + \sum j_i^2 - (\sum j_i)^2\right)\right). \end{aligned}$$

For succinctness, let $d' := d-\ell-1$. We establish the following lemma after concluding this argument

Lemma B.6.4. *If $\ell \le d/4$ and $\lambda(d-4) \ge 3\log(d)$, then*

$$Y_\ell \le 2e^{\lambda/2(d^2 - d(\ell+1))}\left(1 + 2de^{-2\lambda d'}\right)$$

Invoking the above lemma and the previously argued control on Z_ℓ, we get that

$$\begin{aligned} W := \mathbb{E}_P[(Q/P)^2] &= \frac{Z_\ell(\lambda,0)Z_\ell(\lambda,2\mu)}{Z_\ell(\lambda,\mu)^2} \\ &\le \frac{\cosh(\lambda d')\cosh(\lambda d' + 2\mu(\ell+1))}{\cosh^2(\lambda d' + \mu(\ell+1))}\left(\frac{2Y_\ell}{4e^{\lambda/2(d^2-d(\ell+1))}}\right)^2 \\ &\le \left(1 + \frac{\sinh^2(\mu(\ell+1))}{\cosh^2(\lambda d' + \mu(\ell+1))}\right)\left(1 + 2de^{-2\lambda d'}\right)^2 \\ &\le \left(1 + 4\tanh^2(\mu(\ell+1))e^{-2\lambda d'}\right)\left(1 + 2de^{-2\lambda d'}\right)^2 \end{aligned}$$

Under the conditions of the theorem, both $4\tanh^2(\mu(\ell+1))e^{-2\lambda d'}$ and $2de^{-2\lambda d'}$ are smaller than $1/4$. But for x, y, it holds that $(1+x)^2 < 1 + 3x$ and $(1+3x)(1+y) < 1 + 4(x+y) \le 1 + 8\max(x,y)$. Lastly, $4\tanh^2 x \le 4 \le d$, and thus, we have shown the bound

$$W \le 1 + 32de^{-2\lambda(d-\ell-1)}. \qquad \square$$

Proof of Lemma B.6.4. Fix a vector $(j_1, \dots, j_B)$ and let $k := \sum j_i$. We will argue the claim by controlling the terms in Y_ℓ with a given value of k.

Lemma B.6.5. *If $\sum j_i = k \in [2 : d-2]$, $\ell + 1 \le d/4$ and $\lambda(d-4) \ge 3\log(d)$, then*

$$\prod \binom{\ell+1}{j_i} \cdot \exp\left(-2\lambda\left(d'(\sum j_i) + \sum j_i^2 - (\sum j_i)^2\right)\right) \le \frac{1}{d^{\min(k,d-k)}}e^{-2\lambda d'}.$$

Thus, we have the bound

$$\frac{Y_\ell}{e^{\lambda/2(d^2-d(\ell+1))}} \le 2\left(1 + B(\ell+1)e^{-2\lambda d'}\right) + \sum_{k=2}^{d-2} \frac{N_k}{d^{\min(k,d-k)}}e^{-2\lambda d'},$$

where

$$N_k = \left|\left\{j \in [0 : \ell+1]^B : \sum j_i = k\right\}.\right|$$

Notice that $N_k = N_{d-k}$. Further, for $k \le d/2$, by stars and bars,

$$N_k \le \binom{k+B-1}{k} \le (1 + (B-1)/k)^{k-1} \le B^k \le d^k$$

Consequently, $N_k \le d^{\min(k,d-k)}$, and we have established the upper bound

$$\frac{Y_\ell}{2e^{\lambda/2(d^2-d(\ell+1))}} \le 1 + 2de^{-2\lambda d'}. \qquad \square$$

Proof of Lemma B.6.5. Note that $\binom{n}{m} \le n^{\min(m,n-m)}$. Therefore,

$$\prod \binom{\ell+1}{j_i} \le \exp\left(\min(k,d-k)\log(\ell+1)\right).$$

Next, by Cauchy-Schwarz,

$$\sum j_i^2 \ge \frac{(\sum j_i)^2}{B} = k^2\left(1 - \frac{d'}{d}\right).$$

Let LHS, RHS be the left and right hand sides of the inequality claimed in the Lemma. Using the above,

$$\begin{aligned}\log\frac{\text{LHS}}{\text{RHS}} &\le \min(k,d-k)\log(d(\ell+1)) - 2\lambda\left(d'k + k^2 d'/d - d'\right)\\ &= \min(k,d-k)\log(d(\ell+1)) - 2\lambda\frac{d'}{d}\left(k(d-k) - d\right).\end{aligned}$$

Let $f(k)$ be the upper bound above. Notice that $f(k) = f(d-k)$. Thus, it suffices to show that $f(u) \le 0$ for every real number $u \in [2, d/2]$.

For a real number $u \in [2, d/2)$, it holds that $f''(u) = 4\lambda > 0$. It follows that f attains its maxima on $\{2, d/2\}$. Since $\ell + 1 < d/4$, we have $d'/d \ge 3/4$, and thus

$$\begin{aligned}f(2) &= 2\log(d(\ell+1)) - 2\lambda\frac{d'}{d}(d-4) \le 4\log(d) - \frac{3}{2}\lambda(d-4) < 0\\ f(d/2) &= \frac{d}{2}\left(\log(d(\ell+1) - 2\lambda\frac{d'}{d}\cdot\frac{(d-4)}{2}\right) = \frac{d}{4}f(2) < 0. \qquad \square\end{aligned}$$

B.6.2.4 Emmentaler v/s Full Clique

Proof of Proposition B.4.10. Let

$$P_{\ell,\lambda,\eta}(x) := \frac{1}{Z_\ell(\lambda,\eta)}\exp\left(\lambda/2\left(\left(\sum_{i=1}^{B} x_{V_i}\right)^2 - d\right) - (\lambda-\eta)/2\sum_{i=1}^{B}(x_{V_i}^2 - (\ell+1))\right).$$

Then $P_\ell = P_{\ell,\lambda,0}, Q_\ell = P_{\ell,\lambda,\mu}$. Let $d' = d - 1 - \ell$. Developing this a little, one can write

$$Z_\ell(\lambda,\eta) = C_{\ell,\lambda,\eta} \sum_{j_1,\dots,j_B} \prod \binom{\ell+1}{j_i} \cdot e^{-2\lambda\left(d'\sum j_i + \sum j_i^2 - (\sum j_i)^2\right) - 2\eta\left((\ell+1)\sum j_i - \sum j_i^2\right)},$$

where

$$C_{\ell,\lambda,\eta} = \exp\left(\lambda/2(d^2 - d(\ell+1)) + \eta d(\ell+1)/2\right).$$

Notice that

$$\frac{C_{\ell,\lambda,0} C_{\ell,\lambda,2\mu}}{C^2_{\ell,\lambda,\mu}} = 1,$$

and thus

$$W := \mathbb{E}_P[(Q/P)^2] = \frac{Z_\ell(\lambda,0) Z_\ell(\lambda,2\mu)}{Z_\ell(\lambda,\mu)^2} = \frac{\widetilde{Z}_\ell(\lambda,0)\widetilde{Z}_\ell(\lambda,2\mu)}{\widetilde{Z}_\ell(\lambda,\mu)^2},$$

where

$$\widetilde{Z}_\ell(\lambda,\eta) := \frac{Z_\ell(\lambda,\eta)}{C_{\ell\lambda,\eta}} = \sum_{k=0}^{d} e^{-2\lambda\left(d'k - k^2\right) - 2\eta(\ell+1)k} \sum_{\substack{j_1,\dots,j_B \\ \sum j_i = k}} \prod \binom{\ell+1}{j_i} \cdot e^{-2(\lambda-\eta)\sum j_i^2}.$$

Let T_k be the k^{th} term in the above. It holds that $T_k = T_{d-k}$. Indeed, the original terms are invariant under the map $x \mapsto -x$, and for $j = (j_1, \dots, j_B)$, this maps to $(\ell+1)\mathbf{1} - j$ which has the sum $d - k$.

Further, since

$$\sum j_i^2 \le \max_i(j_i) \sum j_i \le (\ell+1)\sum j_i,$$

it holds that each term, which depends on η as $e^{-2\eta((\ell+1)\sum j_i - \sum j_i^2}$ decreases as η increases (or equivalently, $\frac{\partial}{\partial\eta}\widetilde{Z}_\ell(\lambda,\eta) \le 0$)

Due to the above, for $\mu > 0$,

$$\rho_1 := \frac{\widetilde{Z}_\ell(\lambda,0) - \widetilde{Z}_\ell(\lambda,\mu)}{\widetilde{Z}_\ell(\lambda,\mu)} \ge 0$$

$$\rho_2 := \frac{\widetilde{Z}_\ell(\lambda,2\mu) - \widetilde{Z}_\ell(\lambda,\mu)}{\widetilde{Z}_\ell(\lambda,\mu)} \le 0,$$

yielding,

$$W = \frac{\widetilde{Z}_\ell(\lambda, 0)\widetilde{Z}_\ell(\lambda, 2\mu)}{\widetilde{Z}_\ell(\lambda, \mu)^2} \leq 1 + \rho_1 + \rho_2.$$

(For $\mu < 0$, the signs of both ρ_1 and ρ_2 are flipped, giving the same bound.)

We now offer control on $\rho_1 + \rho_2$, to complete the argument. To this end, note that

$$1 - 2e^{-2\mu\left((\ell+1)k - \sum j_i^2\right)} + e^{-4\mu\left((\ell+1)k - \sum j_i^2\right)} = \left(1 - e^{-2\mu\left((\ell+1)k - \sum j_i^2\right)}\right)^2,$$

and thus

$$\begin{aligned}
&\widetilde{Z}_\ell(\lambda, \mu)(\rho_1 + \rho_2) \\
&= \sum_{k=1}^{d-1} \sum_{j : \sum j_i = k} \prod \binom{\ell+1}{j_i} e^{-2\lambda(d'k - k^2 + \sum j_i^2)} \left(1 - e^{-2\mu\left((\ell+1)k - \sum j_i^2\right)}\right)^2 \\
&\leq 2 \sum_{k=1}^{\lfloor d/2 \rfloor} \sum_{j : \sum j_i = k} \prod \binom{\ell+1}{j_i} e^{-2\lambda(d'k - k^2 + \sum j_i^2)} \left(1 - e^{-2\mu\left((\ell+1)k - \sum j_i^2\right)}\right)^2,
\end{aligned}$$

where we have used the symmetry of the $T_k s$ above.

We argue below that the first term in the above strongly dominates all subsequent terms.

Lemma B.6.6. *If $\sum j_i = k \in [2 : \lfloor d/2 \rfloor]$, $\ell + 1 \leq d/4$ and $\lambda(d-4) \geq 3\log(d)$, then*

$$\prod \binom{\ell+1}{j_i} e^{-2\lambda(d'k - k^2 + \sum j_i^2)} \leq \frac{1}{d^k} e^{-2\lambda d'}.$$

Using the above, along with $\sum j_i^2 \geq \sum j_i$ and the fact that the number of B-tuples of whole numbers that sum up to k is at most $\binom{k+B-1}{k} \leq (eB)^k \leq d^k$, we immediately have

$$\widetilde{Z}_\ell(\lambda, \mu)(\rho_1 + \rho_2) \leq 2de^{-2\lambda d'} \sum_{k=1}^{d/2} \left(1 - e^{-2\mu\ell k}\right)^2.$$

We bound the sum above in two ways - firstly, each term is ≤ 1, and so the sum is at most $d/2$. Further, using $1 - e^{-x} \leq x$, the sum is at most $4 \sum \mu^2 \ell^2 k^2 \leq \mu^2 d^5$. This gives ,

$$\widetilde{Z}_\ell(\lambda, \mu)(\rho_1 + \rho_2) \leq 2d^2 \min(1, \mu^2 d^4) e^{-2\lambda(d-1-\ell)}$$

The bound on W now follows since $\widetilde{Z}_\ell(\lambda,\mu) \geq 2$ trivially. □

Proof of Lemma B.6.6. This is essentially the same as Lemma B.6.4, and may be proved similarly. □

B.6.2.5 The Clique versus the Empty Graph in High Temperatures

Proof of Proposition B.4.3. This proof heavily relies on techniques that we encountered in [CNL18]. The principal idea is via the following representation of the law of an Ising model with uniform edge weights, and the subsequent expression (and upper bound) for its partition function, both of which we encountered in the cited paper.

Let $\tau = \tanh(\mu)$. Then the law of the Ising model on a m-vertex graph G with uniform weights α is

$$P(X = x) = \frac{\prod_{(i,j)\in G}(1 + \tau X_i X_j)}{2^m \mathbb{E}_0[\prod_{(i,j)\in G}(1 + \tau X_i X_j)},$$

where $\mathbb{E}_0$ denotes expectation with respect to the uniform law on $\{-1,1\}^m$. This is shown by noticing that $\exp(x) = \cosh(x)(1 + \tanh(x))$, and then observing that for $x = \mu X_i X_j$, since $X_i X_j = \pm 1$, the same is equal to $\cosh(\mu)(1 + \tanh(\mu) X_i X_j)$. The $\cosh(\mu)$ term is fixed for all entries, and thus vanishes under the normalisation. The denominator is simply a restatement of $\sum_{\{-1,1\}^m} \prod_{(i,j)\in G}(1 + \tau X_i X_j)$.

Let the denominator of the above be denoted $2^m \Phi(\tau; G)$. We further have the expansion

$$\Phi(\tau; G) = \sum_{u \geq 0} \mathscr{E}(u, G)\tau^u,$$

where $\mathscr{E}(j, G)$ denotes the number of 'Eulerian subgraphs of G', where we call a graph Eulerian if each of its connected components is Eulerian (and recall that a connected graph is Eulerian if and only if each of its nodes has even degree). This arises by expanding the above product out to get

$$\Phi(\tau; G) = \sum_{u \geq 0} \tau^u \cdot \sum_{\text{choices of } u \text{ edges } (i_1,j_1),(i_2,j_2),\dots(i_u,j_u)} \mathbb{E}_0[X_{i_1} X_{j_1} \dots X_{i_u} X_{j_u}].$$

Now, due to the independence, if any node of the X_is or the X_js appears an odd number of times in the product, the expectation of that term under $\mathbb{E}_0$ is zero. If they all appear an even number of times, the value is of course 1. Thus the inner sum, after expectation, amounts to the number of groups of u edges such that each node occurs an even number of times in this set of edges, which corresponds to the number of Eulerian subgraphs of G, defined in the above way.

A further subsidiary lemma controls the size of $\mathscr{E}(u, G)$ as follows, where we abuse notation and use G to denote the adjacency matrix of the graph G.

$$\mathscr{E}(u, G) \leq (2\|G\|_F)^u.$$

The idea behind this is to first control the number of length-v closed walks in a graph, by noticing that the total number of length v walks from i to i is $(G^v)_{i,i}$, summing which up gives an upper bound on the number of closed length v walks of $\mathrm{Tr}(G^v) \leq \|G\|_F^v$. Next, we note that to get an Eulerian subgraph of G with u edges, we can either take a closed walk of length u in G, or we can add a closed walk of length $v \leq u - 2$ to an Eulerian subgraph with $u - v$ edges. This yields a Grönwall-style inequality that the authors solve inductively. Please see [CNL18, Lemma A.1].

Now, let P be the Ising model K_m with uniform weight α, and let Q be the Ising model on the empty graph on m nodes. Using the above expression for the law of an Ising model, we have

$$1 + \chi^2(Q\|P) = \mathbb{E}_Q[Q/P] = \mathbb{E}_0[\prod_{i<j}(1 + \tau X_i X_j)]\mathbb{E}_0[\prod_{i<j}(1 + \tau X_i X_j)^{-1}],$$

which, by multiplying and dividing each term in the second expression by $1 - \tau X_i X_j$, and noting that $X_i^2 X_j^2 = 1$, may further be written as

$$\begin{aligned} 1 + \chi^2(Q\|P) &= \mathbb{E}[\prod_{i<j}(1 + \tau X_i X_j)]\mathbb{E}\left[\frac{\prod_{i<j}(1 - \tau X_i X_j)}{(1 - \tau^2)^{-\binom{m}{2}}}\right] \\ &= \Phi(\tau; K_m)\Phi(-\tau; K_m)(1 - \tau^2)^{-\binom{m}{2}}. \end{aligned}$$

Since the above expression is invariant under a sign flip of τ, we may assume, without loss of generality, that $\tau \geq 0$. Next, notice, due to the expansion in terms of $\mathscr{E}$ of Φ, that $\Phi(-\tau; K_m) \leq \Phi(\tau; K_m)$ for $\tau \geq 0$. Further, for $\tau \geq 0$, using the bound on $\mathscr{E}(u, G)$,

$$\Phi(\tau; K_m) \leq \mathscr{E}(0; K_m) + t\mathscr{E}(1; K_m) + t^2\mathscr{E}(2; K_m) + \sum_{u \geq 3}(2t\|K_m\|_F)^u.$$

Now notice that $\mathscr{E}(0; K_m) = 1$, and $\mathscr{E}(1; K_m) = \mathscr{E}(2; K_m) = 0$. The first of these is because there is only a single empty graph, while the other two follow since K_m is a simple graph. Further, $\|K_m\|_F = \sqrt{m(m-1)} \le m$. Thus, we have

$$\Phi(\tau; K_m) \le 1 + \sum_{u \ge 3} (2tm)^u.$$

Now, since $2\tanh(\alpha)m \le 2\alpha m \le 1/16 < 1/2$, we sum up and bound the geometric series to conclude that $\Phi(\tau; K_m) \le 1 + 16(tm)^3 \le 1 + (tm)^2$, and as a consequence,

$$\Phi(\tau; K_m)^2 \le (1 + (tm)^2)^2 \le 1 + 3(tm)^2 \le \exp\left(3(tm)^2\right).$$

Further, since $\tau m < 1/32$, and $m \ge 1$, we have $\tau < 1/32$, which in turn implies that $(1-\tau^2)^{-1} \le \exp(2\tau^2)$. Thus, we find that

$$1 + \chi^2(P\|Q) \le \exp\left(3(\tau m)^2\right) \cdot \left(\exp\left(2\tau^2\right)\right)^{m^2/2} \le \exp\left(4(\tau m)^2\right) \le 1 + 8(\tau m)^2,$$

where the final inequality uses the fact that for $x < \ln(2)$, $e^x \le 1 + 2x$, which applies since $4(\tau m)^2 \le 4/(32)^2 < \ln(2)$. □

It is worth noting that Proposition B.4.2 is also shown in the above framework by [CNL18]. The main difference, however, is that in the χ^2 computations, the square of $\prod(1 + \tau X_i X_j)$ appears. The technique the authors use is to extend the notion of $\mathscr{E}$ to multigraphs, and show the same expansion for these, along with the same upper bound for $\mathscr{E}(u, G)$, this time with the entries of G denoting the number of edges between the corresponding nodes.

Appendix C

Appendix to Chapter 4

C.1 Appendix to §4.2

C.1.1 Proof of Proposition 4.2.1

Proof. We recall the notation. $\alpha \in [0,1]^K$ is such that $\sum \alpha_k \leq 1$. The $\mathcal{T}_k^\alpha$ are the optimising solutions to the OSP problems at error $\alpha_k \varepsilon$, i.e.

$$\mathcal{T}_k^\alpha \in \arg\max \mathbb{P}(\mathcal{T}) \text{ s.t. } \mathbb{P}(X \in \mathcal{T}, Y \neq k) \leq \alpha_k \varepsilon,$$

while the $\mathcal{S}_k^\alpha$ are produced by removing the smaller overlap with smaller labels in $\mathcal{T}_k^\alpha$, i.e.

$$\mathcal{S}_k^\alpha = \mathcal{T}_k^\alpha \setminus \bigcup_{k'<k} \mathcal{T}_{k'}^\alpha.$$

We first argue that the total overlap of the $\mathcal{T}$s is small.

Lemma C.1.1. *Let $\mathcal{T}_k^\alpha$ be generated as above. Then*

$$\sum_k \mathbb{P}(\bigcup_{k' \neq k} \mathcal{T}_k^\alpha \cap \mathcal{T}_{k'}^\alpha) \leq 2\varepsilon.$$

Since the total overlap is $\bigcup_k \left(\mathcal{T}_k^\alpha \cap \bigcup_{k' \neq k} \mathcal{T}_{k'}^\alpha\right)$, this also controls the probability of the total overlap, that is,

$$\mathbb{P}(\bigcup_{k,k' \neq k} \mathcal{T}_k^\alpha \cap \mathcal{T}_{k'}^\alpha) \leq 2\varepsilon.$$

This lemma is sufficient to show the claim, since

$$\begin{aligned}\sum_k \mathbb{P}(\mathcal{S}_k) &= \sum_k \mathbb{P}(\mathcal{T}_k^\alpha \setminus \bigcup_{k'<k} \mathcal{T}_{k'}^\alpha) \\ &\geq \sum_k \mathbb{P}(\mathcal{T}_k^\alpha) - \sum_k \mathbb{P}(\mathcal{T}_k^\alpha \cap \bigcup_{k'<k} \mathcal{T}_{k'}^\alpha) \\ &\geq \sum_k \mathbb{P}(\mathcal{T}_k^\alpha) - \sum_k \mathbb{P}(\mathcal{T}_k^\alpha \cap \bigcup_{k'\neq k} \mathcal{T}_{k'}^\alpha) \\ &\geq C(\varepsilon; \mathscr{S}) - \sum_k \mathbb{P}(\mathcal{T}_k^\alpha \cap \bigcup_{k'\neq k} \mathcal{T}_{k'}^\alpha) \\ &\geq C(\varepsilon; \mathscr{S}) - 2\varepsilon,\end{aligned}$$

where we have used that $\sum_k \mathbb{P}(\mathcal{T}_k^\alpha) \geq C(\varepsilon; \mathscr{S})$, which holds because the $\mathcal{T}_k^\alpha$ optimise a relaxation of (SC), and the final inequality is due to the above lemma. □

We conclude by proving the above lemma

Proof of Lemma C.1.1. Since the labels of Y are mutually, exclusive,

$$\sum_k \mathbb{P}(\bigcup_{k'\neq k} \mathcal{T}_k^\alpha \cap \mathcal{T}_{k'}^\alpha) = \sum_k \sum_j \mathbb{P}(\bigcup_{k'\neq k} \mathcal{T}_k^\alpha \cap \mathcal{T}_{k'}^\alpha, Y = j).$$

Applying Fubini's theorem, and recalling that the probability of an intersection of events is smaller than the probability of either of the events, we see that

$$\begin{aligned}\sum_k \mathbb{P}(\bigcup_{k'\neq k} \mathcal{T}_k^\alpha \cap \mathcal{T}_{k'}^\alpha) &= \sum_k \sum_j \mathbb{P}(\mathcal{T}_k^\alpha \cap (\bigcup_{k'\neq k} \mathcal{T}_{k'}^\alpha), Y = j) \\ &\leq \sum_k \left(\sum_{j\neq k} \mathbb{P}(\mathcal{T}_k^\alpha, Y = j) \right) + \mathbb{P}(\bigcup_{k'\neq k} \mathcal{T}_{k'}^\alpha, Y = k),\end{aligned}$$

Now, notice that the sum in the brackets is simply $\mathbb{P}(\mathcal{T}_k^\alpha, Y \neq k)$. Taking the union bound over the second probability, we find the upper bound

$$\begin{aligned}\sum_k \mathbb{P}(\bigcup_{k'\neq k} \mathcal{T}_k^\alpha \cap \mathcal{T}_{k'}^\alpha) &\le \sum_k \mathbb{P}(\mathcal{T}_k^\alpha, Y \neq k) + \sum_k \sum_{k' \neq k} \mathbb{P}(\mathcal{T}_{k'}^\alpha, Y = k)\\ &= \sum_k \mathbb{P}(\mathcal{T}_k^\alpha, Y \neq k) + \sum_{k'} \sum_{k \neq k'} \mathbb{P}(\mathcal{T}_{k'}^\alpha, Y = k)\\ &= \sum_k \mathbb{P}(\mathcal{T}_k^\alpha, Y \neq k) + \sum_{k'} \mathbb{P}(\mathcal{T}_{k'}^\alpha, Y \neq k')\\ &= 2\sum_k \mathbb{P}(\mathcal{T}_k^\alpha, Y \neq k)\\ &\le 2\sum \alpha_k \varepsilon = 2\varepsilon,\end{aligned}$$

where the first equality is by Fubini's theorem again, the second equality is by the disjointness of the values of Y, and the final inequality is due to the constraints of the OSP problems.

□

C.1.2 Proofs of Propositions 4.2.2 and 4.2.3

C.1.2.1 Proofs of Necessity of Finite VC dimension

In both cases, we reduce the problems to realisable PAC learning, and invoke standard bounds for the same, for instance the one of Chapter 3 in the book by Mohri et al. [MRT18, Ch.3]. To this end, suppose $\delta \le 1/100$, and consider the restricted class of joint laws $\mathbb{P}$ such that $\mathbb{P}(Y = k|X = x) = \mathbb{1}\{X \in \mathcal{S}_{k,*}\}$ for some disjoint $\{\mathcal{S}_{k,*}\} \in \mathscr{S}$ that together cover $\mathcal{X}$.[1]

Proof for One-Sided Prediction. Notice that $\mathcal{S}_1^*$ is feasible for OSP-1 for any value of ε. If we can solve OSP-1, then we would have found a set $\mathcal{S}$ such that

$$\begin{aligned}\mathbb{P}(\mathcal{S}) &\ge \mathbb{P}(\mathcal{S}_{1,*}) - \sigma\\ \mathbb{P}(X \in \mathcal{S} \cap \mathcal{S}_{1,*}^c) &= \mathbb{P}(X \in \mathcal{S}, Y = 2) \le \varepsilon + \nu.\end{aligned}$$

[1] Strictly speaking, this requires that $\mathscr{S}$ is rich enough to express such a class. This is a very mild assumption. For the purposes of the lower bound, in fact, this can be weakened still - all we really need is a binary law, and that if $\mathcal{S} \in \mathscr{S}$, then $\mathcal{S}^c \in \mathscr{S}$. Then we can take $\mathbb{P}(Y = 1|X = x) = \mathbb{1}\{X \in \mathcal{S}\}, \mathbb{P}(Y = 2|X = x) = \mathbb{1}\{X \in \mathcal{S}^c\}$, and the entirety of the following argument goes through without change.

Further,

$$\begin{aligned}\mathbb{P}(\mathcal{S}^c) &= \mathbb{P}(\mathcal{S}^c \cap \mathcal{S}_{1,*}) + \mathbb{P}(\mathcal{S}^c \cap \mathcal{S}_{1,*}^c)\\ &= \mathbb{P}(\mathcal{S}^c \cap \mathcal{S}_{1,*}) + \mathbb{P}(\mathcal{S}_{1,*}^c) - \mathbb{P}(\mathcal{S} \cap \mathcal{S}_{1,*}^c).\end{aligned}$$

But $\mathbb{P}(\mathcal{S}^c) = 1 - \mathbb{P}(\mathcal{S}) \le 1 - \mathbb{P}(\mathcal{S}_{1,*}) + \sigma \le \mathbb{P}(\mathcal{S}_{1,*}^c) + \sigma$.
Thus, we have

$$\begin{aligned}&\mathbb{P}(\mathcal{S}^c \cap \mathcal{S}_{1,*}) + \mathbb{P}(\mathcal{S}_{1,*}^c) - \mathbb{P}(\mathcal{S} \cap \mathcal{S}_{1,*}^c) \le \mathbb{P}(\mathcal{S}_{1,*}^c) + \sigma\\ &\qquad\qquad \implies \mathbb{P}(\mathcal{S}^c \cap \mathcal{S}_{1,*}) \le \sigma + \mathbb{P}(\mathcal{S} \cap \mathcal{S}_{1,*}^c) \le \varepsilon + \sigma + \nu.\end{aligned}$$

But then, viewed as a standard classifier for the problem separating the class $\{1\}$ from $[2:K]$, $\mathcal{S}$ has risk at most $2\varepsilon + \sigma + \nu$. Consequently, an algorithm for solving OSP yields an algorithm for realisable PAC learning for this problem. Thus, invoking the appropriate standard lower bound, we conclude that

$$m_{\mathrm{OSP}} \ge \frac{\mathrm{VC}(\mathscr{S}) - 1}{32(2\varepsilon + \sigma + \nu)}. \qquad \square$$

Proof for Learning With Abstention. Notice that $\{\mathcal{S}_{k,*}\}$ serve as a feasible solution for any ε, and have total coverage 1. Thus, if SC is possible, we may recover sets $\{\mathcal{S}_k\}$ such that

$$\begin{aligned}\sum \mathbb{P}(\mathcal{S}_k) &\ge 1 - \sigma\\ \mathbb{P}(\mathcal{E}_{\{\mathcal{S}_k\}}) &\le \varepsilon + \nu\\ \mathbb{P}\left(\bigcup_k (\mathcal{S}_k \cap \bigcup_{k' \neq k} \mathcal{S}_{k'})\right) &\le \nu.\end{aligned}$$

Now notice that $\mathcal{S}_{1,*}$ and $\mathcal{S}_{1,*}^c$ correspond to the realisable classifiers for the binary classification problem separating $\{1\}$ from $[2:K]$.[2] But, in the same way, we may view $\mathcal{S}_1$ and $\mathcal{S}_1^c$ as binary classifiers for this problem. Now notice that for this binary classification problem, $\mathcal{S}_1$ incurs small error. Indeed, denoting $\mathcal{S}_{\neq 1} = \bigcup_{k' \neq 1} \mathcal{S}_{k'}$, we find that

[2] Again, this needs that $\mathscr{S}$ is rich enough to include $\mathcal{S}_{1,*}^c$.

$$\begin{aligned}\mathbb{P}(X \in \mathcal{S}_1, Y \neq 1) + \mathbb{P}(X \in \mathcal{S}_1^c, Y = 1) &= \mathbb{P}(X \in \mathcal{S}_1, Y \neq 1) \\ &\quad + \mathbb{P}(X \in \mathcal{S}_{\neq 1} \cap \mathcal{S}_1^c, Y = 1) \\ &\qquad + \mathbb{P}(X \in \mathcal{S}_{\neq 1}^c \cap \mathcal{S}_1^c, Y = 1) \\ &\leq \mathbb{P}(X \in \mathcal{S}_1, Y \neq 1) \\ &\quad + \mathbb{P}(X \in \mathcal{S}_{\neq 1}, Y = 1) \\ &\qquad + \mathbb{P}(X \in \mathcal{S}_1^c \cap \mathcal{S}_{\neq 1}^c) \\ &\leq \mathbb{P}(\mathcal{E}_{\{\mathcal{S}_k\}}) + (1 - \mathbb{P}(\mathcal{S}_1 \cup \mathcal{S}_{\neq 1}^c))) \\ &\leq \varepsilon + K\nu + \sigma + \zeta, .\end{aligned}$$

where the second inequality is due to non-negativity of probabilities, and the third inequality is due the fact that $\mathbb{P}(\mathcal{E})$ is controlled, and the following inclusion-exclusion argument:

First note that

$$\mathbb{P}(\mathcal{S}_1 \cup \mathcal{S}_{\neq 1}) = \mathbb{P}(\bigcup \mathcal{S}_k) = \sum_k \mathbb{P}(\mathcal{S}_k) - \sum_k \mathbb{P}(\mathcal{S}_k \cap \bigcup_{k'>k} \mathcal{S}_{k'}).$$

Next, observe that if $j > k$, $\mathcal{S}_j \subset \bigcup_{k'>k} \mathcal{S}_{k'}$, and similarly $\bigcup_{k'>j} \mathcal{S}_{k'} \subset \bigcup_{k'>k} \mathcal{S}_k$. Thus,

$$\mathbb{P}(\mathcal{S}_1 \cup \mathcal{S}_{\neq 1}) = \mathbb{P}(\bigcup \mathcal{S}_k) \geq \sum_k \mathbb{P}(\mathcal{S}_k) - K\mathbb{P}(\mathcal{S}_1 \cap \bigcup_{k'>1} \mathcal{S}_{k'}).$$

Now invoking the SC solution conditions, the first sum is at least $1 - \sigma$, while the second probability is bounded by the probability of overlap, giving

$$\mathbb{P}(\mathcal{S}_1 \cup \mathcal{S}_{\neq 1}) \geq 1 - \sigma - K\nu.$$

Thus, a SC yields a realisable PAC learner for the binary classifier problem separating $\{1\}$ from $[2 : K]$, giving the bound

$$m_{\mathrm{SC}} \geq \frac{\mathrm{VC}(\mathscr{S}) - 1}{32(\varepsilon + \sigma + K\nu + \zeta)}. \qquad \square$$

Note that these bounds are likely loose. The problems have plenty of structure that is not exploited in either of the above statements, and tighter inequalities would be of interest. However the point we intend to pursue - that assuming finiteness of VC dimensions in the upper bound analyses is not lossy, is sufficiently made above.

C.1.2.2 Proofs of the Upper Bounds

We mainly make use of the following uniform generalisation bound on the suprema of empirical processes due to the finiteness of VC dimension. This is, again, standard [MRT18].

Lemma C.1.2. *Let $\mathscr{S}$ have finite VC dimension. Then for any distribution $\mathbb{P}$, if $\widehat{P}_m$ denotes the empirical law induced by m i.i.d. samples from $\mathbb{P}$, then with probability at least $1-\delta$ over these samples,*

$$\sup_{\mathcal{S}\in\mathscr{S},k\in[1:K]} |\widehat{\mathbb{P}}_m(X\in\mathcal{S},Y=k)-\mathbb{P}(X\in\mathcal{S},Y=k)| \le C_K\sqrt{\frac{\mathrm{VC}(\mathscr{S})\log m+\log(C/\delta)}{m}},$$

where C_K is a constant independent of $\mathscr{S},\delta,\mathbb{P},m$.

Notice that by summing over the values of Y, this also controls the error in the objects $\mathbb{P}(X\in\mathcal{S})$ and $\mathbb{P}(X\in\mathcal{S},Y\neq k)$, possibly with an error blowup of K, which can be absorbed into C_K.

For the purposes of the following, let $\Delta_{m,\mathscr{S}}(\delta)$ be the value of the upper bound above.

Proof of Upper Bound for OSP. For $\alpha\in[0,1]$, define $\mathscr{S}_\alpha\subset\mathscr{S}$ to be the subset of $\mathcal{S}$s that have $\mathbb{P}(\mathcal{E}^1_{\mathcal{S}})\le\alpha$, and let σ,ν be quantities that we will choose.

We give a two phase scheme - first we collect all sets $\mathcal{S}$ such that $\widehat{P}_m(\mathcal{E}^1_{\mathcal{S}})\le\varepsilon+\nu/2$ into the set $\widehat{\mathscr{S}_{\varepsilon+\nu/2}}$. Notice that as long as $\nu/2>\Delta_{m,\mathscr{S}}(\delta/2)$, we have w.p. at least $1-\delta/2$ that

$$\mathscr{S}_\varepsilon\subset\widehat{\mathscr{S}_{\varepsilon+\nu/2}}\subset\mathscr{S}_{\varepsilon+\nu}.$$

Due to the upper inclusion, with probability at least $1-\delta/2$, every set in $\widehat{\mathscr{S}_{\varepsilon+\nu/2}}$ has error level at most $\varepsilon+\nu$.

Next, we choose the $\mathcal{S}\in\widehat{\mathscr{S}_{\varepsilon+\nu/2}}$ that has the biggest coverage. If $\mathscr{S}_\varepsilon\subset\widehat{\mathscr{S}_{\varepsilon+\nu/2}}$, and $\sigma>\Delta_{m,\mathscr{S}}(\delta/2)$, we are again assured that the selected answer will be at least $\sup_{\mathcal{S}\in\mathscr{S}_\varepsilon}\mathbb{P}(\mathcal{S})-\Delta_{m,\mathscr{S}}(\delta/2)>L_k-\sigma$ with probability at least $1-\delta/2$. By the union bound, these will hold simultaneously with probability at least $1-\delta$. Since we want the smallest σ, ν, but for the arguments to follow we need that these are bigger than $2\Delta_{m,\mathscr{S}}(\delta/2)$, we can set

$$\nu = \sigma = 4\Delta_{m,\mathscr{S}}(\delta/2) = 4C\sqrt{\frac{\mathrm{VC}(\mathscr{S})\log m + \log(2C/\delta)}{m}},$$

concluding the proof. □

Proof of Upper Bound for SC. This proceeds similarly to the above. For the sake of convenience, we let $\mathscr{R} = \{\mathcal{S} : \mathcal{S} = \bigcup_{k,k' \neq k} \mathcal{S}_k \cap \mathcal{S}_{k'}, \{\mathcal{S}_k\} \in \mathscr{S}\}$ be the class of sets obtained by taking pairwise intersection of k-tuples in $\mathscr{S}$. Note that VC dimesnsion of the sets obtained by taking pairwise intersection of sets in $\mathscr{S}$ at most doubles the VC dimesnsion, while taking the $\binom{K}{2}$ unions in turn blows it up by a factor of $O(K^2 \log K$ by Lemma 3.2.3 of Blumer et al. [BEHW89]. Thus $\mathrm{VC}(\mathscr{R}) = O(K^2\mathrm{VC}(\mathscr{S})\log K)$. Now we may proceed as above, first by filtering the pairs of sets that satisfy the intersection constraint with value $\zeta/2$ on the empirical distribution, and then similarly checking the sum-error constraints and finally optimising the sum of their masses. The bounds are the same as the above, except with $\mathrm{VC}(\mathscr{S})$ replaced by $O(K^2\mathrm{VC}(\mathscr{S})\log K)$. □

C.1.2.3 Analyses not pursued here

We first point out that there is nothing special about the VC theoretic analysis here - alternate methods like Rademacher complexity or a covering number analysis may replace Lemma C.1.2. Similarly, the same analysis could be extended, via Rademacher complexities, to the setting of indicators relaxed to Lipschitz surrogates by exploiting Talagrand's lemma.

We note a few further analyses that we do not pursue here - firstly, using the technique of Rigollet & Tong [RT11], it should be possible to give analyses for SC under convex surrogates of the indicator losses and a slight extension of the class $\mathscr{S}$ while directly attaining the constraints (instead of asymptotically) with high prob. Additionally, a number of papers concentrate on deriving fast rates for the excess risks under the assumption of realizability (i.e., under the assumption that level sets of

η can be expressed via $\mathscr{S}$), and that Tsybakov's noise condition holds at the level relevant to the optimal solution.

C.2 Algorithmic rewriting of Section 4.3

We specify the conclusions of §4.3 without any of the justifying development.

Model class and Architecture We use a DNN with the following structure:

- A 'backbone', parametrised by θ, which may have any convenient architecture.
- A 'last layer' with K outputs, denoted f_k, and associated weights w_k for each. We denote $\mathbf{w} = (w_1, \ldots, w_K)$.
- Let $\xi_\theta(x)$ denote the backbone's output on a point x. The DNN's outputs are

$$f(x; \theta, \mathbf{w}) = (f_1, \ldots, f_k)(x; \theta, \mathbf{w}) = \mathrm{softmax}(\langle w_1, \xi_\theta(x)\rangle, \ldots, \langle w_K, \xi_\theta(x)\rangle).$$

Objective function and Training We use the following objective function, where the $\{(x_i, y_i)\}_{i=1}^n$ comprise the training dataset, $\theta, \mathbf{w}$ are model parameters, $\{\varphi_k\}$ are autotuned hyperparameters, $\{\lambda_k\}$ are autotuned multipliers, and μ is the single externally tuned parameter. Similarly to $\mathbf{w}$, we define $\boldsymbol{\varphi} := (\varphi_1, \ldots, \varphi_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$.

$$\widetilde{M}^{\text{res.}}(\theta, \mathbf{w}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \mu) = \sum_{k=1}^{K} \Bigg\{ \frac{\sum_{i:y_i=k} - \log f_k(x_i; \theta, \mathbf{w})}{n_k} \\ + \lambda_k \left(\frac{\sum_{i:y_i \neq k} - \log(1 - f_k(x_i; \theta, \mathbf{w})}{n_{\neq k}} - \varphi_k \right) + \mu \varphi_k \Bigg\},$$

where $n_k := |\{i : y_i = k\}|, n_{\neq k} := |\{i : y_i \neq k\}|$.

The minimax problem we propose is

$$\min_{\theta, \mathbf{w}, \boldsymbol{\varphi}} \max_{\boldsymbol{\lambda}: \forall k, \lambda_k \geq 0} \widetilde{M}^{\text{res.}}(\theta, \mathbf{w}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \mu), \tag{C.1}$$

which is optimised via SGDA in §4.4.

Overall Scheme and Model Selection is presented in Algorithm 5. The subroutine involving the minimax solution requires training data, but this is not mentioned in the same since the focus is on model selection. The training procedure is described in §4.4. $\widehat{\mathbb{P}}_V$ refers to the empirical law on the validation dataset.

Algorithm 5 OSP-Based Selective Classifier: Model Selection

```
**Inputs**: Validation data $\{V\}$, List of $\mu$ values $\mathbf{M}$, List of $t$ values $\mathbf{T}$, Target Error $\varepsilon$.
**for** each $\mu \in \mathbf{M}$, **do**
    $(\theta(\mu), \mathbf{w}(\mu)) \leftarrow$ minimax solution of the program (C.1) with this value of $\mu$.
**for** each $(\mu, t) \in \mathbf{M} \times \mathbf{T}$, **do**
    $\mathcal{S}_k(\mu, t) \leftarrow \{x : k = \arg\max_j f_j(x; \theta(\mu), \mathbf{w}(\mu))\} \cap \{x : f_k(x; \theta(\mu), \mathbf{w}(\mu)) > t\}$.
    $\widehat{E}_V(\mu, t) \leftarrow \widehat{\mathbb{P}}_V(\mathcal{E}_{\{\mathcal{S}_k(\mu,t)\}})$.
    $\widehat{C}_V(\mu, t) \leftarrow \sum_k \widehat{\mathbb{P}}_V(X \in \mathcal{S}_k(\mu, t))$.
$(\mu_*, t_*) = \arg\max_{\mathbf{M}\times\mathbf{T}} \widehat{C}_V(\mu, t)$ s.t. $\widehat{E}_V(\mu, t) \leq \varepsilon$.
**return** $\{\mathcal{S}_k(\mu_*, t_*)\}$.
```

C.3 Experimental Details

The table below presents the values of the various hyperparameters used for the entries of Table 4.2.

Dataset	Algorithm	Hyper-parameters
CIFAR-10	Softmax Response	$t = 0.0445$
	Selective Net	$\lambda = 32, c = 0.51, t = 0.24$
	Deep Gamblers	$o = 1.179, t = 0.03$
	OSP-Based	$\mu = 0.49, t = 0.8884$
SVHN-10	Softmax Response	$t = 0.0224$
	Selective Net	$\lambda = 32, c = 0.79, t = 0.86$
	Deep Gamblers	$o = 1.13, t = 0.23$
	OSP-Based	$\mu = 1.67, t = 0.9762$
Cats v/s Dogs	Softmax Response	$t = 0.029$
	Selective Net	$\lambda = 32, c = 0.7, t = 0.73$
	Deep Gamblers	$o = 1.34, t = 0.06$
	OSP-Based	$\mu = 1.67, t = 0.9532$

Table C.1: Final hyper-parameters used for all the algorithms (at the desired 0.5% error level) in Table 4.2.

The following two tables update the numbers for Deep Gamblers to the case where we scan for 40 values of o in the set $[1, 10)$ (as intended in the specifications) instead of $[1, 2)$.

Dataset	Target	OSP-based		SR		SN		DG	
	Error	Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error
CIFAR-10	2%	**80.6**	1.91	75.1	2.09	73.0	2.31	72.9	1.99
	1%	**74.0**	1.02	67.2	1.09	64.5	1.02	63.5	1.01
	0.5%	**64.1**	0.51	59.3	0.53	57.6	0.48	56.1	0.51
SVHN-10	2%	**95.8**	1.99	95.7	2.06	93.5	2.03	94.7	2.01
	1%	**90.1**	1.03	88.4	0.99	86.5	1.04	89.7	0.99
	0.5%	**82.4**	0.51	77.3	0.51	79.2	0.51	81.4	0.51
Cats & Dogs	2%	**90.5**	1.98	88.2	2.03	84.3	1.94	87.4	1.94
	1%	**85.4**	0.98	80.2	0.97	78.0	0.98	81.7	0.98
	0.5%	**78.7**	0.49	73.2	0.49	70.5	0.46	74.5	0.48

Table C.2: Performance at Low Target Error. This repeats Table 4.2, except that the hyperparameter scan for the DG method is corrected, and the entries in the DG columns are updated to show the resulting values. Notice that the performance in the last column is worse than in Table 4.2.

Dataset	Target	OSP-based		SR		SN		DG	
	Cov.	Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error
CIFAR-10	100%	100	9.74	100	**9.58**	100	11.07	100	10.95
	95%	95.1	**6.98**	95.2	8.74	94.7	8.34	95.0	8.29
	90%	90.0	**4.67**	90.5	6.52	89.6	6.45	90.0	6.28
SVHN-10	100%	100	4.27	100	**3.86**	100	4.27	100	4.01
	95%	95.1	**1.83**	95.1	1.86	95.1	2.53	95.0	2.07
	90%	90.1	**1.01**	90.0	1.04	90.1	1.31	90.0	1.06
Cats & Dogs	100%	100	5.93	100	**5.72**	100	7.36	100	6.16
	95%	95.1	**2.97**	95.0	3.46	95.2	5.1	95.1	4.28
	90%	90.0	**1.74**	90.0	2.28	90.2	3.3	90.0	2.5

Table C.3: Performance at High Target Coverage. Similarly to the previous table, this repeats Table 4.3 but with the scan for the DG method corrected. Again note the reduced performance in the final column relative to Table 4.3.

Appendix D

Appendix to Chapter 5

D.1 Proofs Omitted from the Main Text

D.1.1 Proof of Theorem 5.4.1

Proof of lower bound. Notice that since $\mathcal{H}$ is one-sided learnable, it can learn any $h \in \mathcal{H}$ from below with $\mathsf{L} = 0$. Thus, given $m(\varepsilon, \delta, \lambda, \mathcal{H})$, and samples $(X_i, h(X_i))$ for any $h \in \mathcal{H}$, the scheme $\mathscr{A}$ recovers a function $\widehat{h}$ such that

$$\mu(h = 0, \widehat{h} = 1) \le \lambda$$
$$\mu(h = 0, \widehat{h} = 1) \le \varepsilon.$$

But then $\mu(\widehat{h} \neq h) \le \lambda + \varepsilon$ - i.e. $\mathscr{A}$ also serves as a realisable PAC learner with excess risk bounded by $\lambda + \varepsilon$. Thus, standard lower bounds for realisable PAC-learning can be invoked, for instance, that of §3.4 from the book [MRT18]. □

Proof of Upper Bound. We provide a scheme showing the same. To begin with, suppose that $\mathcal{H}$ is a finite class. Fix g, μ, and let $\mathcal{H}_\eta := \{h \in \mathcal{H} : \mu(g(X) = 0, h(X) = 1) \le \eta\}$. For finite $\mathcal{H}$, the scheme proceeds in two steps:

1. Testing: using m_1 samples (where m_1 is to be specified later), compute the empirical masses $\widehat{\ell}(h) := \widehat{\mu}\{h(X) = 1, g(X) = 0\}$ for every $h \in \mathcal{H}$. Let $\widehat{\mathcal{H}}_\lambda := \{h : \ell(h) < \lambda/2\}$.

2. Optimisation: Using m_2 samples (where m_2 is to be specified later), compute the empirical masses $\widehat{\mathsf{L}(h)} := \widehat{\mu}(h(X) = 0, g(X) = 1)$ for every $h \in \widehat{\mathcal{H}}_\lambda$. Return any $\widehat{h} \in \operatorname{argmin}_{\widehat{\mathcal{H}}_\lambda} \widehat{\mathsf{L}}(h)$.

The correctness of the above procedure is demonstrated by the following lemmata:

Lemma D.1.1. *If*

$$m_1 \ge \frac{24}{\lambda} \log(4|\mathcal{H}|/\delta),$$

then with probability at least $1 - \delta/2$,

$$\mathcal{H}_{\lambda/4} \subset \widehat{\mathcal{H}}_\lambda \subset \mathcal{H}_{3\lambda/4}.$$

The above is proved after the conclusion of this argument.

Lemma D.1.2. *If*

$$m_2 \ge \frac{2}{\varepsilon^2} \log(4|\mathcal{H}|/\delta),$$

then with probability at least $1 - \delta/2$,

$$|\widehat{\mathsf{L}(h)} - \mu(h = 0, g = 1)| \le \varepsilon$$

simultaneously for all $h \in \widehat{\mathcal{H}}_\lambda$.

Proof. The claim follows by Hoeffding's inequality and the union bound, noting that $|\widehat{\mathcal{H}}| \le |\mathcal{H}|$. □

Thus, for finite classes, the claim follows (with $d = \log |\mathcal{H}|$) by an application of the union bound, and noting that $\mathcal{H}_0 = \{h : \mu(h = 1, g = 0) = 0\} \subset \{h : \mu(h = 1, g = 0) \le \lambda/4\} = \mathcal{H}_{\lambda/4}$.

We now appeal to the standard generalisation from finite classes to finite VC-dimension classes. By the Sauer-Shelah lemma (see, e.g., §3.3 of [MRT18]), with m samples, a class of VC-dimension d breaks into at most $(em/d)^d$ equivalence classes of functions that agree on all data points, and the losses of functions in each equivalence class can be simultaneously evaluated and share the same generalisation guarantees. Let $\mathcal{H}'$ be formed by selecting one representative from each such class. We may run the above procedure for $\mathcal{H}'$, and draw the same conclusions so long as

$$\begin{aligned} m &\ge m_1 + m_2 \\ m_1 &\ge \frac{24}{\lambda} \left(d \log(em/d) + \log(4/\delta)\right) \\ m_2 &\ge \frac{1}{2\varepsilon^2} \left(d \log(em/d) + \log(4/\delta)\right) \end{aligned}$$

By crudely upper bounding the right hand sides above, this can be attained if

$$\frac{m}{\log em} \ge 24 \left(\frac{1}{\lambda} + \frac{1}{\varepsilon^2}\right) \left(d + \log(4/\delta)\right),$$

and the conclusion follows on noting that for $v \ge 2$, $u \ge 4v\log(v) \implies u/\log(eu) \ge v$. □

It remains to show Lemma D.1.1.

Proof of Lemma D.1.1. Let $\ell(h) := \mu(h = 1, g = 0)$. Note that for each h, $m_1\widehat{\ell(h)}$ is distributed as $\text{Binomial}(m_1, \ell(h))$. Further, for any $0 \le p \le q \le 1$, and any natural n, the $\text{Binomial}(n, q)$ distribution stochastically dominates $\text{Binomial}(n, p)$.

Thus, for any $h : \ell(h) < \lambda/4$,

$$\mu^{\otimes m_1}(\widehat{\ell}(h) \ge \lambda/2) \le P_{U\sim\text{Binomial}(m_1,\lambda/4)}(U \ge m_1\lambda/2) \le \exp\left(() - 3m_1\lambda/32\right),$$

where the final relation is due to Bernstein's inequality.

Similarly, for any $h : \ell(h) > 3\lambda/4$,

$$\mu^{\otimes m_1}(\widehat{\ell(h)} \le \lambda/2) \le P_{U\sim\text{Binomial}(m_1,3\lambda/4)}(U \le m_1\lambda/2) \le \exp\left(() - m_1\lambda/24\right).$$

For $m_1 \ge 24/\lambda \log(4|\mathcal{H}|/\delta)$, each of the above can be further bounded by $\delta/4|\mathcal{H}|$. The claim follows by the union bound. □

D.1.1.1 Alternate Generalisation Analyses

Note that the above proof utilises the finite VC property only to assert that on a finite sample, the hypotheses to be considered can be reduced to a finite number. Instead of the VC theoretic argument, one can then immediately give analyses via, say, L_1 covering numbers of the sets induced by the functions. Similarly, instead of beginning with finite hypotheses, we may instead directly uniformly control the generalisation error of the estimates for each function via the Rademacher complexity of the class $\mathcal{H}$, thus replacing Lemmas D.1.2, D.1.1 by a bound of the form $m(\varepsilon, \lambda, \delta, \mathcal{H}) \le \inf\{m : \mathfrak{R}_m(\mathcal{H}) + \sqrt{2\log(2/\delta)/m} \le \min(\lambda, \varepsilon)/2\}$, and further extensions via empirical Rademacher complexity. In addition, one can utilise more sophisticated analyses for more sophisticated algorithms.

The point of all this is to underscore that once one adopts the bracketing and

OSP setup, generalisation guarantees, and thus sample complexity bounds, follow the standard approaches in learning theory. This is not to say that these analyses may be trivial - for instance, in the above we have not shown tight sample complexity bounds at all.

D.1.2 Proof of Theorem 5.4.2

Proof of Upper Bound. We note that if $\frac{\mathrm{d}\mu}{\mathrm{dVol}} \geq \rho$, and we can locally predict in a region of volume P, then we can immediately locally predict in a region of μ-mass ρP. Thus, it suffices to argue the claim for the Lebesgue mass on $[0,1]^p$.

Since we have access to κ cuboids in $\mathcal{R}_\kappa^{0,1}$, we can capture any κ of the cuboids induced in the minimal partition aligned with g for any $g \in \mathcal{G}$. In particular, when approximating from below, we will choose h^- to be 1 on some κ of the cuboids contained in $\{g = 1\}$, and 0 otherwise, and similarly for approximating from above (denoted h^+). Naturally, we will 'capture' the cuboids with the biggest volume (more generally, biggest μ mass). Notice that this construction trivially yields $h^- \leq g \leq h^+$.

To finish the argument, fix an arbitrary $g \in \mathcal{G}$. Let $\mathscr{P}$ be a partition aligned with g that is V-regular, and further, has the largest total number of parts possible.[1] Suppose that there are $\mathscr{P}_1$ parts in $\mathscr{P}$ on which g is 1, and $\mathscr{P}_0$ on which it is 0. By the maximality, it must be the case that each rectangle contained in each part of $\mathscr{P}$ has volume less than $2V$, since otherwise we can split this part while maintaining V-regularity. Further, since the mass contained outside of the rectangle in each part is at most V, it follows that $3V(\mathscr{P}_0 + \mathscr{P}_1) \geq 1$ by the union bound. Thus, $\mathscr{P}_0 + \mathscr{P}_1 \geq 1/3V \geq \kappa/3$.

Now, by the above construction, we can capture a volume of at least $(\min(\kappa, \mathscr{P}_0) + \min(\kappa, \mathscr{P}_1))V$, which exceeds $\kappa V/3$. □

Proof of Lower Bound. Divide $[0,1]^p$ into $N = \lfloor 1/V \rfloor$ congruent, disjoint rectangles. Note that since the faces of these rectangles have codimension ≥ 1, they have volume 0. Thus, we need not worry about how they are assigned in the following, and we will omit these irrelevant details in the interest of clarity.

We set $\mathcal{G}$ to be the class of 2^N functions obtained by colouring each of the N boxes as 0 or 1. This class is trivially V-regular.

[1] such a partition exists because V-regularity implies that the number of parts is uniformly bounded by $1/V$.

Now, notice that any time a function h is approximating a function $g \in \mathcal{G}$ from above, it should either attain the value 0 on a whole box, or attain the value 1 on a whole box - if g is 1 on a box, then h is forced to be 1. If g is instead 0, and h dips down to take the value 0 at any point, then rising up to 1 is lossy in that it increases the loss $\mathsf{L}(h, g, \mathrm{Vol})$ while offering no reduction in the expressivity of the class $\mathcal{H}$. Thus, we may restrict attention to classes $\mathcal{H}$ such that all functions contained in them are constant over the boxes described.

Given the above setup, the entire problem is equivalently described by restricting the domains of $\mathcal{G}, \mathcal{H}$ to the centres of the above boxes, and the measure Vol to the uniform measure over these centres. We henceforth work in this space. The domain of the functions in $\mathcal{G}, \mathcal{H}$ is now the abstract set $[1:N]$.

Suppose every $g \in \mathcal{G}$ can be budget learned with budget at most $1 - \Delta/N$ in this measure (where Δ is some integer because the space is discrete and the distribution is rational). Let (h_g^+, h_g^-) be the appropriate bracketing functions that minimise budget for g, and let $\mathcal{I}_g$ be the points where $h_g^+ = h_g^-$. The budget constraint forces that $|\mathcal{I}_g| \geq \Delta$. Notice that outside of $\mathcal{I}_g$, h_g^+ must take the value 1 and h_g^- must take the value 0 - indeed, if $h_g^+(i)$ was 0, then since $0 \leq h_g^-(i) \leq h_g^+(i)$, $h_g^-(i) = 0$, and then $i \in \mathcal{I}_g$.

But, on $[1:N] \sim \mathcal{I}_g$, g must either be predominantly 1 or 0, and then respectively, must agree with h_g^+ or h_g^- on at least $(N - |\mathcal{I}_g|)/2$ points. This means that there exists a $h'_g \in \mathcal{H}$ (which is either h_g^+ or h_g^-) such that

$$|\{i : h'_g(i) = g(i)\} \geq |\mathcal{I}_g| + \frac{N - |\mathcal{I}_g|}{2} \geq \frac{N + \Delta}{2}.$$

With this setup, we invoke the following statement

Lemma D.1.3. *If a class of functions $\mathcal{F}$ on $[1:N]$ is such for every $\{0,1\}$-valued function on $[1:N]$, there exists a $f \in \mathcal{F}$ that agrees with it on at least $(N+\Delta)/2$ points, then*

$$\mathrm{VC}(\mathcal{F}) \geq \frac{3\Delta^2}{2(N+\Delta)\log(eN)} \geq \frac{3\Delta^2}{4N\log(eN)}.$$

Notice that since $N \geq \Delta$, Invoking the above, and the fact that the VC-dimension of $\mathcal{H}$ is at most d, it follows that (for $N \geq 3$)

$$\frac{3\Delta^2}{8N\log N} \leq d \iff \Delta \leq \sqrt{3dN\log N},$$

from which the claim is immediate on recalling that $1/V \geq N \geq 1/V - 1$. □

Proof of Lemma D.1.3. Identify all $\{0,1\}$ labellings as above with the cube $\{0,1\}^N$, and similarly the patterns achieved by $\mathcal{F}$ as a subset of the same. The hypothesis is then equivalent to saying that for every point $p \in \{0,1\}^N$, there exists a point $f \in \mathcal{F}$ such that $d_{\mathrm{H}}(p,f) \leq \frac{N-\Delta}{2}$, were d_{H} is the Hamming distance. But then $\mathcal{F}$ is a $(N-\Delta)/2$-cover of the Boolean hypercube.

By a standard volume argument, it then must hold that

$$|\mathcal{F}| \geq \frac{2^N}{\sum_{i=0}^{(N-\Delta)/2} \binom{N}{i}} \geq e^{+\frac{3}{2}\frac{\Delta^2}{N+\Delta}}$$

where the final inequality follows on noting that the right hand side of the first inequality is 1 divided by a lower tail probability for N independent fair coin flips, and then invoking Bernstein's inequality.

However, by the Sauer-Shelah Lemma, if $d \leq N$ is the VC-dimension of $\mathcal{F}$, then the number of elements in it is at most

$$\sum_{i=0}^{d} \binom{N}{i} \leq \left(\frac{eN}{d}\right)^d.$$

Relating these, we have

$$e^{+\frac{3}{2}\frac{\Delta^2}{N+\Delta}} \leq (eN/d)^d \iff \frac{3\Delta^2}{2(N+\Delta)\log(eN/d)} \leq d.$$ □

D.1.3 Proof of Theorem 5.4.3

These lower bounds are proved similarly to the lower bound from the previous section: principally, they use the fact that any non-trivial budget learner also yields non-trivial coverings, and construct function classes of limited VC dimension with large covering numbers.

Proof of the bound (i). Let $S := \{x_1, \dots, x_D\}$ be a set of shattered points. The measure μ_S is set to the uniform distribution on S. The restriction $\mathcal{G}_{|S}$ consists of all $\{0,1\}$-valued functions on D points. If $\mathcal{H}$ can budget learn this with respect to μ_S with budget $1 - \Delta/D$, then $\mathcal{H}_{|S}$ is a $(D-\Delta)/2$covering of $\{0,1\}^S$. Invoking Lemma

D.1.3 just as in the proof of the lower bound in the previous section, we get that $\frac{\Delta}{D} \ge \sqrt{3\frac{\mathrm{vc}(\mathcal{H})}{D}\log(\frac{eD}{\mathrm{vc}(\mathcal{H})})}$. □

Proof of the bound (ii). We use a class on $[1:N]$, constructed by [Hau95] that is known to have large packing number. Note that the same class is used as an example of a simple budget-learnable class in §5.4.4. The class is defined as follows: Suppose D divides N. Let $\mathcal{F}$ be the class of single thresholds on $[1:N/D]$, i.e.$\mathcal{F} = \{f_k, k \in [0:N/D+1]\}$, where $f_k(i) := \mathbb{1}k \le i$. $\mathcal{F}$ trivially has a VC-dimension of 1. $\mathcal{G}$ is generated as a tensor product of D copies of $\mathcal{F}$ placed on a partition of $[1:N]$. Concretely, we may say that each $g \in \mathcal{G}$ can be represented as D functions $(f_{k_1}, f_{k_2}, \dots, f_{k_D}) \in \mathcal{F}^{\otimes D}$ for some $k_1, \dots k_d \in [0:N/D+1]$ such that for $i \in [jN/D+1:(j+1)N/D]$ for any $j \in [0:D-1]$, $g(i) = f_{k_j}(i)$.

[Hau95] shows that for this class, under the uniform measure on $[1:N]$, the k-packing number is at least $\frac{(1+N/D)^D}{2^D\binom{k+D}{D}}$. Now recall that the $k/2$-covering number must exceed the k-packing number for any set and metric. Further, a budget of $1-\Delta/N$ implies a $\frac{N-\Delta}{2}$-covering. The budget requirement imposes the condition $N - \Delta \le \mathsf{B}N$. Thus, invoking Sauer-Shelah as in the proof of Lemma D.1.3, we obtain

$$\begin{aligned}\left(\frac{eN}{d}\right)^d &\ge \left(\frac{N+D}{2e(N+D-\Delta)}\right)^D \\ &\ge \left(\frac{N+D}{2e(\mathsf{B}N+D)}\right)^D \\ &\ge \left(\frac{1}{4e\mathsf{B}}\right)^D\end{aligned}$$

where we have used that $\mathsf{B}N \ge D$ in the final line. The above bound is non-vacuous only if $4e\mathsf{B} < 1$.

The case $\mathsf{B} < D/N$ is not discussed in the theorem, since it is a vanishingly small budget, but by the above, in this case we get a lower bound of $(N/4eD)^D$ in the above, giving, for $D \lesssim N^{1-\varepsilon}$ for some $\varepsilon > 0$, a bound of $d = \Omega(D)$ in this setting. □

D.1.4 Proofs of budget claims made in §5.4.4

Proof for sparse VC classes. fix any g. We pick the function that is 1 on the d choices of $i \in g^{-1}(1)$ with the largest total μ-mass as the lower approximation, and the constant 1 as the approximation from above. □

Proof for Tensorised class. The class naturally breaks the domain into D equal parts, and places a threshold on each. We choose the $d-1$ parts with largest μ-mass, and place a threshold there. Lastly, we collate the remaining parts into one set, and we place the constant functions 1 and 0 on this. A tensorisation of these function classes demonstrates the claim. □

Proof for Convex Polygons. Instead of approximation from above and from below, we will adopt the more natural terminology of inner and outer approximation. As the class is closed under $f \mapsto 1-f$, to show budget learnability with budget B, it suffices to show that for any polygon P with D vertices and any measure μ, there exist polygons $P_{\text{in}} \subset P \subset P_{\text{out}}$ of d vertices such that $\mu(P_{\text{in}}) \geq (1-\mathsf{B})\mu(P)$ and $\mu(P^c_{\text{out}}) \geq (1-\mathsf{B})\mu(P^c)$. This follows since the cloud query points are precisely those in $P_{\text{out}}/P_{\text{in}})$, which has mass $\mu(P_{\text{out}}) - \mu(P_{\text{in}}) \leq 1 - (1-\mathsf{B})(1-\mu(P)) - (1-\mathsf{B})\mu(P) = 1 - (1-\mathsf{B}) = \mathsf{B}$.

Inner Approximation: We offer a direct proof. Consecutively number the vertices of P as $[1:D]$. Form the d-gon P^1 using the vertices $[1:d]$. Remove this polygon from P and relabel $1 \mapsto 1, d \mapsto 2, \dots, n \mapsto n+2-d, \dots$. Contuining this process $m := \lceil D/d-2 \rceil$ times partitions P into m d-gons $P^1, \dots, P^m$. By the union bound, $\sum \mu(P^i) \geq \mu(P)$. But then there must exist at least one d-gon $P_{\text{in}} \subset P$ such that $\mu(P_{\text{in}}) \geq \mu(P) \geq \frac{1}{\lceil D/d-2 \rceil}\mu(P)$.

Outer Approximation: Recall that $d \geq 4$, and $D \geq d$. We will show that for any D-gon P there exists a d-gon P_{out} containing it such that $\mu(P^c_{\text{out}}) \geq \frac{d-2}{D-2}\mu(P^c)$.

We induct on D. As a base case, for $D = d$, the claim holds trivially since P itself may serve. Let us assume the claim for D-gons, and let P be a $D+1$-gon. Note that since $D \geq 4, D+1 \geq 5$. Thus, P has at most two pairs of consecutive exterior angles that are each exactly $\pi/2$ (since the sum of all exterior angles is 2π, and P has at least 5 exterior angles). For any side such that the two exterior angles are not both $\pi/2$, the sides preceding and following it (in the cyclic order) may be extended to meet at some point. This yields a triangle with this side as a base. Since such an extension can be done for at least $D+1-2 = D-1$ sides, this yields $D+1 \geq J \geq D-1$ triangles $\triangle_1, \triangle_2, \dots, \triangle_J$. Now notice that for each $j \leq J$, $Q_j := P \cup \triangle_j \supset P$ is a D-gon. Further, by the union bound, $\sum \mu(\triangle_j) \leq \mu(P^c)$, and thus there exists a triangle $\triangle_{i^*}$ such that $\mu(\triangle_{i^*}) \leq \mu(P^c)/J$, and thus $\mu(Q^c_{i^*}) \geq \frac{J-1}{J}\mu(P^c) \geq \frac{D-2}{D-1}\mu(P^c)$. Now, by the induction hypothesis, there exists a d-gon P_{out} containing Q_{i^*} (and hence P) such that $\mu(P^c_{\text{out}}) \geq \frac{d-2}{D-2}\mu(Q^c_{i^*}) \geq \frac{d-2}{D-1}\mu(P^c)$. This concludes the argument.

Thus, we can attain the budget

$$\mathsf{B} = 1 - \min\left(\frac{1}{\lceil \frac{D}{d-2} \rceil}, \frac{d-2}{D-2}\right) = 1 - \left\lceil \frac{D}{d-2} \right\rceil^{-1}. \qquad \square$$

D.2 Experiments

D.2.1 Losses and algorithms for methods listed in §5.5

We list the general approach taken for each of the methods we compare to. More precise details very between datasets, and are described in subsequent sections. Note that all models are trained on GPUs using stochastic gradient descent for linear models and ADAM for deep networks. In each case, a multitude of models are trained by scanning over values for the relevant Lagrange multiplier/regularisation weight. The collection of models so obtained is tuned, and then a model finally selected for each target accuracy via procedures detailed in §D.2.5.

Bracketing The general approach, and a formulation for generic loss functions follows the design of Chapter 4. The exact loss formulation used in the experiments is the following,

$$\hat{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} -1_{g(x_i)=1} \log\left(h_\theta(x_i)\right) - \xi 1_{g(x_i)=0} \log\left(1 - h_\theta(x_i)\right) \tag{D.1}$$

where ξ is a hyper parameter between two components of loss function. The term multiplying ξ is the constraint, which imposes a high cost in case of a leakage. The other term in the loss objective pushes the model to increase true positives. For example, if ξ is 0, local model always predicts 1 and it has maximum leakage and minimum budget. If ξ is $+\infty$, the local model always predicts 0 and it has minimum leakage and maximum budget.

Local Thresholding We first train a local predictor using the cross entropy loss and freeze it. We rank the examples based on maximum of the prediction probabilities. We select a threshold and the predictor uses cloud model if its current maximum probability is lower than threshold. We attain different budget values by changing this threshold.

Alternating Minimisation [NS17a] we follow the ADAPT-LIN procedure from this paper, which is an alternative minimisation scheme between an auxiliary q and local predictors & gating. Since we don't have feature costs in our setting, we assumed $\gamma = 0$ in our experiments. We stopped the procedure if the q vector converges, or if a predefined number of iterations - in our case 10 - is exceeded. Different budget values are obtained by sweeping values of the regularisation parameter - in this paper called λ.

Sum relaxation [CDM16] utilising the relaxation as developed in this paper, we use the loss $L_{MH}(h, r, x, y)$ formulated within as a loss function to train a neural network. This is optimised with several values of the regularisation parameter, c, to obtain different usage values.

Selective Net [GE19] we follow the architectural augmentations and losses as prescribed by this paper. We train the network with auxiliary head and ignore this part during inference time. Again, this is performed for several values of the Lagrange multiplier, called c here as well.

D.2.2 Synthetic Data

Cloud Classifier A training dataset of 2.5K points was sampled uniformly from the set $[-10, 10] \times [-10, 10]$. The complex classifier's decision boundary can be expressed as

$$\mathbb{1}x + 4x^2 + 3x^3 + 3x^4 + y + y^2 + y^3 + y^4 + 5xy^2 + 30x^2y < 1000$$

where x, y are the coordinates of the data point.

Local Classifier Weak learners are restricted to axis-aligned conic sections, which may be implemented as linear classifiers which see input features x, y, x^2, y^2.

Training Details Each weak learner model has hyper parameters which are adjusted to observe the power of the methods. As an example, learning rates are chosen in the range of $[10^{-5}, 10^{-2}]$, ξ value for bracketing model is chosen in the range of $[1, 3]$, λ values for alternating minimisation are chosen in the range of $[0.25, 0.75]$ and c values for the sum relaxation method are chosen in the range of $[0, .3]$. After obtaining several models, the best models are reported based on the true error rates and true usages.

D.2.3 MNIST Odd/Even

Cloud Classifier We implement a LeNet architecture with 6 filters in the first convolution layer, 16 filters in the second convolution layer, 120 neurons in the first fully connected layer and 84 neurons in the first fully connected layer. Kernel size for convolution layers is chosen to be 5. Overall, this model has $43.7K$ parameters. Learning rate is chosen to be 10^{-3} and it is halved in every 20 epochs for a total of 60 epochs using 64 as batch size. L_2 regularisation of 10^{-5} is applied. The model attains 99.46% test accuracy.

Local Classifier Linear classifiers are adopted as weak learner architecture - these have $1.57K$ parameters, and no convolutional structure. Half of the training set $(30K)$ is randomly chosen to be weak learner dataset. Within this dataset, 90% $(27K)$ is kept as training set for and 10% ($3K$) as validation. Training and validation sets for each of the methods are kept the same to ensure a fair comparison. The local model attains 89.79% test accuracy.

Training Details For each model, learning rate is chosen to be 10^{-2} and it is halved in every 25 epochs for a total of 120 epochs. Batch size is chosen to be 64 and L_2 regularisation of 10^{-5} is applied. For bracketing, ξ values are chosen in the range of $[0, 24]$ for a total of 21 values. For alternating minimisation, λ values are swept in the range $[0, 1]$ for a total of 25 values and a maximum of 10 alternative minimisation rounds are allowed. For the sum relaxation, c is chosen in the range $[0, .495]$ for a total of 25 values. For the selective net, c values are chosen in range $[0, 1]$ for a total of 25 values. We note here that the auxiliary head in the selective net, which serves in deep networks as a way to improve feature extraction, is ineffective in this linear setting.

D.2.4 CIFAR Random Pair

Cloud Classifier We pick ResNet32 [HZRS16] as the high-powered model and trained it, with configurations as described by [Ide19], on the full multi-class CIFAR training data. This model has $.46M$ parameters.

Local Classifiers We pick a narrow LeNet model as weak learner that has 3 filters in the first and second convolution layers, and 15 neurons in the first fully connected layer. Kernel size for convolution layers is chosen to be 5. Overall, this weak model has $1,628$ parameters.

Procedure for training For each run, we choose 2 classes out of 10 CIFAR classes randomly and extract the subset of the dataset corresponding to this couple. The cloud classifier is obtained using the pre-trained ResNet32 and only retraining the prediction layer while keeping the backbone frozen for this binary dataset. Learning rate is chosen to be 10^{-2} and it is halved after 50 epochs for a total of 100 epochs. Batch size is chosen to be 64 and L_2 regularisation of 10^{-5} is applied. The model attains on average 98.38% test accuracy.

For the weak learners, 60% (6K points) of the training set is randomly chosen to be the training dataset. From this, 83.3% (5K) is kept as training set for and 16.7% (1K) culled for validation. The model attains on average 90.94% test accuracy. Training and validation set are kept the same across methods to have a fair comparison.

Training Details Learning rate is chosen to be 10^{-3} and it is halved in every 75 epochs for a total of 300 epochs. Batch size is chosen to be 64 and L_2 regularisation of 10^{-5} is applied. For bracketing model, values for ξ are chosen in the range of $[0, 65]$ for a total of 36 values. For alternating minimisation λs are swept in the range $[0, 1]$ for a total of 40 values and a maximum of 10 alternative minimisation rounds are allowed. For the sum relaxation method c values are chosen in range $[0, .495]$. For each of the above methods, all the networks are warm started using the parameters of the local model. Note each of the previous methods implement two Narrow LeNets - for bracketing these are the two one-sided learners, while for the other two, these are gates and predictors. For the selective net, c values are chosen in range $[0, 1]$ for a total of 40 values. Warm starting this network leads to lowered performance than random initialisation, and so the latter values are reported.

The above procedure is performed for 10 trials of random classes of CIFAR. These classes are listed in Table D.1 below, along with usages attained for the bracketing and selective net methods in these cases. Only these two methods are reported here since they are the most competitive of the five.

D.2.5 Model Selection Process

For each value of the Lagrange multiplier/regularisation constant chosen in the above training methods, we receive a model (or a pair of models, as appropriate). Let this collection of models be $\mathcal{M}$. These models have real valued outputs in the range $[0,1]$, and a decision needs to be extracted from these. In order to provide sufficient granularity to the models that they be able to match any required target accuracy, we vary the threshold of output value at which the models' decisions go from 0 to 1. This process differs in details for different methods. The tuning is performed

Local Thresholding In this case $\mathcal{M}$ is a singleton. We compute the cross entropy of the classifier's output and abstain if this cross entropy is larger than a threshold τ that is selected as follows: the values of τ considered are obtained by computing the cross entropies of the model outputs on each of the training points. On validation data, usages and accuracy are computed for the models which thresholds at each of the considered thresholds. At a given target accuracy, the value of τ which yields at least this accuracy on the validation data with the smallest usage is selected.

Bracketing Note that each $m \in \mathcal{M}_{\text{bracketing}}$ contains two models $(m_{\text{below}}, m_{\text{above}})$ which are respectively approximations from above and below - these may be trained with different ξ, thus giving a total of $|\Xi|^2$ models. Suppose the target accuracy is $1-\alpha$. Let the training data have size T. Using the training data, for every $i \in [0 : \alpha T]$, we determine pairs of thresholds $\tau_m(i) = (\tau^m_{\text{below}}(i), \tau^m_{\text{above}}(i))$ such that the leakages of $(m_{\text{below}}, m_{\text{above}})$ on the training data are exactly i/T each. This then gives us a total of at most $|\Xi|^2 \times \alpha T$ possible model-threshold pairs, represented as $(m, \tau_m(i))$.

Now, each of these tuples is evaluated on the validation data, with usages and accuracies computed. Again, the pair of models and thresholds with the smallest usage that exceeds the target accuracy on the validation set is selected.

Alternating Minimisation *and* Sum Relaxation *and* Selective Net Each $m \in \mathcal{M}$ is a pair (γ, π), where the former is the gate. Again, on the training data, the value taken by γ on each training point is recorded. This gives all the thresholds that may be selected for the gating function. Now, each m and corresponding choice of threshold may be evaluated on the validation set, and we select the ones which match the accuracy requriement and show the lowest usage.

D.2.6 Tables Omitted from the Main Text

Class Pair	Bracketing	Sel. Net.	Gain
0 - 3	0.304	0.364	1.199×
6 - 4	0.452	0.526	1.164×
5 - 2	0.616	0.631	1.026×
6 - 1	0.095	0.122	1.296×
9 - 3	0.220	0.211	0.961×
8 - 1	0.235	0.381	1.619×
7 - 4	0.615	0.646	1.050×
8 - 7	0.059	0.091	1.538×
4 - 0	0.195	0.315	1.620×
6 - 7	0.152	0.179	1.177×

Table D.1: Usages and relative gain for bracketing and selective net [GE19] methods at 99% target accuracy for 10 CIFAR random pairs. These two methods uniformly have the lowest usages, and hence the others are omitted. All models achieve test accuracy in the range 98.1-99.3% test accuracy. Notice that the gains have a large variance, but with a skew towards entries greater than 1.

Task	Target	Bracketing			Local Thr.			Alt. Min.			Sum relax.			Sel. Net.			Gain
	Acc.	Acc.	Usg.	ROL	Acc.	Usg.	ROL	Acc.	Usg.	ROL	Acc.	Usg.	ROL	Acc.	Usg.	ROL	
MNIST Odd/Even	0.995	0.994	0.457	2.19	0.995	0.653	1.53	0.991	0.830	1.20	0.997	0.785	1.27	0.996	0.658	1.52	1.431×
	0.990	0.990	0.387	2.58	0.991	0.515	1.94	0.985	0.740	1.35	0.992	0.651	1.54	0.992	0.544	1.84	1.332×
	0.980	0.982	0.299	3.35	0.983	0.358	2.79	0.974	0.604	1.66	0.992	0.651	1.54	0.985	0.423	2.37	1.199×
CIFAR Random Pair	0.995	0.991	0.363	4.01	0.996	0.510	2.25	0.991	0.854	1.19	0.997	0.620	2.07	0.992	0.436	3.04	1.280×
	0.990	0.986	0.294	5.66	0.991	0.399	3.41	0.986	0.754	1.40	0.994	0.488	3.31	0.987	0.347	4.30	1.265×
	0.980	0.975	0.214	9.97	0.983	0.276	6.38	0.975	0.611	1.87	0.986	0.345	5.81	0.977	0.257	11.67	1.195×

Table D.2: Performances on BL tasks studied. This table repeats the entries of Table 5.2, with the addition of a column indicating the test accuracy attained by the models. Note that these fluctuate in the range -0.05 to +0.02 of the target, as can be expected from any selection method.

Appendix E

Appendix to Chapter 6

E.1 An Adversarial Anytime Uniform Law of Large Numbers For Probing Binary Sequences

E.1.1 Proofs of Lemma6.3.1

We begin with a simple lemma that underlies the remaining argument. Below, κ is chosen so that $\kappa''(0) = 1$.

Lemma E.1.1. *Let $\mathscr{F}_t, U_t, B_t, W_t, \widetilde{W}_t$ be as in Lemma 6.3.1. Let $\overline{p} = 1 - p$. Then for any $\eta \in \mathbb{R}$, the process*

$$\xi_t^\eta := \exp\left(\eta(W_t - \widetilde{W}_t/p) - \kappa(\eta)V_t\right)$$

is a non-negative, $\mathscr{F}_t$-adapted martingale, where

$$\begin{aligned} V_t &= \frac{\overline{p}}{p}W_t, \\ \kappa(\eta) &= \frac{p}{\overline{p}}\log\left(pe^{-\eta\overline{p}/p} + \overline{p}e^{\eta}\right). \end{aligned}$$

Proof. The nonnegativity of ξ_t^η is trivial, and it is $\mathscr{F}_t$-adapted since it is a deterministic function of the adapted processes $W_t, \widetilde{W}_t$. We need to argue that ξ is a martingale. To this end, observe that since $W_t = \sum_{s<t} U_s, \widetilde{W}_t = \sum_{s<t} U_s B_s$,

$$\xi_t^\eta = \xi_{t-1}^\eta \cdot \exp\left(\eta U_t(1 - B_t/p - \overline{p}\kappa(\eta)/p)\right).$$

Due to the independence of B_t from $\sigma(U_t, \mathscr{F}_{t-1})$, we have

$$\begin{aligned}
&\mathbb{E}[\exp\left(\eta U_t(1 - B_t/p)\right) | \mathscr{F}_{t-1}, U_t] \\
&= \left(p e^{-\eta U_t \overline{p}/p} + \overline{p} e^{\eta U_t}\right) \\
&\overset{*}{=} \left(p e^{-\eta \overline{p}/p)} + \overline{p} e^{\eta}\right)^{U_t} = \exp\left(\frac{\overline{p}}{p} U_t \kappa(\eta)\right),
\end{aligned}$$

where the equality marked $*$ exploits the fact that U_t is $\{0,1\}$-valued. Rearranging, we have

$$\mathbb{E}\left[\exp\left(\eta U_t(1 - B_t/p) - \frac{\overline{p}}{p} U_t \kappa(\eta)\right) \middle| \mathscr{F}_{t-1}, U_t\right] = 1,$$

and exploiting the tower rule, we conclude that

$$\mathbb{E}[\xi_t^\eta | \mathscr{F}_{t-1}] = \xi_{t-1}^\eta \mathbb{E}\left[\mathbb{E}\left[\exp\left(\eta U_t(1 - B_t/p) - \frac{\overline{p}}{p} U_t \kappa(\eta)\right) \middle| \mathscr{F}_{t-1}, U_t\right] \middle| \mathscr{F}_{t-1}\right] = \xi_{t-1}^\eta.$$

□

The following argument heavily uses the techniques of Howard et al. [HRMS20], and assumes familiarity with the same. It also exploits the property that only the upper tail of Δ_t is being controlled, although this is extended in the following section.

Proof of Lemma 6.3.1. We define the deviation of W_t from $\widetilde{W}_t$ as

$$\Delta_t := W_t - \frac{\widetilde{W}_t}{p}.$$

Notice that $\Delta_0 = 1$. As a result of the above lemma, Δ_t is a 1-sub-κ process with the associated variance process V_t, in the sense of Definition 1 of Howard et al. [HRMS20]. In particular, since κ is the (normalised) cumulant generating function of a centred Bernoulli random variable taking values $\{-\overline{p}/p, 1\}$, the process is sub-binary. Further, since $p < {}^1\!/_2, \overline{p}/p > 1$, and thus the process is sub-gamma, with the scale parameter $c = 0$. [HRMS20, §3.1, and Prop.2].

We can thus invoke the line-crossing inequality of Corollary 1, part c) of Howard et al., instantiated with $c = 0$ to find that for any $x, m > 0$

$$\mathbb{P}\left(\exists t : \Delta_t \geq x + \mathfrak{s}(x/m)(V_t - m)\right) \leq \exp\left(-\frac{x^2}{2m}\right),$$

where [HRMS20, Table 2]

$$\mathfrak{s}(x/m) = \frac{x}{2m}.$$

Plugging these in, we observe that

$$\mathbb{P}\left(\exists t : \Delta_t \geq \frac{x}{2} + \frac{x}{2m} V_t\right) \leq \exp\left(-\frac{x^2}{2m}\right).$$

Now notice that if $V_t \geq m$, then $x/2 + (x/2m)V_t \leq (x/m)V_t$. Therefore, we can conclude that

$$\mathbb{P}\left(\exists t : \Delta_t \geq \frac{x}{m} V_t, V_t \geq m\right) \leq \exp\left(-\frac{x^2}{2m}\right),$$

and substituting $V_t = \frac{\overline{p}}{p} W_t, \Delta_t = W_t - \widetilde{W}_t/p$,

$$\mathbb{P}\left(\exists t : \frac{\widetilde{W}_t}{p} \leq \frac{mp - x\overline{p}}{pm} W_t, W_t \geq \frac{pm}{\overline{p}}\right) \leq \exp\left(-\frac{x^2}{2m}\right).$$

Now, if we choose $m = \frac{\overline{p}}{p}(x + 1/p)$ it follows that

$$\forall W_t \geq \frac{p}{\overline{p}} m, \frac{pm - x\overline{p}}{mp} W_t \geq \frac{1}{p},$$

and thus

$$\mathbb{P}\left(\exists t : \frac{\widetilde{W}_t}{p} \leq \frac{1}{p}, W_t \geq \frac{1}{p} + x\right) \leq \exp\left(-\frac{px^2}{2(1/p + x)\overline{p}}\right),$$

and choosing $x \geq 1/p$ further ensures that

$$\mathbb{P}\left(\exists t : \widetilde{W}_t \leq 1, W_t \geq 2x\right) \leq \exp\left(-\frac{px}{4\overline{p}}\right).$$

Now, setting $x = \max\left(\frac{1}{p}, \frac{4\overline{p}}{p}\log(1/\delta)\right)$ leaves us with

$$\mathbb{P}\left(\exists t : \widetilde{W}_t \leq 1, W_t \geq \max\left(\frac{2}{p}, \frac{8\overline{p}}{p}\log(1/\delta)\right)\right) \leq \delta.$$

The conclusion follows on observing since $p < 1/2, 8\overline{p} \geq 4$, and thus, for $\log(1/\delta) \geq 1/2$, $\frac{2}{p} \leq 8\frac{\overline{p}}{p}\log(1/\delta)$. □

E.1.2 An improved ALLN via a Self-Normalised Law of Iterated Logarithms

The line-crossing inequalities we utilised in the previous subsection can be stitched together, by picking an exponentially increasing set of xs, and optimising the ms at each, to yield a curve crossing inequality, which in effect determines a curve that the deviations are unlikely to cross. We use the results of Howard et al. [HRMS18] that produce non-asymptotic constructions.

For our purposes, note that the processes Δ_t and $-\Delta_t$ are both sub-Gamma with variance process V_t, with the scale parameters $c_+ = 0$ and $c_- = \frac{1}{3} \cdot \frac{1-2p}{p}$ respectively. The former property is useful for controlling the upper deviations of Δ_t, and the latter for the lower deviations. Note that since the scale parameter c_+ is 0, the upper tails in the following can be improved, but for ease of presentation we will just set $c = |c_+| = c_-$ in the following.

Using Theorem 1 of Howard et al. [HRMS18] twice - for Δ_t and $-\Delta_t$, and instantiating it with $\eta = e, h(k) = \frac{\pi^2 k^2}{6}$ yields that for the sub-gamma process Δ_t with scale parameter $\leq c$, and variance process V_t, and any constant $m > 0$, and for the functions

$$\begin{aligned} S_{m,\delta}(v) &= 2\sqrt{v\ell_{m,\delta}(v)} + c\ell_{m,\delta}(v), \\ \ell_{m,\delta}(v) &= \log\frac{\pi^2}{6} + 2\log\log\frac{v}{m} + \log\frac{2}{\delta}, \end{aligned}$$

the following bound holds true

$$\mathbb{P}(\exists t : |\Delta_t| \geq \mathcal{S}_{m,\delta}(\max(V_t, m)) \leq \delta.$$

The curve $S(\max(V_t, m))$ can be simplified upon observing that

$$\{\exists t : V_t \geq m, |\Delta_t| \geq \mathcal{S}_{m,\delta}(V_t)\} \subset \{\exists t : |\Delta_t| \geq \mathcal{S}_{m,\delta}(\max(V_t, m))\}.$$

With the above in hand, set $m = \overline{p}/p$, so that $W_t \ge 1 \iff V_t \ge m$, and observe that $\log(\pi^2/6) < 1$. The following bound is immediate upon recalling that $V_t = \frac{\overline{p}}{p}W_t, \Delta = W_t - \widetilde{W_t}/p$.

Theorem E.1.2. *In the setting of Lemma 6.3.1,*

$$\mathbb{P}\left(\exists t : W_t \ge 1, |W_t - \widetilde{W_t}/p| \ge 2\sqrt{\frac{\overline{p}W_t}{p}\left(2\log\left(\frac{2e}{\delta}\log W_t\right)\right)} + \frac{2\log\left(\frac{2e}{\delta}\log W_t\right)}{3p}\right) \le \delta.$$

Technically, the $\log\log W_t$ is not always defined in the above. This should be read as $\log(\max(1, \log W_t))$ to handle edge cases - alternately, it can be handled by replacing $W_t \ge 1$ by $W_t \ge 3 > e$ in the above.[1]

Notice that the bound above has the correct form when taking into account the behaviour of binomial tails, which $\widetilde{W_t}$ behaves like. Indeed, if W is some natural number valued random variable, and $\widetilde{W}|W \sim \mathrm{Bin}(W, p)$, then Bernstein's inequality [BLM13, Ch. 2] states that

$$\mathbb{P}\left(|W - \widetilde{W}/p| \ge C\sqrt{\overline{p}\frac{W}{p}\log(2/\delta)} + C\log(2/\delta)\right) \le \delta,$$

which entirely parallels the form of the above theorem, barring the $\log\log W_t$ blowup due to the uniformity over time.

The above analysis was inspired by studying the recent work of Ben-Eliezer and Yogev [BY20], on adversarial sketching - their goal was to maintain an estimate of the incidence of a process within a given set (and more generally, within sets in a given system) while using limited memory, and they analysed a similar sampling approach, showing via an application of Freedman's inequality that [BY20, Lemma 4.1]

$$\mathbb{P}\left(|W_T - \widetilde{W}_T/p| \ge C\sqrt{\frac{T}{p}\log(2/\delta)} + C\frac{\log(2/\delta)}{p}\right) \le \delta.$$

[1] In a similar vein of edge-cases, if $W_t < 1 \implies W_t = 0$, then $0 \le \widetilde{W_t} \le W_t = 0$, and thus the bound extends to all possible values of W_t.

This essentially amounts to using the crude bound $W_T \le T$. The same paper, in Theorem 1.4 and associated lemmata argues that the Reservoir Sampler [BY20, §2] of size $\sim pT$ controls deviations uniformly over time at scale $\sqrt{\frac{T}{p}\log\frac{\log T}{\delta}}$, and it was asserted that the Bernoulli Sampler cannot attain such a 'continuous robustness'[BY20, §1]. The above result improves upon this in a few ways - firstly, the result applies to the simpler Bernoulli sampler, and improves the deviation control to $O(\sqrt{W_t})$ instead of $O(\sqrt{T})$. This has the further advantage that if one is concerned with the number of samples queried along with the memory, the Bernoulli sampler only queries $\sim pT$ times with high probability, while the reservoir sampler queries about $pT\log T$ times. Secondly, it shows that the Bernoulli sampler *does* offer continuous robustness, but up to a flattening of the deviation control for sets of small incidence (small W_t). Ben-Eliezer & Yogev show a number of applications of such bounds to sketching, and Alon et al. have recently applied this to tightly characterise the regret in online classification [Alo+21], using techniques of Rakhlin et al. [RST15a; RST15b]. We believe that self-normalised bounds as above can contribute to showing adaptive versions of these results.

E.2 Analysis of VUE Against Adaptive Adversaries

This section serves to show Theorems 6.3.2 and 6.5.4. We will analyse the excess abstention, and the mistakes separately. Both deviations are controlled with probability $1-\delta/2$, and so a union bound completes the argument. The excess abstention control is common to both, and exploits Bernstein's inequality.

Proof of excess abstention bound. Notice that the procedure only abstains if $C_t = 1$ or if $\widehat{\mathcal{Y}}_t = \{\perp\}$. In the latter case, the competitor also abstains, and thus no excess abstention is incurred. Therefore, the net excess abstention is bounded as $A_T - A_T^* \le \sum C_t$. Now, $\sum C_t$ is a Binomial random variable with parameters T, p.

By Bernstein's inequality [BLM13, Ch. 2],

$$\mathbb{P}\left(\sum C_t \geq pT + 2\sqrt{p(1-p)T\log(2/\delta)} + 2\log(2/\delta)\right) \leq \frac{\delta}{2}. \qquad \square$$

We move on to bounding mistakes in a N-dependent way.

Proof of mistake bound from Theorem 6.3.2. As in the main text, consider the filtration $\{\mathscr{F}_t\} = \{\sigma(\mathscr{H}_t^{\mathfrak{A}})\}$, $U_t^f := \mathbb{1}\{f(X_t) \notin \{\perp, Y_t\}\}$, and consider the processes $W_t^f = \sum_{s<t} U_t^f$, $B_t = C_t$, $\widetilde{W}_t^f = U_t^f C_t$. Note that since $N \geq 2$, $\frac{\delta}{2N} \leq \frac{1}{4} \leq \frac{1}{\sqrt{e}}$.

Note that for every f, U_t^f and C_t satisfy the requirements of Lemma 6.3.1, since C_t is tossed independently of $\mathscr{H}_{t-1}^{\mathfrak{A}}$. Therefore, we may invoke Lemma 6.3.1 to find that

$$\mathbb{P}\left(\exists t : \widetilde{W}_t^f = 0, W_t^f \geq \frac{8}{p}\log({}^{2N}\!/\!{}_{\delta})\right) \leq \frac{\delta}{2N},$$

and applying a union bound over $f \in \mathcal{F}$, we conclude that

$$\mathbb{P}\left(\exists t, f : \widetilde{W}_t^f = 0, W_t^f \geq \frac{8}{p}\log({}^{2N}\!/\!{}_{\delta})\right) \leq \frac{\delta}{2},$$

Notice that if $\widetilde{W}_{t-1}^f$ is non-zero, then $f \notin \mathcal{V}_t$ since we've seen it make a mistake prior to the time t. Now define the stopping times $\tau_f := \max\{t : f \in \mathcal{V}_t\} = \max\{t : \widetilde{W}_{t-1}^f = 0\}$. We observe that

$$\begin{aligned}
M_T = \sum_t \mathbb{1}\{\widehat{Y}_t \notin \{\perp, Y_t\}\} &\leq \sum_t \mathbb{1}\{\exists f \in \mathcal{V}_t : f(X_t) \notin \{\perp, Y_t\}\} \\
&\leq \sum_f \sum_t \mathbb{1}\{f \in \mathcal{V}_t, f(X_t) \notin \{\perp, Y_t\}\} \\
&= \sum_f \sum_t \mathbb{1}\{t \leq \tau_f\} U_t^f.
\end{aligned}$$

Next, define the event

$$\mathsf{E} := \left\{\exists t, f : f \in \mathcal{V}_t, W_{t-1}^f \geq 8\log(2N/\delta)/p\right\}.$$

Since $f \in \mathcal{V}_t \iff \widetilde{W}_{t-1}^f = 0 \iff t \leq \tau_f$. Also recall that $W_{t-1}^f =$

$\sum_{s<t} \mathbb{1}\{f(X_s) \notin \{\perp, Y_s\}$. Therefore, given E^c,

$$\sum_t \mathbb{1}\{t \le \tau_f, f(X_t) \notin \{\perp, Y_t\}\} \le 8\frac{\log(2N/\delta)}{p} + 1,$$

since on E^c, $t \le \tau_f \implies \widetilde{W}^f_{t-1} = 0 \implies \sum_{s<t} U_t^f \le \frac{8\log(2N/\delta)}{p}$, and the additional 1 arises since E^c does not control behaviour at τ_f. We conclude that given E^c, we have

$$M_T \le \sum_f 9\frac{\log(2N/\delta)}{p} = 9\frac{N\log(2N/\delta)}{p}.$$

But E occurs with probability at most $\delta/2$, and we have shown that

$$\mathbb{P}\left(M_T > \frac{9N\log({}^{2N}/{}_{\delta})}{p}\right) \le \frac{\delta}{2}. \qquad \square$$

As discussed in §6.5, the $\mathcal{X}$-dependent argument proceeds similarly.

Proof of mistake bound from Theorem 6.5.4. Again, we will work with the the natural filtration $\{\mathscr{F}_t\} = \{\sigma(\mathscr{H}_t^{\mathfrak{A}})\}$. Define $\widehat{\mathcal{Y}}_t^x = \{f(x) : f \in \mathcal{V}_t\}$, and the process $U_t^x := \mathbb{1}\{X_t = x, \widehat{Y}_t \notin \{\perp, Y_t\}\}$, and consider the processes $W_t^x = \sum_{s<t} U_t^x, B_t = C_t, \widetilde{W}_t^x = U_t^x C_t$. Again, since $|\mathcal{X}| \ge 2$, $\frac{\delta}{2|\mathcal{X}|} \le \frac{1}{4} \le \frac{1}{\sqrt{e}}$.

Invoking Lemma 6.3.1, since C_t is tossed independently of $\mathscr{H}_{t-1}^{\mathfrak{A}}$, we find that

$$\mathbb{P}\left(\exists t : \widetilde{W}_t^x \le 1, W_t^x \ge \frac{8}{p}\log({}^{2|\mathcal{X}|}/{}_{\delta})\right) \le \frac{\delta}{2|\mathcal{X}|},$$

and applying a union bound over $x \in \mathcal{X}$, we conclude that

$$\mathbb{P}\left(\exists t, x : \widetilde{W}_t^x \le 1, W_t^x \ge \frac{8}{p}\log({}^{2|\mathcal{X}|}/{}_{\delta})\right) \le \frac{\delta}{2},$$

Now, from the argument in the main text, $U_t^x \ge 0 \implies |\widehat{\mathcal{Y}}_t^x| \ge 2 \iff W_{t-1}^x \le 1$. So, define the stopping times

$$\tau_x := \max\{t : |\widehat{\mathcal{Y}}_t^x| \ge 2\} = \max\{t : W_{t-1}^x \le 1\}.$$

We have that

$$\begin{aligned}M_T &= \sum_t \mathbb{1}\{\widehat{Y}_t \not\in \{\perp, Y_t\}\} \\ &= \sum_x \sum_t \mathbb{1}\{|\widehat{\mathcal{Y}}_t^x| \geq 2\} U_t^x \\ &= \sum_x \sum_t \mathbb{1}\{t \leq \tau_x\} U_t^x.\end{aligned}$$

Defining the event

$$\mathsf{E} := \left\{\exists t, x : t \leq \tau_x, W_{t-1}^x \geq 8 \log({}^{2|\mathcal{X}|}/_{p})\right\},$$

we again observe that given E^c,

$$\sum_t \mathbb{1}\{t \leq \tau_x\} U_t^x \leq 1 + 8\frac{\log(2N/\delta)}{p},$$

since on E^c, $t \leq \tau_x \iff \widetilde{W}_{t-1}^x \leq 1 \implies \sum_{s \leq t-1} U_s^x \leq \frac{8 \log(2|\mathcal{X}|/\delta}{p}$. We thus conclude that

$$M_T \leq \sum_x 9\frac{\log(2|\mathcal{X}|/\delta)}{p} = \frac{9|\mathcal{X}|\log(2|\mathcal{X}|/\delta)}{p}.$$

But E occurs with probability at most $\delta/2$, and we have shown that

$$\mathbb{P}\left(M_T > \frac{9|\mathcal{X}|\log({}^{2|\mathcal{X}|}/_{\delta})}{p}\right) \leq \frac{\delta}{2}. \qquad \square$$

E.3 Stochastic Adversaries

This section contains proofs omitted from §6.4.

E.3.1 Performance of VUE-PROD

This section consitutes a proof of Theorem 6.4.1. We begin by controlling the excess abstentions.

Proof of excess abstention bound. We begin by analysing the PROD algorithm for the setting where decision sets may shrink with time. For succinctness, denote $a_t^f = \mathbb{1}\{f(X_t) = \perp\}, A_t^f := \sum_{s \leq t} a_t^f$.

Lemma E.3.1. *Let π_t^f be as in Algorithm 3. If $\eta \le 1/2$, then for any $g \in \mathcal{V}_T$, it holds that*

$$\sum_{t,f} \pi_t^f a_t^f \le \frac{\log N}{\eta} + A_T^g + \eta \sum_{t \le T} (a_t^g)^2.$$

Proof. We follow the standard analysis of PROD, updated slightly to account for versioning. Consider the potential $W_t := \sum_{f \in \mathcal{V}_t} w_t^f$, where recall that $w_t^f = \prod_{s<t}(1 - \eta a_s^f)$. Since the weights are always non-negative, for any $g \in \mathcal{V}_T$, we have that

$$W_{T+1} \ge \prod_{t \le T} (1 - \eta a_t^g).$$

Therefore, we have the lower bound

$$\log \frac{W_{T+1}}{W_1} \ge -\log N + \sum \log(1 - \eta a_t^g) \ge -\log N - \sum \eta a_t^g - \sum (\eta a_t^g)^2,$$

which exploits the fact that for $z \le 1/2, \log(1-z) \ge -z - z^2$.

To upper bound the same quantity, notice that for any t,

$$W_{t+1} = \sum_{f \in \mathcal{V}_{t+1}} w_{t+1}^f \le \sum_{f \in \mathcal{V}_t} w_t^f (1 - \eta a_t^f) = W_t \left(1 - \eta \sum_f \pi_t^f a_t^f\right),$$

which again exploits that weights are non-negative, and that $\mathcal{V}_t$ is a non-increasing sequence of sets. Taking ratios and bounding $\log(1-z)$ by $-z$, and finally summing over $t = 1 : T$, we have

$$\log \frac{W_{T+1}}{W_1} = \sum_t \log \frac{W_{t+1}}{W_t} \le -\eta \sum_t \sum_f \pi_t^f a_t^f.$$

Rearranging the inequality obtained by sandwiching $\log \frac{W_{T+1}}{W_1}$ yields the bound. □

Note that the above lemma holds generically, for any loss $\ell_t^f \le 1$, and any sequence of shrinking decision sets. We will exploit this fact later.

For our purposes, observe that since a_t^f is an indicator, $(a_t^f)^2 = a_t^f$. Thus, using Lemma E.3.1 for $g = f^* \in \mathcal{V}_T$,

$$\sum_{t,f} \pi_t^f a_t^f \le \frac{\log N}{\eta} + A_T^* + \eta A_T^*.$$

Now, the total abstention incurred by the learner is

$$A_T = \sum \mathbb{1}\{C_t = 1\} + \mathbb{1}\{C_t = 0, f_t(X_t) = \perp\}.$$

Exploiting the independence of the exploratory coin, we find that

$$\mathbb{E}[A_T] = pT + (1-p)\mathbb{E}[\sum_{t,f} \pi_t^f a_t^f].$$

Invoking the above bound on $\sum_{t,f} \pi_t^f a_t^f$ and rearranging then yields that

$$\mathbb{E}[A_T] \le pT + \frac{(1-p)\log N}{\eta} + (1-p)\mathbb{E}[A_T^*] + \eta(1-p)\mathbb{E}[A_T^*].$$

Now, if $\eta = p$, then $\eta(1-p) - p = -p^2 < 0$, and then exploiting that $A_T^* \ge 0$ yields the bound

$$\mathbb{E}[A_T - A_T^*] \le pT + \frac{\log N}{p}. \qquad \square$$

This leaves the mistake control. The argument we present critically relies on the law π_t^f being chosen independently of X_t, given $\mathscr{H}_{t-1}^{\mathfrak{L}}$. This is ultimately a source of inefficiency - for instance, if π_t^f were allowed to depend also on X_t, then we could enforce that non-abstaining actions are not played when $C_t = 0$, and drop the second $\log(N)/p$ term from the excess abstention bound. However, we were unable to show mistake control with only logarithmic dependence on N in this situation.

Proof of mistake bound. The mistake control proceeds by partitioning the class $\mathcal{F}$ according to the mistake rates of individual $\mathcal{F}s$ and arguing that whole groups of these are simultaneously, and quickly, eliminated from the version space without incurring too many mistakes. This fundamentally exploits the stochasticity of the setting.

To this end, define

$$\mathcal{F}_\zeta := \{f \in \mathcal{F} : 2^{-\zeta} \le P(f(X_t) \notin \{?, Y_t\}) \le 2^{1-\zeta}$$
$$\overline{\mathcal{F}}_\zeta := \{f \in \mathcal{F} : P(f(X_t) \notin \{?, Y_t\}) \le 2^{-\zeta}\}.$$

In the following, ζ_0 is a parameter for the purposes of analysis, that will be chosen later. Notice that $\mathcal{F} = \bigcup_{\zeta \le \zeta_0} \mathcal{F}_\zeta \cup \overline{\mathcal{F}}_{\zeta_0}$.

We'll argue that all $f \in \mathcal{F}_\zeta$ are eliminated quickly (for small ζ). For this, it is useful to define the stopping times

$$\tau_\zeta := \max\{t : \exists f \in \mathcal{F}_\zeta \cap \mathcal{V}_t\}.$$

Notice that for any $f \in \mathcal{F}_\zeta$,

$$P(C_t = 1, f(X_t) \not\in \{\perp, Y_t\}) \geq 2^{-\zeta} p.$$

As a consequence of this and the union bound, we have the following tail inequality.

Lemma E.3.2. *For any* $\delta \in (0,1)$*,*

$$\mathbb{P}\left(\exists \zeta \leq \zeta_0 : \tau_\zeta > \sigma_{\delta,\zeta_0}(\zeta)\right) \leq \delta,$$

where

$$\sigma_{\delta,\zeta_0}(\zeta) := \frac{2^\zeta}{p} \log(\zeta_0 N/\delta).$$

With this in hand, notice that

$$\begin{aligned} M_T &= \sum_t \sum_f \mathbb{1}\{f_t = f\}\mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\} \\ &= \sum_t \sum_{\zeta \leq \zeta_0} \sum_{f \in \mathcal{F}_\zeta} \mathbb{1}\{f_t = f\}\mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\}\} \\ &\qquad + \sum_t \sum_{f \in \overline{\mathcal{F}}_{\zeta_0}} \mathbb{1}\{f_t = f\}\mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\}\}. \end{aligned}$$

Next, we observe that

$$\begin{aligned} \mathbb{E}\left[\sum_{f \in \mathcal{F}_\zeta} \mathbb{1}\{f_t = f\}\mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\}\middle| \mathscr{H}^{\mathfrak{L}}_{t-1}\right] &= \sum_{f \in \mathcal{F}_\zeta} \pi_t^f P(f(X_t) \not\in \{\perp, Y_t\}) \\ &\leq 2^{1-\zeta} \pi_t(f_t \in \mathcal{F}_\zeta) \\ &\leq 2^{1-\zeta} \mathbb{1}\{t \leq \tau_\zeta\}, \end{aligned}$$

where the first equality is because π_t^f is predictable given $\mathscr{H}^{\mathfrak{L}}_{t-1}$, the second uses the definition of $\mathcal{F}_\zeta$, and the final inequality is because π_t is a distribution that is supported on $\mathcal{V}_t$, and thus has total mass at most 1, and mass 0 when $\mathcal{F}_\zeta \cap \mathcal{V}_t = \varnothing$. In much the same way, also notice that

$$\mathbb{E}\left[\sum_{f \in \overline{\mathcal{F}}_{\zeta_0}} \mathbb{1}\{f_t = f, f(X_t) \notin \{\perp, Y_t\}\} \middle| \mathscr{H}^{\mathfrak{L}}_{t-1}\right] \le 2^{-\zeta_0}.$$

Exploiting both the linearity of expectations and the tower rule,

$$\begin{aligned}\mathbb{E}[M_T] &\le \sum_t \sum_{\zeta \le \zeta_0} 2^{1-\zeta} P(\tau_\zeta \ge t) + 2^{-\zeta_0} T \\ &\le \sum_{\zeta \le \zeta_0} \left(2^{1-\zeta} \sum_{t \le \sigma_{\delta,\zeta_0}(\zeta)} 1 + \sum_{t > \sigma_{\delta,\zeta_0}(\zeta)} \delta \right) + 2^{-\zeta_0} T \\ &\le 2\zeta_0 \frac{\log(\zeta_0 N/\delta)}{p} + 2\delta T + 2^{-\zeta_0} T.\end{aligned}$$

Now set $\zeta_0 = \lfloor \log T \rfloor, \delta = 1/T$. Since $\zeta_0 N/\delta \le N^2 T^2$, we find that

$$\mathbb{E}[M_T] \le 4 \frac{\log T \log(NT)}{p} + 4,$$

and finally since $p \le 1$, $\frac{4}{p} \ge 4$, leading to the claimed bound (for $T \ge 3$). □

E.3.2 Lower Bound

Proof of Theorem 6.4.2. Without loss of generality, assume $f_2(x) = 1$. Recall that $f_1(x) = \perp$. We describe the two adversaries -

- P_1^γ is supported on $\{(x, 1)\}$, so that for each time $X_t = x$, and the label $Y_t = 1$.
- P_2^γ is supported on $\{(x, 1), (x, 2)\}$ such that for each time $X_t = x$, while the label is drawn iid from the law $Y_t = \begin{cases} 1 & \text{w.p. } 1-\gamma \\ 2 & \text{w.p. } \gamma \end{cases}$.

Notice that against P_1^γ, the competitor is f_2, which attains $A_T^{(P_1^\gamma)} = 0$, while against P_2^γ, the competitor is f_1, which attains $A_T^{(P_2^\gamma)} = T$. Observe further that since $\gamma < 1/2$, if any learner does not play $\perp$, it is advantageous for it to play 1 and never play 2.[2] We thus lose no generality in assuming that the learner's actions lie in $\{\perp, 1\}$. Now,

[2] More formally, given any leaner, we can create the better—in expectation—learner that abstains when the given one does, and predicts 1 when the given one plays something other than $\perp$.

run two coupled versions of the learner, so that if these observe the same Z_ts, they produce identical actions. Feed the first of these data generated from P_1^γ, and the second of these data generated from P_2^γ.

Let η_1 be the (random) number of abstentions that the first version of the learner makes - this means that it must have played 1 $T - \eta_1$ times. Denote the number of mistakes that the second version of the learner makes as η_2. Given η_1, the second version gets exactly the same sequence as the first with probability $(1-\gamma)_1^\eta$ - indeed, due to the coupling, they first abstain together, and then receive the same label with probability $1-\gamma$. Conditioned on this, they again abstain together, and then receive the same label with probability $1-\gamma$ and so on, η_1 times. This means that, given η_1, and the event that they get the same sequence, the second version of the learner plays $T - \eta_1$ '1' actions. Since each of these is wrong with probability γ, independently and identically,

$$\mathbb{E}[\eta_2|\eta_1] \geq (1-\gamma)^{\eta_1}\gamma(T-\eta_1).$$

Notice that $(1-\gamma)^{\eta_1}$ is a convex function of η_1. Thus, $\mathbb{E}[(1-\gamma)_1^\eta] \geq (1-\gamma)^{\mathbb{E}[\eta_1]} = (1-\gamma)^K$. Further, $\mathbb{E}[-(1-\gamma)^{\eta_1}\eta_1] \geq \mathbb{E}[-\eta_1] = -K$, and finally, for $\gamma \leq 1/2$, $(1-\gamma) \geq e^{-2\gamma}$. It follows that

$$\mathbb{E}[\eta_2] \geq (1-\gamma)^K\gamma T - \gamma K = \gamma(e^{-2\gamma K}T - K). \qquad \square$$

While here, let us also comment that the proof of Corollary 6.4.3 is mildly incomplete, since the argument requires that $\varphi \geq 2$. If instead $\varphi < 2$, then notice that setting $\gamma = 1/2$ in the above, and using that $\mathbb{E}[\eta_2] \geq \gamma((1-\gamma)^K T - K)$, we have $\psi \geq 2^{-\varphi}\frac{T}{2} - \frac{2}{2} \geq \frac{T}{8} - 1$, which grows linearly with T.

E.4 Analysis of MIXED-LOSS-PROD Against Adaptive Adversaries

This section provides a proof of Theorem 6.5.1, and describes an adaptive variant of the same scheme, based on a doubling trick, that serves to show Theorem 6.5.3.

Proof of Theorem 6.5.1. Recall that the scheme runs PROD with the loss

$$\ell_t^f := \mathbb{1}\{C_t = 1\}\mathbb{1}\{f(X_t) \notin \{\perp, Y_t\}\} + \lambda\mathbb{1}\{f(X_t) = \perp\}.$$

We first observe that repeating the proof of Lemma E.3.1 with a_t^f replaced by ℓ_t^f gives us that for any $g \in \mathcal{V}_T$,

$$\sum_{t,f} \pi_t^f \ell_t^f \le \frac{\log N}{\eta} + \sum_t \ell_t^g + \eta \sum (\ell_t^g)^2. \tag{E.1}$$

Note that this relation holds given the context and label processes. For $g = f^* \in \mathcal{V}_T$, we observe that $\ell_t^{f^*} = \lambda \mathbb{1}\{f^*(X_t) = \perp\}$, since by definition f^* makes no mistakes. Instantiating the above with f^*, and noting $\sum \mathbb{1}\{f^*(X_t) = \perp\} = A_T^*$, we conclude that

$$\sum_{t,f} \pi_t^f \ell_t^f \le \frac{\log N}{\eta} + \lambda A_T^* + \eta \lambda^2 A_T^*. \tag{E.2}$$

We proceed to characterise the mistakes and abstentions that the learner makes in terms of $\sum_{t,f} \ell_t^f$. To this end, notice that

$$M_T = \sum_{t,f} \mathbb{1}\{f_t = f\} \cdot \mathbb{1}\{C_t = 0\} \cdot \mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\}\}.$$

As a result, integrating over the randomness of the algorithm, but not over the contexts or labels, we find that

$$\begin{aligned}\mathbb{E}[M_T] &= \mathbb{E}\left[\sum_{t,f} \mathbb{E}[\mathbb{1}\{f_t = f\}\mathbb{1}\{C_t = 0\}\mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\}\}|\mathscr{H}_{t-1}^{\mathfrak{A}}, X_t, Y_t]\right] \\ &= \sum_{t,f} \mathbb{E}\left[\pi_t^f (1-p)\mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\}\right].\end{aligned}$$

But, observe that

$$\begin{aligned}\mathbb{E}[\pi_t^f \ell_t^f] &= \mathbb{E}\left[\mathbb{E}[\pi_t^f C_t \mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\} + \lambda \pi_t^f \mathbb{1}\{f(X_t) = Y_t\}|\mathscr{H}_{t-1}^{\mathfrak{A}}]\right] \\ &= \mathbb{E}[p\pi_t^f \mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\}\}] + \lambda \mathbb{E}[\pi_t^f \mathbb{1}\{f(X_t) = \perp\}].\end{aligned}$$

Therefore,

$$\mathbb{E}[M_T] = \sum_{t,f} \mathbb{E}\left[\frac{(1-p)}{p}\left(\pi_t^f \ell_t^f - \pi_t^f \lambda \mathbb{1}\{f(X_t) = \perp\}\right)\right]. \tag{E.3}$$

Further, notice that

$$A_T = \sum_t \mathbb{1}\{C_t = 1\} + \sum_{t,f} \mathbb{1}\{C_t = 0\}\mathbb{1}\{f_t = f\}\mathbb{1}\{f(X_t) = \perp\},$$

and thus,

$$\mathbb{E}[A_T] = \mathbb{E}\left[pT + (1-p)\sum_{t,f} \pi_t^f \mathbb{1}\{f(X_t) = \perp\}\right].$$

Moving the negative terms in (E.3) to the left hand side, and exploiting the above, we find that

$$\mathbb{E}[M_T] + \frac{\lambda}{p}\mathbb{E}[A_T - pT] = \frac{1-p}{p}\mathbb{E}\left[\sum_{t,f} \pi_t^f \ell_t^f\right],$$

where we note that both the terms $\mathbb{E}[M_T]$ and $\mathbb{E}[A_T - pT]$ are non-negative.

Exploiting the inequality E.2 and the above relation, we conclude that

$$\mathbb{E}[M_T] + \mathbb{E}\left[\frac{\lambda}{p}(A_T - pT)\right] \le \mathbb{E}\left[\frac{1-p}{p}\left(\frac{\log N}{\eta} + \lambda A_T^* + \eta\lambda^2 A_T^*\right)\right]. \tag{E.4}$$

The required bounds are now forthcoming. Dropping the M_T term in the left hand side of (E.4), and pushing the constants N, η, p, λ through the expectations,

$$\begin{aligned} \frac{\lambda}{p}\mathbb{E}[A_T - pT] &\le \frac{(1-p)\log N}{p\eta} + \frac{(1-p)\lambda}{p}\mathbb{E}[A_T^*] + \frac{\eta(1-p)\lambda^2}{p}\mathbb{E}[A_T^*] \\ \iff \mathbb{E}[A_T - pT] &\le \frac{(1-p)\log N}{\eta\lambda} + (1-p)\mathbb{E}[A_T^*] + \eta\lambda(1-p)\mathbb{E}[A_T^*] \\ \iff \mathbb{E}[A_T - A_T^*] &\le pT + \frac{\log N}{\eta\lambda} + (\eta\lambda - p)\mathbb{E}[A_T^*]. \end{aligned}$$

Taking $\eta = 1/2, \lambda \le p$, observe that the last term is negative (since $A_T^* \ge 0$). Thus, making these substitutions and dropping the final term gives the required excess abstention control.

In a similar way, dropping the $\mathbb{E}[A_T - pT]$ term in (E.4) gives

$$\mathbb{E}[M_T] \le \frac{\log N}{p\eta} + \frac{\lambda(1+\eta\lambda)}{p}\mathbb{E}[A_T^*].$$

The claim follows on setting $\eta = 1/2$, and observing that $\eta\lambda \le 1$. □

E.4.1 Adapting Rates for small A_T^*

E.4.1.1 Deriving the form of $\widetilde{\alpha}$

We first describe a derivation of the form of $\widetilde{\alpha}$. As noted, the relevant parametrisation is $p = T^{-u}, \lambda = T^{-(u+v)}$, for $u, v \geq 0$. This, with the bounds of the previous section gives the control

$$\begin{aligned}\mathbb{E}[M_T] &\leq 2T^u \log N + T^{\alpha^*-v} \\ \mathbb{E}[A_T - A_T^*] &\leq T^{1-u} + 2T^{u+v}\log N + T^{\alpha^*-u-v}.\end{aligned}$$

Notice that $\alpha^* - u - v \leq 1 - u - v \leq 1 - u$, since $\alpha^* \leq 1, v \geq 0$. Thus, we have the rate bounds

$$\begin{aligned}\mu &= \max(u, \alpha^* - v) \\ \alpha &= \max(1-u, u+v)\end{aligned}$$

Deriving the optimal α attainable for a fixed μ then amounts to the following convex program

$$\begin{aligned}&\min \max(1-u, u+v) \\ &\text{s.t. } 0 \leq u \leq \mu \\ &\qquad \max(0, \alpha^* - \mu) \leq v\end{aligned}$$

Notice that the objective is a non-decreasing function of v, so the optimal choice of the same is $(\alpha^* - \mu)_+$, the smallest value it may take. This leaves us with trying to minimise $\max(1-u, u + (\alpha^* - \mu)_+)$ for $0 \leq u \leq \mu$. The unconstrained minimum of this function occurs at $u_0 = \frac{1-(\alpha^*-\mu)_+}{2}$, which is feasible if $\mu \geq u_0$. If on the other hand $\mu < u_0$, then the max-affine function is in the decreasing branch $1 - u$, and the optimal choice of u is just μ. Thus, the optimum is achieved at

$$\begin{aligned} v &= (\alpha^* - \mu)_+ \\ u &= \begin{cases} \frac{1-(\alpha^*-\mu)_+}{2} & 1-(\alpha^*-\mu)_+ \leq 2\mu \\ \mu & 1-(\alpha^*-\mu)_+ > 2\mu \end{cases} = \frac{\min(1-(\alpha^*-\mu)_+, 2\mu)}{2}. \end{aligned}$$

Correspondingly, $\widetilde{\alpha}$ takes the form

$$\widetilde{\alpha}(\mu;\alpha^*) = \begin{cases} \frac{1+(\alpha^*-\mu)_+}{2} & 1-(\alpha^*-\mu)_+ \leq 2\mu \\ \max(1-\mu, \mu+(\alpha^*-\mu)_+) & 1-(\alpha^*-\mu)_+ > 2\mu \end{cases}.$$

But,

$$1-(\alpha^*-\mu)+ > 2\mu \iff 1-\mu \geq \mu + (\alpha^*-\mu)_+,$$

and therefore

$$\widetilde{\alpha}(\mu;\alpha^*) = \begin{cases} \frac{1+(\alpha^*-\mu)_+}{2} & 1-(\alpha^*-\mu)_+ \leq 2\mu \\ 1-\mu & 1-(\alpha^*-\mu)_+ > 2\mu \end{cases} = \max\left(1-\mu, \frac{1+(\alpha^*-\mu)_+}{2}\right).$$

E.4.1.2 Adaptive Scheme and Proofs

We start by recalling the definition of B_t^*

$$B_t^* = \min_{f \in \mathcal{V}_t} \sum_{s \leq t} \mathbb{1}\{f(X_t) = \perp\}.$$

We will also use the term

$$\beta_t^* := \frac{\log B_t^*}{\log T}.$$

For the remainder of this section, let $\kappa := \frac{\lambda}{p}$. Recall that the optimal behaviour is attained by setting $p = T^{-u}, \kappa = T^{-v}$, where

$$u = \frac{\min(1-(\alpha^*-\mu)_+, 2\mu)}{2}$$
$$v = (\alpha^*-\mu)_+.$$

Algorithm 6 essentially consitutes a doubling trick by setting p and κ in phases, which are indexed by non-negative integers, n. The scheme is parametrised by a scale parameter, θ.

- We begin in the zeroth phase, with $\kappa = 1, p = T^{-\min(1,2\mu)/2}$ This phase ends when β^* first exceeds μ, at which point the first phase begins.
- At the beginning of each phase, we re-initialise the scheme.
- For $n \geq 1$, the nth phase ends when (the reinitialised) β^* first exceeds $\mu + n\theta$.
- Each time the nth phase ends, we restart the scheme, with $\kappa = T^{-(n+1)\theta}$, $p = T^{-\min(1-(n+1)\theta, 2\mu)/2}$.

Since the scheme is restarted in each phase, we may analyse each phase separately. Note that if $A_T \leq T^{\alpha^*}$ almost surely, then the index of the largest phase is at most $n^* = \lfloor \frac{(\alpha^*-\mu)_+}{\theta} \rfloor$ phases, since $\beta_t^* \leq \alpha^*$ always. For convenience, we set T_n to be the length of the nth phase. Times t_n correspond to rounds within the nth phase, and $M_{T_n}^n, A_{T_n}^n$ are the number of mistakes and abstentions incurred by the learner in the nth phase, while , $A_{T_n}^{*,n}$ is the number of abstentions incurred by f^* in the nth phase.

Consider the behaviour in the nth phase. Let g_n be the function that minimises

$$\sum_{s_n \leq T_n} \mathbb{1}\{g(X_t) = \perp\} \text{ subject to } \sum_{s_n \leq T_n} C_t \mathbb{1}\{g(X_t) \notin \{\perp, Y_t\} = 0,$$

and set the value of this optimum to $B_{T_n}^{*,n}$ By exploiting inequality (E.1) instantiated with g_n, and setting $\eta = 1/2$, we may infer that

$$\sum_{t_n \le T_n} \pi^f_{t_n} \ell^f_{t_n} \le 2\log N + p_n \kappa_n B^{*,n}_{T_n} + \frac{p_n^2 \kappa_n^2}{2} B^{*,n}_{T_n}.$$

As a result, reiterating the previous analysis over the nth phase, the number of mistakes and abstentions incurred in this phase

$$\mathbb{E}[M^n_{T_n}] \le \frac{2\log N}{p_n} + 2\mathbb{E}[\kappa_n B^{*,n}_{T_n}]$$
$$\mathbb{E}[A^n_{T_n} - B^{*,n}_{T_n}] \le \mathbb{E}[p_n T_n + 2\frac{\log N}{\kappa_n p_n}]$$

Further, notice that in each phase, $B^{*,n}_{T_n} \le T^{\mu+(n+1)\theta}$, $\kappa_n = T^{-n\theta}$, and $p_n = T^{-\min(1-n\theta, 2\mu)/2}$. Substituting these into the above bounds, we have

$$\mathbb{E}[M^n_{T_n}] \le 2T^{\min(1-n\theta,2\mu)/2} \log N + 2T^{\mu+\theta} \le 4T^{\mu+\theta} \log N$$
$$\mathbb{E}[A^n_{T_n} - B^{*,n}_{T_n}] \le T^{-\min(1-n\theta,2\mu)/2} \mathbb{E}[T_n] + T^{n\theta+\min(1-n\theta,2\mu)/2} \log N$$

But then, summing over the phases,

$$\begin{aligned} \mathbb{E}[M_T] &= \sum_{0 \le n \le n^*} \mathbb{E}[M^n_{T_n}] \\ &\le 4T^\mu \log N \cdot (n^*+1) T^\theta \\ &\le 4T^\mu \log N \cdot \frac{T^\theta}{\theta}. \end{aligned}$$

Further,

$$\begin{aligned}
\mathbb{E}[A_T - A_T^*] &= \mathbb{E}[\sum_{n\leq n^*} A^n_{T_n} - A^{*,n}_{T_n}] \\
&\leq \mathbb{E}[\sum_{0\leq n\leq n^*} A^n_{T_n} - B^{*,n}_{T_n}] \\
&\leq \mathbb{E}[\sum_{0\leq n\leq n^*} T^{-\min(\mu, 1-n\theta/2)} T_n] + \log N \sum_{0\leq n\leq n^*} T^{n\theta + \min(1-n\theta/2, \mu)} \\
&\leq \left(\sum_{n=0}^{n^*} T^{1-\min(\mu,1-n\theta/2)} + \sum_{n=0} T^{n\theta+\min(1-n\theta/2,\mu)}\right) \log N.
\end{aligned}$$

To simplify the above, let $n_0 = \lfloor \frac{1-2\mu}{\theta} \rfloor$, so that $\min(\mu, \frac{1-n\theta}{2}) = \mu$ for $n \leq n_0$. Notice that n_0 may be bigger or smaller than n^*. We can then write the bound as

$$\begin{aligned}
\frac{\mathbb{E}[A_T - T_T^*]}{\log N} \leq &\sum_{n=0}^{\min(n^*,n_0)} T^{1-\mu} + \sum_{n=\min(n^*,n_0)+1}^{n^*} T^{\frac{1+n\theta}{2}} \\
&+ \sum_{n=0}^{\min(n^*,n_0)} T^{n\theta+\mu} + \sum_{n=\min(n^*,n_0)+1}^{n^*} T^{\frac{1+n\theta}{2}},
\end{aligned}$$

where we interpret $\sum_{n=i}^{j} = 0$ for $i > j$. This can further be simplified to

$$\begin{aligned}
\frac{\mathbb{E}[A_T - A_T^*]}{\log N} \leq &\min(n^*+1, n_0+1) T^{1-\mu} \\
&+ \frac{T^\mu}{T^\theta - 1} T^{(\min(n^*,n_0)+1)\theta)} + 2\mathbf{1}\{n_0 < n^*\} \frac{T^{\frac{1+(n^*+1)\theta}{2}}}{T^{\theta/2} - 1}.
\end{aligned}$$

If we further assume that θ is chosen so that $T^{\theta/2} \geq 2$, we can lower bound $T^{\theta/2} - 1 \geq T^{\theta/2}/2, T^\theta - 1 \geq T^\theta/2$ which gives the bound

$$\frac{\mathbb{E}[A_T - A_T^*]}{4\log N} \leq (\min(n_0, n_*) + 1) \left(T^{1-\mu} + T^{\mu+\min(n_0,n^*)\theta} + \mathbb{1}\{n_0 < n^*\} T^{(1+n^*\theta)/2}\right),$$

from which we can derive the rate control

$$\alpha \leq \zeta(\mu, n_0, n^*, \theta) = \max(1-\mu, \mu + \min(n_0, n^*)\theta, \mathbb{1}\{n_0 < n^*\}(1+n^*\theta)/2)$$

The exact statement of the theorem is now straightforward to prove

Proof of Theorem 6.5.3. We run the above procedure with $\theta = \frac{2\ln 2}{\log T}$. Notice that $T^{\theta/2} \geq 2$, and that $T^\theta/\theta \leq \frac{2}{\ln 2}\log T \leq T^\varepsilon$ for large enough T. Therefore, mistakes are controlled at $O(T^{\mu+\varepsilon})$.

Further, for the abstention control, again $\min(n^*, n_0) + 1 \leq n_0 + 1 \leq \frac{1}{\theta} = \frac{\log T}{2\ln 2}$. Recall the abstention rate bound ζ above. It suffices to argue that $\zeta \leq \widetilde{\alpha} + \theta$, since $T^\theta = 4 = O(1)$.

To this end, first notice that

$$n_0 < n^* \iff \lfloor \frac{1-2\mu}{\theta} \rfloor < \lfloor \frac{(\alpha^* - \mu)_+}{\theta} \rfloor \implies 1 - 2\mu < (\alpha^* - \mu)_+.$$

In this case,

$$\begin{aligned}\zeta &= \max\left(1-\mu, \mu + n_0\theta, \frac{1+n^*\theta}{2}\right)\\ &\leq \max\left(1-\mu, \mu + \frac{(1-2\mu)}{\theta}\cdot\theta, \frac{1 + \frac{(\alpha^*-\mu)_+}{\theta}\cdot\theta}{2}\right)\\ &= \max\left(1-\mu, \frac{1+(\alpha^*-\mu)_+}{2}\right)\\ &= \widetilde{\alpha}(\mu;\alpha^*).\end{aligned}$$

On the other hand, if $n_0 \geq n^*$ then we have that

$$\frac{(\alpha^*-\mu)_+}{\theta} - 1 \leq \frac{(1-2\mu)}{\theta} \iff \mu \leq \frac{1+\theta-(\alpha^*-\mu)_+}{2}.$$

As a result, in this case,

$$\begin{aligned}\zeta &\leq \max\left(1-\mu, \mu + n^*\theta\right)\\ &\leq \max\left(1-\mu, \mu + (\alpha^*-\mu)_+\right)\\ &\leq \max\left(1-\mu, \frac{1+(\alpha^*-\mu)_+ + \theta}{2}\right)\\ &\leq \widetilde{\alpha}(\mu;\alpha^*) + \theta/2\end{aligned}$$

□

Algorithm 6 ADAPTIVE-MIXED-LOSS-PROD

```
**Inputs**: $\mathcal{F}$, Time $T$, Mistake rate $\mu$, Scale $\theta$.
**Initialise**: $n \leftarrow 0; n_{\max} \leftarrow \lceil 1/\theta \rceil; \forall f \in \mathcal{F}, w_1^f \leftarrow 1; \forall n \leq n_{\max}, \tau_n \leftarrow T$.
**for** $t \in [1:T]$ **do**
    $u \leftarrow \min(1 - n\theta, 2\mu)/2, v \leftarrow n\theta$
    $p \leftarrow T^{-u}, \lambda \leftarrow T^{-(u+v)}$.
    Sample $f_t \sim \pi_t = {w_t^f}/{\sum w_t^f}$.
    Toss $C_t \sim \mathrm{Bern}(p)$.
    **if** $C_t = 1$ **then**
        $\widehat{Y}_t \leftarrow \perp$
    **else**
        $\widehat{Y}_t \leftarrow f_t(X_t)$
    $\forall f \in \mathcal{F}$, evaluate
        $$\ell_t^f = C_t \mathbb{1}\{f(X_t) \notin \{\perp, Y_t\}\} + \lambda \mathbb{1}\{f(X_t) = \perp\}$$
    $w_{t+1}^f \leftarrow w_t^f (1 - \eta \ell_t^f)$.
    Compute
        $$B^* = \min_{g \in \mathcal{F}} \sum_{\tau_n < s \leq t} \mathbb{1}\{g(X_s) = \perp\}$$
        $$\text{s.t.} \sum_{\tau_n < s \leq t} C_s \mathbb{1}\{g(X_s) \notin \{\perp, Y_t\} = 0.$$
    **if** $\log B^* \geq (\mu + n\theta) \log T$ **then**
        $n \leftarrow n + 1$
        $\tau_{n+1} \leftarrow t$
        $\forall f \in \mathcal{F}, w_{t+1}^f \leftarrow 1$.
```

E.5 Details of Experiments.

N.B. Code required to reproduce the experiments is provided at https://github.com/anilkagak2/Online-Selective-Classification.

E.5.1 Dataset Details

GAS [Ver+12] dataset is a 6-way classification task based on the 16 chemical sensors data. These sensors are used to discriminate 6 gases at various levels of concentrations. The data consists of these sensor readings for over a period of 36 months divided into

10 batches. There are $13,910$ data points in this dataset. We use the first 7 batches as training set and the remaining 3 batches as test set. This split results in train and test sets with 9546 and 4364 data points respectively. The gas task contains data from 16 sensors (each of which gives 8 numbers). The standard error attained by the class we use (see below) on this is $\approx 87\%$. For the selective classification task, we use only the data from the first 8 sensors (and thus only 64 out of 128 features). The standard error attainable for this is $\approx 67\%$. Importantly, for the GAS task, the selective classification setting we study only demands matching the performance of the best classifier with the full 16-sensor data, and thus supervision for the 8-sensor function is according to this best function. To be more concrete, denote the training data as $\{(X_i^1, X_i^2, Y_i)\}$, where X^1 and X^2 are the features from the first and second 8 sensors respectively, and Y is the label. We train a classifier g on this whole dataset. Then we produce the labelled dataset $\{(X_i^1, g(X_i^1, X_i^2))\}$, and train selective classifiers on this dataset. The online problem then takes the test dataset, and gives to the learner only the X^1 features from it. If the learner abstains, then the label $Y_t = g(X_t^1, X_t^2)$ is given to the learner.

CIFAR-10 [Kri09] dataset is a popular image recognition dataset that consists of 32×32 pixels RGB images of 10 classes. It contains $50,000$ training and $10,000$ test images. We use standard data augmentations (shifting, mirroring and mean-std gaussian normalisation) for preprocessing the datasets. The best standard error attainable for this task by the models we use (see below) is $\approx 90\%$. This experiment is more straightforward to describe- selective classifiers are trained on the whole dataset. For the online problem, the test image is supplied to the learner, and if it abstains, then the true label of that image is provided as feedback.

E.5.2 Training Experts

[GKS21] proposed a scheme to train classifiers with an in-built abstention option. This scheme provides a loss function, which takes a single hyper-parameter μ, and is trained as a minimax program using gradient ascent-descent. The scheme then uses the outputs of this training with a second hyper-parameter t to provide classification or abstention decisions. Therefore, the scheme utilises two hyper-parameters (μ, t) to control the classification accuracy and abstentions.

We trained selective classifiers using this scheme. As per their recommendation, we used 30 values of μ with 10 values equally spaced in $[0.01, 1]$ and remaining 20 values in the $[1, 16]$. For the threshold parameter t, we used 20 equally spaced values in $[0.2, 0.95)$. The minimax program was run with the learning rates $(10^{-4}, 10^{-6})$ for the descent and ascent respectively. Notice that the resulting set of classifiers have $20 \times 30 = 600$ functions.

Note that classification on CIFAR-10 is a relatively difficult task than GAS. Hence, we used a simpler 3-layer fully connected neural network architecture for the GAS dataset, and a Resnet32 architecture [Ide19; HZRS16] for the CIFAR-10 dataset.

E.5.3 Algorithm implementation, Hyper-parameters, Compute requirements

We implemented Algorithm 7 (which relaxes the versioning in 3) using Python constructs. It has three hyper-parameters: (a) T denoting the number of rounds, (b) the exploration rate p, and (c) ε controlling the mistake tolerance. For each run, the test data points were randomly permuted, and the first T of them were presented to the algorithm.

There are two main departures from the scheme in the main text. Firstly, rather than only using feedback gained when $C_t = 1$, the version space is refined whenever $\widehat{Y}_t = \perp$, allowing faster learning. Secondly, the versioning is relaxed as already described, to only exclude functions that make too many mistakes, as determined by ε.

An important implementation detail is that for very small ε, the version space may get empty before the run concludes. This is particularly relevant for small values of ε. As a simple fix, we modify the versioning rule so that if the version space were to become empty at the end of a round, it is not updated (and, indeed, the state of the scheme is retained, see below).

Since our experiments are CPU compute bounded, we used a machine with two Intel Xeon 2.60 GHz CPUs providing 40 cores. Both the regret-with-varying-time experiments took about 1 hour compute time, and the operating point experiments took nearly 5 hours each.

E.5.4 Regret Behaviour as Time-horizon in Varied.

We use the hyperparameter $\varepsilon = 0.01$. For the sake of efficiency, we use the adaptive scheme Algorithm 8 that adapts to the time horizon, that instead varies p with the number of rounds as $p_t = \min(0.1, \frac{1}{\sqrt{t}}), \eta_t = p_t$. This adaptation strategy is a standard way to handle varying horizons, and the observations obtained via this represent (and slightly overestimate) the regrets for when Algorithm 7 is run with $p = \eta = \frac{1}{\sqrt{T}}$. A major advantage is that this significantly increases the efficiency of the procedure, since instead of re-starting the experiment for each time horizon, we can now run for one single time horizon, and obtain representative values of regret at smaller horizons by recording the values at checkpoints corresponding to these. In the plots, we ran for $T = 4000$, and checkpointed every 250 rounds.

E.5.4.1 Excess Abstention Behaviour

As noted in the main text, the excess abstention regret for both datasets is negative. This remains consistent with the theory, and likely arises since these datasets are, of course, not the worst case distributions. The excess abstentions regret are plotted below.

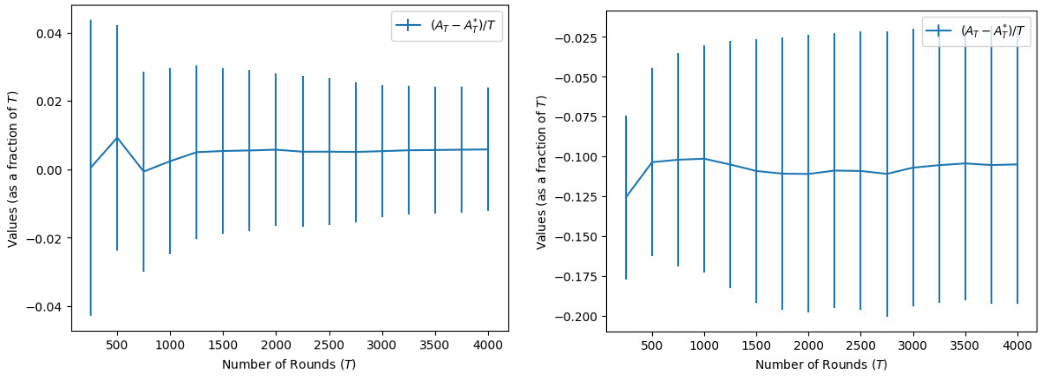


Figure E·1: Excess abstention regret, normalised by T, in the setting of Figure 6·2 for CIFAR-10 (left) and GAS (right). The plots are averaged over 100 runs, and one-standard-deviation error bars are drawn. Notice that the values are negative for GAS, and strongly dominated by the MMEA for CIFAR.

E.5.5 Achievable Operating Points of Mistakes and Abstentions

We use Algorithm 7, instantiated with $T = 500$, and always choosing $\eta = p$. The particular values of p, ε that are scanned are, as listed in the main text, 20 equally spaced values of p in the range $[0.015, 0.285]$, and 10 equally spaced values of ε in the range $[0.001, 0.046]$, giving in total 200 values of (p, ε) pairs that are scanned over.

The post-hoc batch operating points are obtained as follows: We first find the largest value of the number of mistakes that are attained by the online learner for some choice of (p, ε). Call this M. The values attained were $M_{\text{CIFAR}} = 50$ and $M_{\text{GAS}} = 120$. Then, we then instantiated the set $\mathcal{M}_{\text{CIFAR}} = \{2, 3, \dots, 50\}$, and for $\mathcal{M}_{\text{GAS}} = \{2, 7, \dots, 117\}$. The density was chosen lower for GAS for visual pleasantness. Finally, for each $m \in \mathcal{M}_*$, we run the post-hoc optimisation

$$a(m) := \min_{f \in \mathcal{F}} \sum_t \mathbb{1}\{f(X_t) = \perp\} \quad \text{s.t.} \quad \sum_t \mathbb{1}\{f(X_t) \notin \{\perp, Y_t\}\} \leq m.$$

The resulting points $(a(m), m)$ are plotted as black triangles.

Definition of MMEA As stated in the main text, the mistake matched competitor is defined as follows: suppose that the scheme makes M mistakes and A abstentions over a stream. If the following program is feasible, then we define

$$A^*(m) = \min_{f \in \mathcal{F}} \sum \mathbb{1}\{f(X_t) = \perp\} \text{ s.t. } \sum \mathbb{1}\{f(X_t) \not\in \{\perp, Y_t\}\} \le M.$$

If not, then we take $A^*(M)$ to be the abstentions made by the least mistake f, which is the competitor in the rest of the section. Then we define

$$\mathrm{MMEA} = A - A^*(M).$$

E.5.6 Sensitivity of the scheme to hyperparameters

Working in the setting of Figure 6·2, we show how the excess mistake and abstention regrets vary at $T = 4000$ (the final point) as ε is varied in Figure E·2. As expected, the excess mistakes increase roughly linearly with large ε, but the data reflects subtle non-monotonicities in the same. The variation in abstentions is, as expected, essentially opposite to that of the mistakes.

Similarly, in Figure E·3, we show the operating points that can be achieved by varying ε for a fixed p, and by varying p for a fixed ε. We observe first that the variation with ε for a fixed p is relatively regular, with larger ε increasing mistakes but decreasing abstentions at roughly the same rate, up to small variations. On the other, the behaviour with increasing p for a fixed ε is much more subtle, and indicates that a sweet-spot of the coin-based exploration rate exists for each tolerance level.

Together, these plots indicate that the optimal tuning of ε and p together can be subtle, and exploring how one can execute the same in an online way is an interesting open problem.

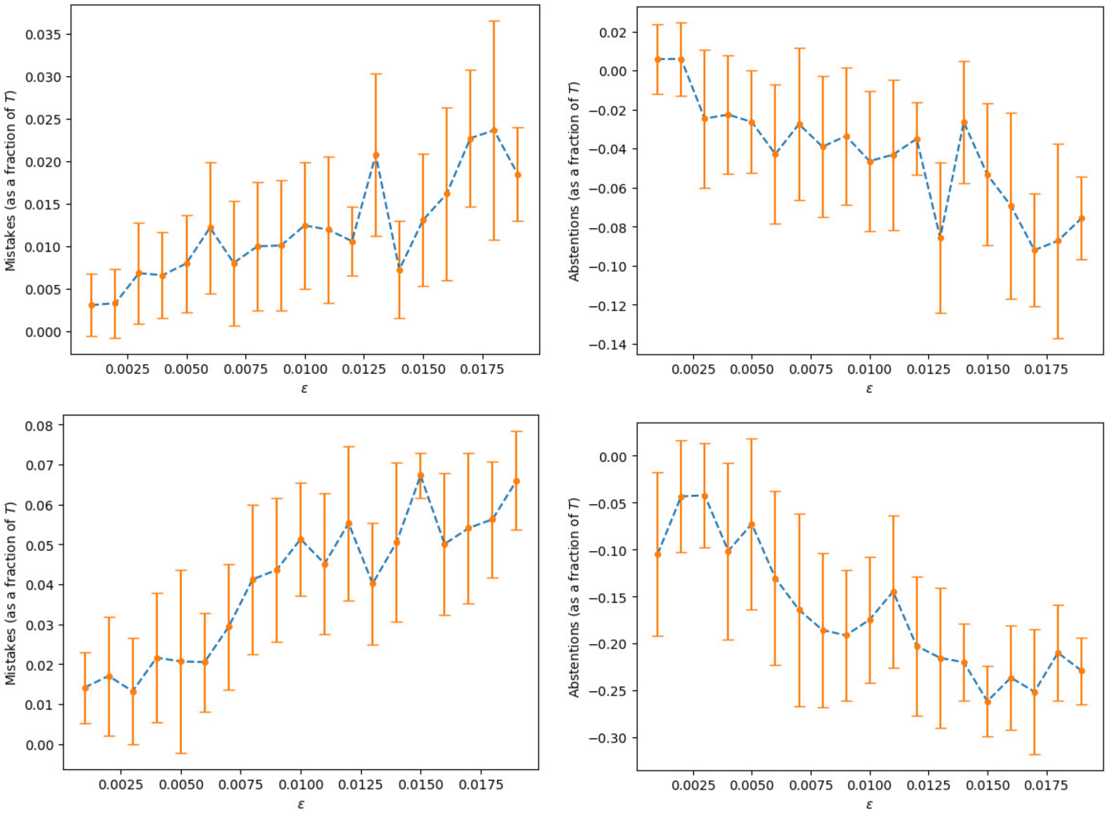


Figure E·2: Senstivity with ε of the excess mistakes (left) and excess abstention (right) regrets at $T = 4000$ for CIFAR (top) and GAS (bottom) datasets. Points are averaged over 100 runs, and one-standard-deviation error bars are included.

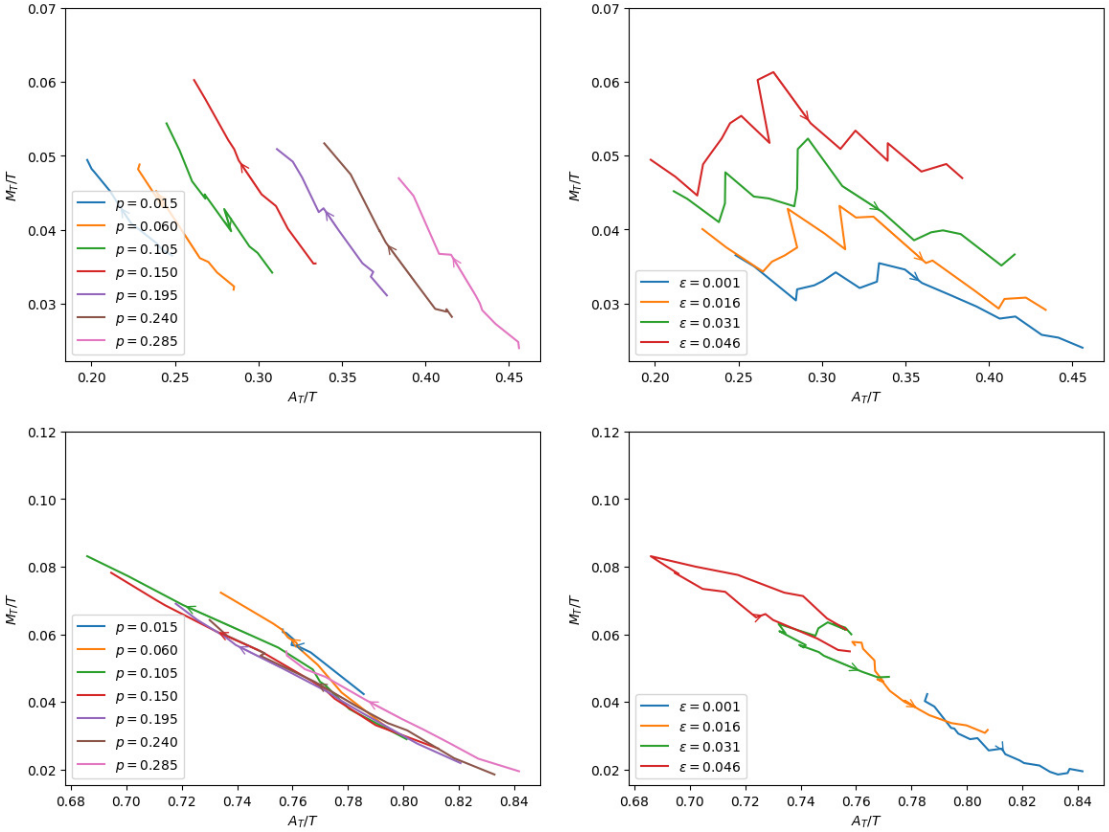


Figure E·3: Illustration of how operating points achieved by the scheme vary as p is changed for fixed values of ε (left) and as ε is changed for fixed values of p (right), in the CIFAR (top) and GAS (bottom) datasets. The sets of εs and ps marking the traces is reduced with respect to Figure 6·3 for the sake of legibility. The arrow denotes the direction of increasing the varied parameter.

Algorithm 7 VUE-PROD-RELAXED

```
Inputs: $\mathcal{F}$, Exploration rate $p$, Learning rate $\eta$, Tolerance $\varepsilon$.
Initialise: $\mathcal{V}_1 \leftarrow \mathcal{F}; \forall t, \mathcal{U}_t \leftarrow \varnothing; \forall f \in \mathcal{F}, w_1^f \leftarrow 1, o_0^f \leftarrow 0; \mathrm{Ctr}_0 \leftarrow 0.$
for $t \in [1:T]$ do
    Sample $f_t \sim \pi_t = \frac{w_t^f \mathbb{1}\{f \in \mathcal{V}_t\}}{\sum_{f \in \mathcal{V}_t} w_t^f}$.
    Toss $C_t \sim \mathrm{Bern}(p)$.
    $\widehat{Y}_t \leftarrow \begin{cases} \perp & C_t = 1 \\ f_t(X_t) & C_t = 0 \end{cases}$.
    if $\widehat{Y}_t = \perp$ then    ▷ Refine the version space if the exploratory coin is heads
        $\mathrm{Ctr}_t \leftarrow \mathrm{Ctr}_{t-1} + 1$.
        for $f \in \mathcal{V}_t$ do
            $o_t^f \leftarrow o_{t-1}^f + \mathbb{1}\{f(X_t) \notin \{\perp, Y_t\}\}$
            if $o_t^f \le \varepsilon \mathrm{Ctr}_t + \sqrt{2\varepsilon \mathrm{Ctr}_t}$ then ▷ Retain all $f$s that have error rate $< \varepsilon$ w.h.p.
                $\mathcal{U}_t \leftarrow \mathcal{U}_t \cup \{f\}$.
        $\mathcal{V}_{t+1} = \mathcal{V}_t \cap \mathcal{U}_t$.
    else
        $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_t$.
        $\forall f \in \mathcal{V}_{t+1}, o_t^f \leftarrow o_{t-1}^f$
        $\mathrm{Ctr}_t \leftarrow \mathrm{Ctr}_{t-1}$.
    if $\mathcal{V}_{t+1} \neq \varnothing$ then    ▷ Penalise Abstentions if the version space is non-empty
        for $f \in \mathcal{V}_{t+1}$ do
            $a_t^f \leftarrow \mathbb{1}\{f(X_t) = \perp\}$
            $w_{t+1}^f \leftarrow w_t^f \cdot (1 - \eta a_t^f)$.
    else                                ▷ $\mathcal{V}_{t+1} = \varnothing$, and so revert the state
        $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_t$.
        $\mathrm{Ctr}_t \leftarrow \mathrm{Ctr}_{t-1}$.
        for $f \in \mathcal{V}_{t+1}$ do
            $o_t^f \leftarrow o_{t-1}^f$.
            $w_{t+1}^f \leftarrow w_t^f$.
```

Algorithm 8 VUE-PROD-RELAXED-TIME-ADAPTED

```
**Inputs**: $\mathcal{F}$, Tolerance $\varepsilon$.
**Initialise**: $\mathcal{V}_1 \leftarrow \mathcal{F}; \forall t, \mathcal{U}_t \leftarrow \varnothing; \forall f \in \mathcal{F}, w_1^f \leftarrow 1, o_0^f \leftarrow 0; \mathrm{Ctr}_0 \leftarrow 0$.
**for** $t \in [1:T]$ **do**
    $p_t \leftarrow \min(0.1, 1/\sqrt{t})$.
    $\eta_t \leftarrow p_t$.
    Sample $f_t \sim \pi_t = \frac{w_t^f \mathbb{1}\{f \in \mathcal{V}_t\}}{\sum_{f \in \mathcal{V}_t} w_t^f}$.
    Toss $C_t \sim \mathrm{Bern}(p_t)$.
    $\widehat{Y}_t \leftarrow \begin{cases} \bot & C_t = 1 \\ f_t(X_t) & C_t = 0 \end{cases}$.
    **if** $\widehat{Y}_t = \bot$ **then**    ▷ Refine the version space if the exploratory coin is heads
        $\mathrm{Ctr}_t \leftarrow \mathrm{Ctr}_{t-1} + 1$.
        **for** $f \in \mathcal{V}_t$ **do**
            $o_t^f \leftarrow o_{t-1}^f + \mathbb{1}\{f(X_t) \notin \{\bot, Y_t\}\}$
            **if** $o_t^f \le \varepsilon \mathrm{Ctr}_t + \sqrt{2\varepsilon \mathrm{Ctr}_t}$ **then** ▷ Retain all $f$s that have error rate $< \varepsilon$ w.h.p.
                $\mathcal{U}_t \leftarrow \mathcal{U}_t \cup \{f\}$.
        $\mathcal{V}_{t+1} = \mathcal{V}_t \cap \mathcal{U}_t$.
    **else**
        $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_t$.
        $\forall f \in \mathcal{V}_{t+1}, o_t^f \leftarrow o_{t-1}^f$
        $\mathrm{Ctr}_t \leftarrow \mathrm{Ctr}_{t-1}$.
    **if** $\mathcal{V}_{t+1} \neq \varnothing$ **then**    ▷ Penalise Abstentions if the version space is non-empty
        **for** $f \in \mathcal{V}_{t+1}$ **do**
            $a_t^f \leftarrow \mathbb{1}\{f(X_t) = \bot\}$
            $w_{t+1}^f \leftarrow w_t^f \cdot (1 - \eta_t a_t^f)$.
    **else**    ▷ $\mathcal{V}_{t+1} = \varnothing$, and so revert the state
        $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_t$.
        $\mathrm{Ctr}_t \leftarrow \mathrm{Ctr}_{t-1}$.
        **for** $f \in \mathcal{V}_{t+1}$ **do**
            $o_t^f \leftarrow o_{t-1}^f$.
            $w_{t+1}^f \leftarrow w_t^f$.
```

Bibliography

[Abb18] Emmanuel Abbe. “Community Detection and Stochastic Block Models: Recent Developments”. In: *Journal of Machine Learning Research* 18.177 (2018), pp. 1–86 (cit. on pp. 5, 6, 20, 23, 38, 40).

[ABDK18] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. “Learning and Testing Causal Models with Interventions”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 9447–9460. URL: http://papers.nips.cc/paper/8155-learning-and-testing-causal-models-with-interventions.pdf (cit. on p. 46).

[ABFX08] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. “Mixed membership stochastic blockmodels”. In: *Journal of machine learning research* 9.Sep (2008), pp. 1981–2014 (cit. on p. 40).

[ABH16] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. “Exact recovery in the stochastic block model”. In: *IEEE Transactions on Information Theory* 62.1 (2016), pp. 471–487 (cit. on p. 5).

[ABS17] Emmanuel Abbe, Francois Baccelli, and Abishek Sankararaman. “Community detection on euclidean random graphs”. In: *arXiv preprint arXiv:1706.09942* (2017) (cit. on p. 40).

[ACDK15] Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. “Online learning with feedback graphs: Beyond bandits”. In: *Conference on Learning Theory*. PMLR. 2015, pp. 23–35 (cit. on p. 122).

[AD89] Rudolf Ahlswede and Gunter Dueck. “Identification via channels”. In: *IEEE Transactions on Information Theory* 35.1 (1989), pp. 15–29 (cit. on p. 42).

[AG05] Lada A Adamic and Natalie Glance. “The political blogosphere and the 2004 US election: divided they blog”. In: *Proceedings of the 3rd international workshop on Link discovery*. ACM. 2005, pp. 36–43 (cit. on pp. 20, 22, 33, 34, 37).

[AGS20] Durmus Alp Emre Acar, Aditya Gangrade, and Venkatesh Saligrama. “Budget Learning via Bracketing”. In: *International Conference on Artificial Intelligence and Statistics*. 2020, pp. 4109–4119 (cit. on pp. 76, 97).

[AKPS19] Radosław Adamczak, Michał Kotowski, Bartłomiej Polaczyk, and Michał Strzelecki. "A note on concentration for polynomials in the Ising model". In: *Electronic Journal of Probability* 24 (2019) (cit. on pp. 63, 211).

[Alo+21] Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. "Adversarial Laws of Large Numbers and Optimal Regret in Online Classification". In: *arXiv preprint arXiv:2101.09054* (2021) (cit. on pp. 17, 120, 127, 279).

[AM13] Emmanuel Abbe and Andrea Montanari. "Conditional random fields, planted constraint satisfaction and entropy concentration". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. Springer, 2013, pp. 332–346 (cit. on p. 41).

[AV14] Ery Arias-Castro and Nicolas Verzelen. "Community Detection in Dense Random Networks". In: *The Annals of Statistics* 42.3 (2014), pp. 940–969 (cit. on p. 23).

[Bas+08] Danielle S Bassett, Edward Bullmore, Beth A Verchinski, Venkata S Mattay, Daniel R Weinberger, and Andreas Meyer-Lindenberg. "Hierarchical organization of human cortical networks in health and schizophrenia". In: *Journal of Neuroscience* 28.37 (2008), pp. 9239–9248 (cit. on p. 20).

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. "Learnability and the Vapnik-Chervonenkis dimension". In: *Journal of the ACM (JACM)* 36.4 (1989), pp. 929–965 (cit. on p. 252).

[Bez+19] Ivona Bezáková, Antonio Blanca, Zongchen Chen, Daniel Štefankovič, and Eric Vigoda. "Lower bounds for testing graphical models: colorings and antiferromagnetic Ising models". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. 2019, pp. 283–298 (cit. on pp. 46, 218, 223).

[BK20] Guy Bresler and Mina Karzand. "Learning a tree-structured Ising model in order to make predictions". In: *The Annals of Statistics* 48.2 (2020), pp. 713–737 (cit. on pp. 48, 61, 197).

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013 (cit. on pp. 278, 280).

[BMNN16] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. "Information-theoretic thresholds for community detection in sparse networks". In: *Conference on Learning Theory*. 2016, pp. 383–416 (cit. on pp. 24, 169).

[BN18] Guy Bresler and Dheeraj Nagaraj. "Optimal Single Sample Tests for Structured versus Unstructured Network Data". In: *Conference On Learning Theory*. 2018, pp. 1657–1690 (cit. on p. 48).

[BN19] Guy Bresler and Dheeraj Nagaraj. “Stein’s method for stationary distributions of Markov chains and application to Ising models”. In: *Annals of Applied Probability* 29.5 (Oct. 2019), pp. 3230–3265. DOI: 10.1214/19-AAP1479. URL: https://doi.org/10.1214/19-AAP1479 (cit. on p. 48).

[Bre15] Guy Bresler. “Efficiently Learning Ising Models on Arbitrary Graphs”. In: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing (STOC 2015)*. Portland, Oregon, USA, 2015 (cit. on p. 5).

[BRS16] Quentin Berthet, Philippe Rigollet, and Piyush Srivastava. “Exact recovery in the Ising blockmodel”. In: *arXiv preprint arXiv:1612.03880* (2016) (cit. on pp. 68, 69).

[BRS19] Quentin Berthet, Philippe Rigollet, and Piyush Srivastava. “Exact recovery in the Ising blockmodel”. In: *Annals of Statistics* 47.4 (2019), pp. 1805–1834 (cit. on p. 67).

[BS16] Peter J Bickel and Purnamrita Sarkar. “Hypothesis testing for automated community detection in networks”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.1 (2016), pp. 253–273 (cit. on p. 24).

[BVB16] Eugene Belilovsky, Gaël Varoquaux, and Matthew B Blaschko. “Testing for differences in Gaussian graphical models: applications to brain connectivity”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 595–603 (cit. on p. 44).

[BW08] Peter L Bartlett and Marten Wegkamp. “Classification with a reject option using a hinge loss”. In: *Journal of Machine Learning Research* 9.Aug (2008), pp. 1823–1840 (cit. on pp. 74, 120).

[BWDS17] Tolga Bolukbasi, Joseph Wang, Ofer Dekel, and Venkatesh Saligrama. “Adaptive neural networks for efficient inference”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 527–536 (cit. on p. 101).

[BY20] Omri Ben-Eliezer and Eylon Yogev. “The adversarial robustness of sampling”. In: *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 2020, pp. 49–62 (cit. on pp. 17, 120, 127, 278, 279).

[BZN18] Kelly Bodwin, Kai Zhang, and Andrew Nobel. “A testing based approach to the discovery of differentially correlated variable sets”. In: *The Annals of Applied Statistics* 12.2 (2018), pp. 1180–1203 (cit. on p. 44).

[CCMW18] Jérémie Chalopin, Victor Chepoi, Shay Moran, and Manfred K Warmuth. “Unlabeled sample compression schemes and corner peelings for ample and maximum classes”. In: *arXiv preprint arXiv:1812.02099* (2018) (cit. on p. 113).

[CCZS21] Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. “Classification with rejection based on cost-sensitive classification”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 1507–1517 (cit. on p. 75).

[CDH19] Evgenii Chzhen, Christophe Denis, and Mohamed Hebiri. “Minimax semi-supervised confidence sets for multi-class classification”. In: *arXiv preprint arXiv:1904.12527* (2019) (cit. on p. 75).

[CDKS17] Clement L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. “Testing Bayesian Networks”. In: *Conference on Learning Theory*. 2017, pp. 370–448 (cit. on p. 46).

[CDM16] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. “Learning with rejection”. In: *International Conference on Algorithmic Learning Theory*. Springer. 2016, pp. 67–82 (cit. on pp. 74, 101, 114, 120, 267).

[CHHS02] Adam Cannon, James Howse, Don Hush, and Clint Scovel. “Learning with the Neyman-Pearson and min-max criteria”. In: *Los Alamos National Laboratory, Tech. Rep. LA-UR* (2002), pp. 02–2951 (cit. on pp. 102, 106).

[Cho57] C Chow. “An optimum character recognition system using decision functions”. In: *IRE Transactions on Electronic Computers* EC-6.4 (1957), pp. 247–254 (cit. on pp. 12, 72, 118).

[Cho70] C Chow. “On optimum recognition error and reject tradeoff”. In: *IEEE Transactions on Information Theory* 16.1 (1970), pp. 41–46 (cit. on pp. 12, 72, 118).

[Chu10] Fan Chung. “Graph theory in the information age”. In: *Notices of the AMS* 57.6 (2010), pp. 726–732 (cit. on p. 21).

[CL06] Fan Chung and Linyuan Lu. *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics)*. Boston, MA, USA: American Mathematical Society, 2006. ISBN: 0821836579 (cit. on pp. 143, 144, 160, 167).

[Cli+07] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, et al. “Integration of biological networks and gene expression data using Cytoscape”. In: *Nature protocols* 2.10 (2007), p. 2366 (cit. on p. 4).

[CLMX19] TT Cai, H Li, J Ma, and Y Xia. “Differential Markov random field analysis with an application to detecting differential microbial community networks”. In: *Biometrika* 106.2 (2019), pp. 401–416 (cit. on pp. 44, 48).

[CLS05] Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. “Minimizing regret with label efficient prediction”. In: *IEEE Transactions on Information Theory* 51.6 (2005), pp. 2152–2162 (cit. on pp. 122, 126).

[CMS07] Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. “Improved second-order bounds for prediction with expert advice”. In: *Machine Learning* 66.2 (2007), pp. 321–352 (cit. on pp. 131, 133).

[CNL18] Yuan Cao, Matey Neykov, and Han Liu. “High Temperature Structure Detection in Ferromagnets”. In: *arXiv preprint arXiv:1809.08204* (2018) (cit. on pp. 48, 53, 58, 214, 215, 243–245).

[Cor+18] Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. “Online learning with abstention”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1059–1067 (cit. on p. 120).

[Cor19] Apple Inc CoreML. *CoreML Documentation*. Note: Product documentation, not peer-reviewed. Accessed on 2020-2-28. 2019. URL: https://developer.apple.com/documentation/coreml (cit. on p. 99).

[Cos+10] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice LY Koh, Kiana Toufighi, and Sara Mostafavi. “The genetic landscape of a cell”. In: *science* 327.5964 (2010), pp. 425–431 (cit. on p. 4).

[CRV15] Peter Chin, Anup Rao, and Van Vu. “Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery”. In: *Conference on Learning Theory*. 2015, pp. 391–423 (cit. on pp. 5, 22, 23, 31, 69, 158, 160).

[CY06] Jingchun Chen and Bo Yuan. “Detecting functional modules in the yeast protein–protein interaction network”. In: *Bioinformatics* 22.18 (2006), pp. 2283–2290. DOI: 10.1093/bioinformatics/btl370 (cit. on p. 20).

[DAM17] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. “Asymptotic mutual information for the balanced binary stochastic block model”. In: *Information and Inference: A Journal of the IMA* 6.2 (2017), pp. 125–170. DOI: 10.1093/imaiai/iaw017 (cit. on p. 23).

[DDK16] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. “Testing Ising Models”. In: *arXiv preprint arXiv:1612.03147* (2016) (cit. on p. 223).

[DDK17] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. “Concentration of multilinear functions of the Ising model with applications to network data”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 12–23 (cit. on pp. 46, 55).

[DDK19] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. "Testing Ising models". In: *IEEE Transactions on Information Theory* 65.11 (2019), pp. 6829–6852 (cit. on pp. 46, 58, 66).

[Den+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 11).

[DH17] Christophe Denis and Mohamed Hebiri. "Confidence Sets with Expected Sizes for Multiclass Classification". In: *Journal of Machine Learning Research* 18.1 (Jan. 2017), pp. 3571–3598. ISSN: 1532-4435 (cit. on pp. 75, 82).

[DH19] Christophe Denis and Mohamed Hebiri. "Consistency of plug-in confidence sets for classification in semi-supervised learning". In: *Journal of Nonparametric Statistics* (2019), pp. 1–31 (cit. on p. 75).

[DKW18] Constantinos Daskalakis, Gautam Kamath, and John Wright. "Which distribution distances are sublinearly testable?" In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2018, pp. 2747–2764 (cit. on pp. 10, 63).

[DM10] Amir Dembo and Andrea Montanari. "Gibbs measures and phase transitions on sparse random graphs". In: *Brazilian Journal of Probability and Statistics* 24.2 (2010), pp. 137–211 (cit. on p. 36).

[DM17] Mathias Drton and Marloes H Maathuis. "Structure learning in graphical modeling". In: *Annual Review of Statistics and Its Application* 4 (2017), pp. 365–393 (cit. on p. 5).

[DZ13] Erik D Demaine and Morteza Zadimoghaddam. "Learning Disjunctions: Near-Optimal Trade-off between Mistakes and "I Don't Knows"". In: *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2013, pp. 1369–1379 (cit. on p. 122).

[EW10] Ran El-Yaniv and Yair Wiener. "On the foundations of noise-free selective classification". In: *Journal of Machine Learning Research* 11.May (2010), pp. 1605–1641 (cit. on pp. 74, 120).

[FB16] Farideh Fazayeli and Arindam Banerjee. "Generalized Direct Change Estimation in Ising Model Structure". In: *Proceedings of The 33rd International Conference on Machine Learning*. Vol. 48. 2016, pp. 2281–2290 (cit. on pp. 44, 46, 47).

[FC19] Yingjie Fei and Yudong Chen. "Exponential error rates of SDP for block models: Beyond Grothendieck's inequality". In: *IEEE Transactions on Information Theory* 65.1 (2019), pp. 551–571 (cit. on pp. 22, 23, 69, 160, 163).

[For10] Santo Fortunato. “Community detection in graphs”. In: *Physics reports* 486.3-5 (2010), pp. 75–174 (cit. on p. 20).

[Gal68] Robert G Gallager. *Information theory and reliable communication*. Vol. 588. Springer, 1968 (cit. on p. 171).

[GE17] Yonatan Geifman and Ran El-Yaniv. “Selective classification for deep neural networks”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 4878–4887 (cit. on pp. 74, 90, 101).

[GE19] Yonatan Geifman and Ran El-Yaniv. “SelectiveNet: A Deep Neural Network with an Integrated Reject Option”. In: *International Conference on Machine Learning*. 2019, pp. 2151–2159 (cit. on pp. 74, 79, 90, 94, 101, 114, 120, 267, 272).

[GG16] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *International Conference on Machine Learning*. 2016, pp. 1050–1059 (cit. on p. 76).

[GGCV20] Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike Von Luxburg. “Two-sample Hypothesis Testing for Inhomogeneous Random Graphs”. In: *The Annals of Statistics* 48.4 (2020), pp. 2208–2229 (cit. on pp. 24, 25).

[GGCvL17] Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike von Luxburg. “Two-Sample Tests for Large Random Graphs Using Network Statistics”. In: *Proceedings of the 2017 Conference on Learning Theory*. 2017, pp. 954–977 (cit. on p. 24).

[GKCS21] Aditya Gangrade, Anil Kag, Ashok Cutkosky, and Venkatesh Saligrama. “Online Selective Classification with Limited Feedback”. In: *Advances in Neural Information Processing Systems* 34 (2021) (cit. on p. 118).

[GKK19] Surbhi Goel, Daniel Kane, and Adam Klivans. “Learning Ising Models with Independent Failures”. In: *Conference on Learning Theory*. Vol. 99. 2019 (cit. on p. 48).

[GKS21] Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. “Selective Classification via One-Sided Prediction”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2179–2187 (cit. on pp. 72, 75, 120, 136, 298).

[GL17] Chao Gao and John Lafferty. “Testing network structure using relations between small subgraph probabilities”. In: *arXiv:1704.06742* (2017) (cit. on p. 24).

[GLP18] Reza Gheissari, Eyal Lubetzky, and Yuval Peres. “Concentration inequalities for polynomials of contracting Ising models”. In: *Electronic Communications in Probability* 23 (2018) (cit. on pp. 46, 55).

[GMPS19] Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. "Connectivity of Random Annulus Graphs and the Geometric Block Model". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*. Ed. by Dimitris Achlioptas and László A. Végh. Vol. 145. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, 53:1–53:23. ISBN: 978-3-95977-125-2. DOI: 10.4230/LIPIcs.APPROX-RANDOM.2019.53. URL: http://drops.dagstuhl.de/opus/volltexte/2019/11268 (cit. on p. 40).

[GMZZ18] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. "Community detection in degree-corrected block models". In: *The Annals of Statistics* 46.5 (2018), pp. 2153–2185 (cit. on p. 40).

[GN02] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826 (cit. on p. 20).

[GNS17] Aditya Gangrade, Bobak Nazer, and Venkatesh Saligrama. "Lower bounds for two-sample structural change detection in Ising and Gaussian models". In: *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2017, pp. 1016–1025 (cit. on pp. 43, 58).

[GNS18] Aditya Gangrade, Bobak Nazer, and Venkatesh Saligrama. "Two-Sample Testing can be as Hard as Structure Learning in Ising Models: Minimax Lower Bounds". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 6931–6935 (cit. on p. 43).

[GNS20] Aditya Gangrade, Bobak Nazer, and Venkatesh Saligrama. "Limits on Testing Structural Changes in Ising Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 9878–9889 (cit. on p. 43).

[Gol17] Oded Goldreich. *Introduction to property testing*. Cambridge University Press, 2017 (cit. on p. 10).

[GPMS18] S. Galhotra, S. Pal, A. Mazumdar, and B. Saha. "The Geometric Block Model and Applications". In: *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2018, pp. 1147–1150 (cit. on p. 40).

[Gri69] Robert B Griffiths. "Rigorous results for Ising ferromagnets of arbitrary spin". In: *Journal of Mathematical Physics* 10.9 (1969), pp. 1559–1565 (cit. on p. 207).

[Gun11] Adityanand Guntuboyina. "Lower bounds for the minimax risk using f-divergences, and applications". In: *IEEE Transactions on Information Theory* 57.4 (2011), pp. 2386–2399 (cit. on pp. 55, 188, 189).

[Gup+17] Chirag Gupta, Arun Sai Suggala, Ankit Goyal, Harsha Vardhan Simhadri, Bhargavi Paranjape, Ashish Kumar, Saurabh Goyal, Raghavendra Udupa, Manik Varma, and Prateek Jain. "ProtoNN: Compressed and Accurate kNN for Resource-scarce Devices". In: *International Conference on Machine Learning*. 2017, pp. 1331–1340 (cit. on p. 101).

[GvL18] Debarghya Ghoshdastidar and Ulrike von Luxburg. "Practical methods for graph two-sample testing". In: *Advances in Neural Information Processing Systems*. 2018, pp. 3019–3028 (cit. on pp. 24, 25).

[GVNS19] Aditya Gangrade, Praveen Venkatesh, Bobak Nazer, and Venkatesh Saligrama. "Efficient Near-Optimal Testing of Community Changes in Balanced Stochastic Block Models". In: *Advances in Neural Information Processing Systems*. 2019, pp. 10364–10375 (cit. on p. 20).

[Hau95] David Haussler. "Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension". In: *Journal of Combinatorial Theory, Series A* 69.2 (1995), pp. 217–232 (cit. on p. 264).

[HKM17] Linus Hamilton, Frederic Koehler, and Ankur Moitra. "Information theoretic properties of Markov random fields, and their algorithmic applications". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2463–2472 (cit. on p. 48).

[Hoe63] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30. DOI: `10.1080/01621459.1963.10500830` (cit. on pp. 154, 172).

[Hol17] Matthijs Hollemans. *Machine learning on mobile: on the device or in the cloud?* Note: Blog post, not peer-reviewed. Accessed on 2020-2-28. Feb. 2017. URL: `http://machinethink.net/blog/machine-learning-device-or-cloud/` (cit. on pp. 12, 99).

[HRMS18] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. "Time-uniform, nonparametric, nonasymptotic confidence sequences". In: *arXiv preprint arXiv:1810.08240* (2018) (cit. on pp. 127, 277).

[HRMS20] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. "Time-uniform Chernoff bounds via nonnegative supermartingales". In: *Probability Surveys* 17 (2020), pp. 257–317 (cit. on pp. 127, 275, 276).

[Hun07] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing In Science & Engineering* 9.3 (2007), pp. 90–95. DOI: `10.5281/zenodo.1098480` (cit. on p. 177).

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015) (cit. on pp. 11, 99, 101).

[HW06] Radu Herbei and Marten Wegkamp. “Classification with reject option”. In: *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* (2006), pp. 709–721 (cit. on pp. 74, 120).

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 89, 269, 298).

[Ide19] Yerlan Idelbayev. *Proper ResNet Implementation for CIFAR10/CIFAR100 in pytorch*. Accessed on 2020-2-28. 2019. URL: https://github.com/akamaster/pytorch_resnet_cifar10 (cit. on pp. 269, 298).

[IK12] Trey Ideker and Nevan J Krogan. “Differential network biology”. In: *Molecular systems biology* 8.1 (2012) (cit. on pp. 5, 6).

[IS12] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. Springer Science & Business Media, 2012 (cit. on pp. 10, 55, 187).

[JaJ85] Joseph JaJa. “Identification is easier than decoding”. In: *26th Annual Symposium on Foundations of Computer Science (FOCS 1985)*. IEEE. 1985, pp. 43–50 (cit. on p. 42).

[JOP01] Eric Jones, Travis Oliphant, and Pearu Peterson. *SciPy: Open source scientific tools for Python*. [Online]. 2001. URL: http://www.scipy.org/ (cit. on p. 177).

[JPL19a] Jaromír Janisch, Tomáš Pevný, and Viliam Lisý. “Classification with Costly Features as a Sequential Decision-Making Problem”. In: *arXiv preprint arXiv:1909.02564* (2019) (cit. on p. 101).

[JPL19b] Jaromír Janisch, Tomáš Pevný, and Viliam Lisý. “Classification with costly features using deep reinforcement learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 3959–3966 (cit. on p. 101).

[JTZ04] Daxin Jiang, Chun Tang, and Aidong Zhang. “Cluster analysis for gene expression data: A survey”. In: *IEEE Transactions on Knowledge & Data Engineering* 11 (2004), pp. 1370–1386 (cit. on p. 20).

[JY16] Antony Joseph and Bin Yu. “Impact of regularization on spectral clustering”. In: *The Annals of Statistics* 44.4 (2016), pp. 1765–1791. DOI: 10.1214/16-AOS1447 (cit. on p. 177).

[KGV17] Ashish Kumar, Saurabh Goyal, and Manik Varma. “Resource-efficient Machine Learning in 2 KB RAM for the Internet of Things”. In: *International Conference on Machine Learning*. 2017, pp. 1935–1944 (cit. on pp. 11, 99, 101).

[Kim18] Chiheon Kim. “Statistical limits of graphical channel models and a semidefinite programming approach”. PhD thesis. Massachusetts Institute of Technology, 2018 (cit. on p. 41).

[KKM12] Adam Tauman Kalai, Varun Kanade, and Yishay Mansour. “Reliable agnostic learning”. In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1481–1495 (cit. on pp. 75, 120).

[KLK19] Byol Kim, Song Liu, and Mladen Kolar. “Two-sample inference for high-dimensional markov networks”. In: *arXiv preprint arXiv:1905.00466* (2019) (cit. on pp. 44, 46).

[KLSS20] Holger Knöpfel, Matthias Löwe, Kristina Schubert, and Arthur Sinulis. “Fluctuation results for general block spin Ising models”. In: *Journal of Statistical Physics* (2020), pp. 1–26 (cit. on p. 68).

[KM17] Adam Klivans and Raghu Meka. “Learning graphical models using multiplicative weights”. In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 343–354 (cit. on pp. 5, 48).

[KN11] Brian Karrer and Mark EJ Newman. “Stochastic blockmodels and community structure in networks”. In: *Physical review E* 83.1 (2011), p. 016107 (cit. on p. 40).

[Kri09] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. 2009. URL: http://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf (cit. on pp. 89, 136, 297).

[Lei14] Jing Lei. “Classification with confidence”. In: *Biometrika* 101.4 (Oct. 2014), pp. 755–769. ISSN: 0006-3444. DOI: 10.1093/biomet/asu038 (cit. on pp. 75, 79, 120).

[Lei16] Jing Lei. “A goodness-of-fit test for stochastic block models”. In: *The Annals of Statistics* 44.1 (2016), pp. 401–424 (cit. on pp. 24, 35, 37).

[LFS17] Song Liu, Kenji Fukumizu, and Taiji Suzuki. “Learning sparse structural changes in high-dimensional Markov networks”. In: *Behaviormetrika* 44.1 (2017), pp. 265–286 (cit. on p. 46).

[Lin+20] Ji Lin, Wei-Ming Chen, Yujun Lin, Chuang Gan, Song Han, et al. “MCUNet: Tiny Deep Learning on IoT Devices”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 11711–11722 (cit. on p. 11).

[Liu+14] Song Liu, John A Quinn, Michael U Gutmann, Taiji Suzuki, and Masashi Sugiyama. “Direct learning of sparse changes in Markov networks by density ratio estimation”. In: *Neural computation* 26.6 (2014), pp. 1169–1197 (cit. on pp. 44, 46, 47).

[Liu+17] Song Liu, Taiji Suzuki, Raissa Relator, Jun Sese, Masashi Sugiyama, and Kenji Fukumizu. “Support consistency of direct sparse-change learning in Markov networks”. In: *The Annals of Statistics* 45.3 (2017), pp. 959–990. DOI: 10.1214/16-AOS1470. URL: http://dx.doi.org/10.1214/16-AOS1470 (cit. on pp. 44, 46, 47).

[Liu+19] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. “Deep Gamblers: Learning to Abstain with Portfolio Theory”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 10622–10632 (cit. on pp. 74, 90, 94).

[LJJ19] Tianyi Lin, Chi Jin, and Michael I Jordan. “On gradient descent ascent for nonconvex-concave minimax problems”. In: *arXiv preprint arXiv:1906.00331* (2019) (cit. on p. 92).

[LL18] Yezheng Li and Hongzhe Li. “Two-sample Test of Community Memberships of Weighted Stochastic Block Models”. In: *arXiv preprint arXiv:1811.12593* (2018) (cit. on pp. 24, 25).

[LLV17] Can M Le, Elizaveta Levina, and Roman Vershynin. “Concentration and regularization of random graphs”. In: *Random Structures & Algorithms* 51.3 (2017), pp. 538–561 (cit. on p. 25).

[LLWS11] Lihong Li, Michael L Littman, Thomas J Walsh, and Alexander L Strehl. “Knows what it knows: a framework for self-aware learning”. In: *Machine learning* 82.3 (2011), pp. 399–443 (cit. on p. 121).

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 6402–6413 (cit. on p. 76).

[LR06] Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer Science & Business Media, 2006 (cit. on p. 147).

[LS18] Matthias Löwe and Kristina Schubert. “Fluctuations for block spin Ising models”. In: *Electronic Communications in Probability* 23 (2018), 12 pp. DOI: 10.1214/18-ECP161. URL: https://doi.org/10.1214/18-ECP161 (cit. on p. 68).

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020 (cit. on p. 122).

[LVMC18] Andrey Y Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. "Optimal structure and parameter learning of Ising models". In: *Science advances* 4.3 (2018), e1700791 (cit. on p. 48).

[Mal12] Yu. V. Malykhin. "Bracketing entropy and VC-dimension". In: *Mathematical Notes* 91.5 (May 2012), pp. 800–807. ISSN: 1573-8876. DOI: 10.1134/S0001434612050264 (cit. on p. 111).

[ML 19] Google LLC ML Kit. *ML Kit Documentation.* Note: Product documentation, not peer-reviewed. Accessed on 2020-2-28. 2019. URL: https://developers.google.com/ml-kit (cit. on p. 99).

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. "Reconstruction and estimation in the planted partition model". In: *Probability Theory and Related Fields* 162.3 (2015), pp. 431–461. ISSN: 1432-2064. DOI: 10.1007/s00440-014-0576-6 (cit. on p. 5).

[Moh+16] Ali I Mohammed, Howard J Gritton, Hua-an Tseng, Mark E Bucklin, Zhaojie Yao, and Xue Han. "An integrative approach for analyzing hundreds of neurons in task performing mice using wide-field calcium imaging". In: *Scientific reports* 6 (2016), p. 20986 (cit. on p. 6).

[MRT18] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning.* Adaptive Computation and Machine Learning series. MIT Press, 2018. ISBN: 9780262039406 (cit. on pp. 248, 251, 258, 259).

[MS11] Shie Mannor and Ohad Shamir. "From Bandits to Experts: On the Value of Side-Observations". In: *Advances in Neural Information Processing Systems.* Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger. Vol. 24. Curran Associates, Inc., 2011 (cit. on p. 122).

[MT99] Enno Mammen and Alexandre Tsybakov. "Smooth discrimination analysis". In: *The Annals of Statistics* 27.6 (1999), pp. 1808–1829 (cit. on p. 110).

[NCHS19] Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. "On the Calibration of Multiclass Classification with Rejection". In: *Advances in Neural Information Processing Systems 32.* Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 2586–2596 (cit. on p. 74).

[Net+11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. "Reading Digits in Natural Images with Unsupervised Feature Learning". In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011.* 2011. URL: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf (cit. on p. 89).

[New06] Mark EJ Newman. “Modularity and community structure in networks”. In: *Proceedings of the national academy of sciences* 103.23 (2006), pp. 8577–8582 (cit. on p. 20).

[NL19] Matey Neykov and Han Liu. “Property testing in high-dimensional Ising models”. In: *The Annals of Statistics* 47.5 (2019), pp. 2472–2503 (cit. on pp. 48, 58, 223, 224).

[Nor19] Hellen Norman. *Living on the Edge: Why On-Device ML is Here to Stay*. Note: Popular article, not peer-reviewed. Accessed on 2020-2-28. Apr. 2019. URL: https://community.arm.com/developer/ip-products/processors/b/ml-ip-blog/posts/why-on-device-ml-is-here-to-stay (cit. on pp. 12, 99).

[NP15] Mark EJ Newman and Tiago P Peixoto. “Generalized communities in networks”. In: *Physical review letters* 115.8 (2015), p. 088701 (cit. on p. 40).

[NRWY12] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. “A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers”. In: *Statistical Science* 27.4 (2012), pp. 538–557 (cit. on p. 47).

[NS17a] Feng Nan and Venkatesh Saligrama. “Adaptive classification for prediction under a budget”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 4727–4737 (cit. on pp. 74, 99, 101, 114, 118, 267).

[NS17b] Feng Nan and Venkatesh Saligrama. “Dynamic model selection for prediction under a budget”. In: *arXiv preprint arXiv:1704.07505* (2017) (cit. on pp. 74, 101).

[NWS16] Feng Nan, Joseph Wang, and Venkatesh Saligrama. “Pruning random forests for prediction on a budget”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 2334–2342 (cit. on p. 101).

[NZ20] Gergely Neu and Nikita Zhivotovskiy. “Fast rates for online prediction with abstention”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 3030–3048 (cit. on p. 122).

[Oli06] Travis E Oliphant. *A guide to NumPy*. 2006 (cit. on p. 177).

[OR02] Evelien Otte and Ronald Rousseau. “Social network analysis: a powerful strategy, also for the information sciences”. In: *Journal of information Science* 28.6 (2002), pp. 441–453 (cit. on p. 4).

[OR94] Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994 (cit. on pp. 16, 120).

[Orl+15] Javier G. Orlandi, Bisakha Ray, Demian Battaglia, Isabelle Guyon, Vincent Lemaire, Mehreen Saeed, Alexander Statnikov, Olav Stetter, and Jordi Soriano. “First Connectomics Challenge: From Imaging to Connectivity”. In: *Proceedings of the Neural Connectomics Workshop at ECML 2014*. Ed. by Demian Battaglia, Isabelle Guyon, Vincent Lemaire, and Jordi Soriano. Vol. 46. Proceedings of Machine Learning Research. 2015, pp. 1–22 (cit. on p. 4).

[PDXL21] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. “Meta pseudo labels”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 11557–11568 (cit. on p. 11).

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 177).

[PF95] Eric M Phizicky and Stanley Fields. “Protein-protein interactions: methods for detection and analysis.” In: *Microbiology and Molecular Biology Reviews* 59.1 (1995), pp. 94–123 (cit. on p. 4).

[Pne+16] Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, and Weijian Yang. “Simultaneous denoising, deconvolution, and demixing of calcium imaging data”. In: *Neuron* 89.2 (2016), pp. 285–299 (cit. on p. 35).

[PTLC18] Yu-Shao Peng, Kai-Fu Tang, Hsuan-Tien Lin, and Edward Chang. “Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 7322–7331 (cit. on p. 101).

[RST15a] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. “Online learning via sequential complexities.” In: *Journal of Machine Learning Research* 16.1 (2015), pp. 155–186 (cit. on p. 279).

[RST15b] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. “Sequential complexities and uniform martingale laws of large numbers”. In: *Probability Theory and Related Fields* 161.1-2 (2015), pp. 111–153 (cit. on p. 279).

[RT11] Philippe Rigollet and Xin Tong. “Neyman-pearson classification, convexity and stochastic constraints”. In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2831–2855 (cit. on pp. 83, 252).

[RTA18] Harish G. Ramaswamy, Ambuj Tewari, and Shivani Agarwal. “Consistent algorithms for multiclass classification with an abstain option”. In: *Electronic Journal of Statistics* 12.1 (2018), pp. 530–554. DOI: 10.1214/17-EJS1388 (cit. on p. 74).

[SB17] A. Sankararaman and F. Baccelli. “Community detection on euclidean random graphs”. In: *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2017, pp. 510–517 (cit. on p. 40).

[SC16] Jonathan Scarlett and Volkan Cevher. “On the difficulty of selecting Ising models with approximate recovery”. In: *IEEE Transactions on Signal and Information Processing over Networks* 2.4 (2016), pp. 625–638 (cit. on pp. 48, 58).

[SGJ19] Shubhanshu Shekhar, Mohammad Ghavamzadeh, and Tara Javidi. “Binary Classification with Bounded Abstention Rate”. In: *arXiv preprint arXiv:1905.09561* (2019) (cit. on p. 83).

[Sho20] Ali Shojaie. “Differential network analysis: A statistical perspective”. In: *WIREs Computational Statistics* (2020). DOI: 10.1002/wics.1508. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1508. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1508 (cit. on p. 48).

[SLW19] Mauricio Sadinle, Jing Lei, and Larry Wasserman. “Least Ambiguous Set-Valued Classifiers With Bounded Error Levels”. In: *Journal of the American Statistical Association* 114.525 (2019), pp. 223–234. DOI: 10.1080/01621459.2017.1395341 (cit. on pp. 75, 82).

[SN05] Clayton Scott and Robert Nowak. “A Neyman-Pearson approach to statistical learning”. In: *IEEE Transactions on Information Theory* 51.11 (2005), pp. 3806–3819 (cit. on pp. 102, 106).

[SS11] István Szita and Csaba Szepesvári. “Agnostic KWIK learning and efficient approximate reinforcement learning”. In: *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings. 2011, pp. 739–772 (cit. on p. 121).

[SW12] Narayana P Santhanam and Martin J Wainwright. “Information-theoretic limits of selecting binary graphical models in high dimensions”. In: *IEEE Transactions on Information Theory* 58.7 (2012), pp. 4117–4134 (cit. on pp. 5, 48, 51, 58, 63, 186, 208–210, 222).

[SZB10] Amin Sayedi, Morteza Zadimoghaddam, and Avrim Blum. “Trading off Mistakes and Don’t-Know Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 23. 2010 (cit. on p. 122).

[Tak07] Gábor Takács. "The vapnik-chervonenkis dimension of convex n-gon classifiers". In: *Hungarian Electronic Journal of Sciences*. 2007 (cit. on p. 112).

[Tan+17] Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, Youngser Park, and Carey E. Priebe. "A Semiparametric Two-Sample Hypothesis Testing Problem for Random Graphs". In: *Journal of Computational and Graphical Statistics* 26.2 (2017), pp. 344–354. DOI: 10.1080/10618600.2016.1193505 (cit. on p. 24).

[Ton13] Xin Tong. "A Plug-in Approach to Neyman-Pearson Classification". In: *Journal of Machine Learning Research* 14.56 (2013), pp. 3011–3040. URL: http://jmlr.org/papers/v14/tong13a.html (cit. on p. 83).

[TS13] Kirill Trapeznikov and Venkatesh Saligrama. "Supervised Sequential Classification Under Budget Constraints". In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. Vol. 31. Proceedings of Machine Learning Research. PMLR, 2013, pp. 581–589 (cit. on p. 101).

[TSRD14] Rashish Tandon, Karthikeyan Shanmugam, Pradeep K Ravikumar, and Alexandros G Dimakis. "On the information theoretic limits of learning Ising models". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2303–2311 (cit. on p. 58).

[Tsy04] Alexandre Tsybakov. "Optimal aggregation of classifiers in statistical learning". In: *The Annals of Statistics* 32.1 (2004), pp. 135–166 (cit. on p. 110).

[VA15] Nicolas Verzelen and Ery Arias-Castro. "Community detection in sparse random networks". In: *The Annals of Applied Probability* 25.6 (2015), pp. 3465–3510 (cit. on p. 23).

[Vap00] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2000. ISBN: 9781441931603. DOI: 10.1007/978-1-4757-3264-1 (cit. on p. 76).

[vdVW96] Aad W. van der Vaart and Jon A Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996 (cit. on p. 102).

[Ver+12] Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margaret A Ryan, Margie L Homer, and Ramón Huerta. "Chemical gas sensor drift compensation using classifier ensembles". In: *Sensors and Actuators B: Chemical* 166 (2012), pp. 320–329 (cit. on pp. 136, 296).

[vHan13] Ramon van Handel. "The universal Glivenko–Cantelli property". In: *Probability Theory and Related Fields* 155.3 (2013), pp. 911–934. ISSN: 1432-2064. DOI: 10.1007/s00440-012-0416-5 (cit. on pp. 102, 111).

[vLux07] Ulrike von Luxburg. “A tutorial on spectral clustering”. In: *Statistics and Computing* 17.4 (2007), pp. 395–416. ISSN: 1573-1375. DOI: 10.1007/s11222-007-9033-z (cit. on p. 177).

[VMLC16] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. “Interaction screening: Efficient and sample-optimal learning of Ising models”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 2595–2603 (cit. on p. 5).

[WE11] Yair Wiener and Ran El-Yaniv. “Agnostic selective classification”. In: *Advances in Neural Information Processing Systems*. 2011, pp. 1665–1673 (cit. on pp. 74, 120).

[Weg07] Marten Wegkamp. “Lasso type classifiers with a reject option”. In: *Electronic Journal of Statistics* 1 (2007), pp. 155–168 (cit. on p. 74).

[WJ08] Martin J. Wainwright and Michael I. Jordan. “Graphical Models, Exponential Families, and Variational Inference”. In: *Foundations and Trends® in Machine Learning* 1.1–2 (2008), pp. 1–305. ISSN: 1935-8237. DOI: 10.1561/2200000001 (cit. on p. 36).

[WSD19] Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. “Sparse logistic regression learns all discrete pairwise graphical models”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 8069–8079 (cit. on p. 48).

[WTS15] Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. “Efficient Learning by Directed Acyclic Graph For Resource Constrained Prediction”. In: *Advances in Neural Information Processing Systems 28*. 2015 (cit. on p. 101).

[Wu+19] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. “Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10734–10742 (cit. on pp. 11, 99, 101).

[WWR10] Wei Wang, Martin J Wainwright, and Kannan Ramchandran. “Information theoretic bounds on model selection for Gaussian Markov random fields”. In: *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE. 2010, pp. 1373–1377 (cit. on p. 36).

[WX18] Yihong Wu and Jiaming Xu. “Statistical Problems with Planted Structures: Information-Theoretical and Computational Limits”. In: *arXiv preprint arXiv:1806.00118* (2018) (cit. on pp. 32, 169, 170).

[WY11] Marten Wegkamp and Ming Yuan. “Support vector machines with a reject option”. In: *Bernoulli* 17.4 (2011), pp. 1368–1385 (cit. on p. 74).

[XCC15] Yin Xia, Tianxi Cai, and T Tony Cai. “Testing differential networks with applications to the detection of gene-gene interactions”. In: *Biometrika* 102.2 (2015), pp. 247–266 (cit. on p. 44).

[Xu+14] Zhixiang (Eddie) Xu, Matt J. Kusner, Kilian Q. Weinberger, Minmin Chen, and Olivier Chapelle. “Classifier Cascades and Trees for Minimizing Feature Evaluation Cost”. In: *Journal of Machine Learning Research* 15 (2014), pp. 2113–2144. URL: http://jmlr.org/papers/v15/xu14a.html (cit. on pp. 72, 99, 101, 118).

[Ye21] Min Ye. “Exact recovery and sharp thresholds of Stochastic Ising Block Model”. In: *IEEE Transactions on Information Theory* (2021) (cit. on p. 70).

[Yu97] Bin Yu. “Assouad, Fano, and Le Cam”. In: *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*. Ed. by David Pollard, Erik Torgersen, and Grace L. Yang. New York, NY: Springer New York, 1997, pp. 423–435. ISBN: 978-1-4612-1880-7. DOI: 10.1007/978-1-4612-1880-7_29. URL: https://doi.org/10.1007/978-1-4612-1880-7_29 (cit. on pp. 55, 187).

[YW10] Ming Yuan and Marten Wegkamp. “Classification Methods with Reject Option Based on Convex Risk Minimization”. In: *Journal of Machine Learning Research* 11.5 (2010), pp. 111–130. URL: http://jmlr.org/papers/v11/yuan10a.html (cit. on p. 74).

[YY17] Weijian Yang and Rafael Yuste. “In vivo imaging of neural activity”. In: *Nature methods* 14.4 (2017), p. 349 (cit. on p. 4).

[ZC16] Chicheng Zhang and Kamalika Chaudhuri. “The extended littlestone’s dimension for learning with mistakes and abstentions”. In: *Conference on Learning Theory*. PMLR. 2016, pp. 1584–1616 (cit. on p. 122).

[ZCL14] Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. “Direct estimation of differential networks”. In: *Biometrika* 101.2 (2014), pp. 253–268 (cit. on p. 44).

[Zha+08] Bai Zhang, Huai Li, Rebecca B. Riggins, Ming Zhan, Jianhua Xuan, Zhen Zhang, Eric P. Hoffman, Robert Clarke, and Yue Wang. “Differential dependency network analysis to identify condition-specific topological changes in biological networks”. In: *Bioinformatics* 25.4 (2008), pp. 526–532. DOI: 10.1093/bioinformatics/btn660 (cit. on p. 20).

[Zha+19] Xiao-Fei Zhang, Le Ou-Yang, Shuo Yang, Xiaohua Hu, and Hong Yan. “DiffNetFDR: differential network analysis with false discovery rate control”. In: *Bioinformatics* (2019) (cit. on p. 44).

[Zhu+19] Pengkai Zhu, Durmus Alp Emre Acar, Nan Feng, Prateek Jain, and Venkatesh Saligrama. “Cost aware inference for iot devices”. In: *The 22nd International Conference on Artificial Intelligence and Statistics.* 2019, pp. 2770–2779 (cit. on pp. 12, 72, 99, 115).

[ZWTD19] Li Zhou, Hao Wen, Radu Teodorescu, and David HC Du. “Distributing deep neural networks with containerized partitions at the edge”. In: *2nd {USENIX} Workshop on Hot Topics in Edge Computing (HotEdge 19).* 2019 (cit. on pp. 11, 98).

[ZZ16] Anderson Y Zhang and Harrison H Zhou. “Minimax rates of community detection in stochastic block models”. In: *The Annals of Statistics* 44.5 (2016), pp. 2252–2280 (cit. on pp. 22, 23).

Curriculum Vitae

[redacted]

[redacted]