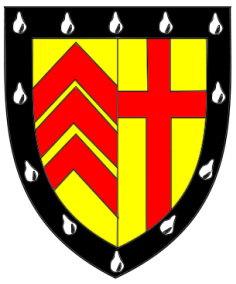
Structure in Machine Learning:

Graphical Models and Monte Carlo Methods



Mark Daniel Rowland

Department of Pure Mathematics and Mathematical Statistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare College

December 2018

Abstract

Structure in Machine Learning: Graphical Models and Monte Carlo Methods

Mark Daniel Rowland

This thesis is concerned with two main areas: approximate inference in discrete graphical models, and random embeddings for dimensionality reduction and approximate inference in kernel methods. Approximate inference is a fundamental problem in machine learning and statistics, with strong connections to other domains such as theoretical computer science. At the same time, there has often been a gap between the success of many algorithms in this area in practice, and what can be explained by theory; thus, an important research effort is to bridge this gap. Random embeddings for dimensionality reduction and approximate inference have led to great improvements in scalability of a wide variety of methods in machine learning. In recent years, there has been much work on how the stochasticity introduced by these approaches can be better controlled, and what further computational improvements can be made.

In the first part of this thesis, we study approximate inference algorithms for discrete graphical models. Firstly, we consider linear programming methods for approximate MAP inference, and develop our understanding of conditions for exactness of these approximations. Such guarantees of exactness are typically based on either structural restrictions on the underlying graph corresponding to the model (such as low treewidth), or restrictions on the types of potential functions that may be present in the model (such as log-supermodularity). We contribute two new classes of exactness guarantees: the first of these takes the form of particular hybrid restrictions on a combination of graph structure and potential types, whilst the second is given by excluding particular substructures from the underlying graph, via graph minor theory. We also study a particular family of transformation methods of graphical models, *uprooting* and *rerooting*, and their effect on approximate MAP and marginal inference methods. We prove new theoretical results on the behaviour of particular approximate inference methods under these transformations,

in particular showing that the triplet relaxation of the marginal polytope is unique in being universally rooted. We also introduce a heuristic which quickly picks a rerooting, and demonstrate benefits empirically on models over several graph topologies.

In the second part of this thesis, we study Monte Carlo methods for both linear dimensionality reduction and approximate inference in kernel machines. We prove the statistical benefit of coupling Monte Carlo samples to be almost-surely orthogonal in a variety of contexts, and study fast approximate methods of inducing this coupling. A surprising result is that these approximate methods can simultaneously offer improved statistical benefits, time complexity, and space complexity over i.i.d. Monte Carlo samples. We evaluate our methods on a variety of datasets, directly studying their effects on approximate kernel evaluation, as well as on downstream tasks such as Gaussian process regression.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Mark Daniel Rowland
December 2018

Acknowledgements

First and foremost, I'd like to thank my supervisors, Rich Turner and John Aston, for their support and guidance throughout the course of my PhD. I'd also like to thank Adrian Weller – we've collaborated closely on much of the work in this thesis, and it has been a great pleasure working with him. Adrian also introduced me to Krzysztof Choromanski when he visited Cambridge in 2016 – collaborating with Krzysztof has been a fantastic experience, and has led to the two papers that make up the second part of this thesis. I'd also like to thank my examiners, Robin Evans and Sergio Bacallado, for an enjoyable viva and many valuable comments on this thesis.

I'm very grateful to have been able to be part of the Statslab and MLG research communities whilst at Cambridge. It's been particularly enjoyable to have been part of such a collaboration-driven research environment at MLG, and to have worked with so many great people there, including Thang Bui, Zoubin Ghahramani, José Miguel Hernández-Lobato, Jiri Hron, Yingzhen Li, María Lomelí, Alex Matthews, and Nilesh Tripuraneni. Thanks also to Wessel Bruinsma, Jiri Hron, and Dave Janz for valuable feedback on this thesis.

I'd also like to thank the directors at the Cambridge Centre for Analysis, past and present. Particular thanks to Nigel Peake and James Norris, who admitted me to the CCA. Thanks also to David Waymont and Sumeet Singh, who supervised my first-year CCA research projects on neural network optimisation and sequential Monte Carlo methods, which were my first forays into machine learning and computational statistics.

I've also been fortunate to have been able to experience research in several industry environments during the course of my PhD, and would like to thank David Stern, Simon Setzer and Andrew Naish at G-Research, and Yee Whye Teh, Marc Bellemare, Will Dabney, Rémi Munos and many others at DeepMind for two very enjoyable internships. Both have had significant impacts on my approach to research in general.

Finally, thank you to my family and to Kristin, for all their love and support – without them, this thesis would not have been possible.

Publications

This thesis is based on the following publications:

- Weller, A., Rowland, M., and Sontag, D. (2016). Tightness of LP relaxations for almost balanced models. In *Artificial Intelligence and Statistics (AISTATS)*.
- Rowland, M., Pacchiano, A., and Weller, A. (2017). Conditions beyond treewidth for tightness of higher-order LP relaxations. In *Artificial Intelligence and Statistics (AISTATS)*.
- Choromanski, K.*, Rowland, M.*, and Weller, A. (2017). The unreasonable effectiveness of structured random orthogonal embeddings. In *Neural Information Processing Systems (NIPS)*. [*=equal contribution].
- Rowland, M.* and Weller, A.* (2017). Uprooting and rerooting higher-order graphical models. In *Neural Information Processing Systems (NIPS)*. [*=equal contribution].
- Choromanski, K.*, Rowland, M.*, Sarlos, T., Sindhwani, V., Turner, R. E., and Weller, A. (2018a). The geometry of random features. In *Artificial Intelligence and Statistics (AISTATS)*. [*=equal contribution].

During my time as a PhD student, I have also contributed to the following publications:

- Hernández-Lobato, J. M., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., and Turner, R. E. (2016). Black-box alpha divergence minimization. In *International Conference on Machine Learning (ICML)*.
- Tripuraneni, N., Rowland, M., Ghahramani, Z., and Turner, R. E. (2017). Magnetic Hamiltonian Monte Carlo. In *International Conference on Machine Learning (ICML)*.
- Dabney, W., Rowland, M., Bellemare, M. G., and Munos, R. (2017). Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence (AAAI)*.

- Rowland, M., Bellemare, M. G., Dabney, W., Munos, R., and Teh, Y. W. (2018). An analysis of categorical distributional reinforcement learning. In *Artificial Intelligence and Statistics (AISTATS)*.
- Matthews, A. G. D. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations (ICLR)*.
- Choromanski, K.*, Rowland, M.*, Sindhwani, V., Turner, R. E., and Weller, A. (2018b). Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning (ICML)*. [*=equal contribution].

Contents

Part I

Inference in Discrete Graphical Models

Chapter 1

Graphical Models and Approximate Inference

1.1 Graphical models

Graphical models provide a flexible language for describing probability distributions over collections of random variables, and are therefore popular throughout machine learning and statistics. This framework allows for rich modelling of probabilistic phenomena, representation of uncertainty, and in many cases also informs the design of efficient inference algorithms for these models. Graphical models are widely used in practice, and continue to be an important area of theoretical research. Areas in which they have found application include modelling molecular interaction, such as in protein folding (Jaimovich et al., 2006), natural language processing (Blei et al., 2003), speech recognition (Gales and Young, 2007), image segmentation (Zheng et al., 2015), target tracking (Stone et al., 1999), and coding and digital communication (MacKay and Neal, 1995), amongst many others. In addition, theoretical questions concerning inference and computation with graphical models have fundamental connections with a diverse range of disciplines, including statistical physics (Mezard and Montanari, 2009) and communication theory (MacKay, 2002).

In this thesis, we focus on discrete undirected graphical models, in which each random variable takes values in a finite set. We begin with a formal definition.

Definition 1.1 (**Discrete undirected graphical models**)**.** A *discrete undirected graphical model* M is a tuple $(V, E, (\mathcal{X}_v | v \in V), (\theta_{\mathcal{E}} | \mathcal{E} \in E))$ given by: (i) a hypergraph $G = (V, E)$ with

vertex set V and hyperedge set $E \subseteq \mathcal{P}(V)$, where $\mathcal{P}(V)$ is the powerset of V; (ii) a finite domain $\mathcal{X}_v$ for each vertex $v \in V$; and (iii) a *potential function* $\theta_{\mathcal{E}} : \mathcal{X}_{\mathcal{E}} \to \mathbb{R}$ for each hyperedge $\mathcal{E} \in E$, where $\mathcal{X}_A = \prod_{v \in A} \mathcal{X}_v$ for all $A \subseteq V$. This collection of objects gives rise to a probability distribution for a collection of random variables $(X_v | v \in V)$, given by

$$\mathbb{P}(X_V = x_V) = \frac{1}{Z} \exp\left(\sum_{\mathcal{E} \in E} \theta_{\mathcal{E}}(x_{\mathcal{E}}) \right) \qquad \forall x_V \in \mathcal{X}_V . \tag{1.1}$$

Here we adopt the notation $X_U = (X_v | v \in U)$ and $x_U = (x_v | v \in U)$ for all $U \subseteq V$. The *normalising constant* (or *partition function*) Z, appearing in Equation (1.1), is given by

$$Z = \sum_{x_V \in \mathcal{X}_V} \exp\left(\sum_{\mathcal{E} \in E} \theta_{\mathcal{E}}(x_{\mathcal{E}}) \right) . \tag{1.2}$$

The argument of the exponential function in Equation (1.1),

$$\sum_{\mathcal{E} \in E} \theta_{\mathcal{E}}(x_{\mathcal{E}}) , \tag{1.3}$$

is referred to as the *score function* associated with the model, and the value of this function for a particular configuration $x_V \in \mathcal{X}_V$ is referred to as the *score* of the configuration x_V. Thus, for a particular configuration, a higher score leads to a higher probability.

What distinguishes a discrete undirected graphical model M from a mere joint distribution of a set of discrete random variables X_V is the topological information supplied by the hypergraph G and the corresponding factorisation of the joint probability mass function over the set of hyperedges E of G. The potential functions $(\theta_{\mathcal{E}} | \mathcal{E} \in E)$ may be thought of as specifying "local" preferences for the configurations of the random variables $X_{\mathcal{E}}$; the global probability distribution in Equation (1.1) represents the cumulative effects of these local preferences. This is particularly attractive from the point of view of probabilistic modelling, as it means that a practitioner may introduce inductive biases and domain expertise into their models, which can often result in sample-efficient learning.

The availability of the topological information is also extremely useful from the point of view of performing inference on a given model. In general, it is computationally infeasible to perform inference on joint distributions over anything but the smallest sets of random variables (in a sense that we make precise in the sections that follow), but often sparsity and structure in the topology of G, or regularity in the potential functions $(\theta_{\mathcal{E}} | \mathcal{E} \in E)$, can be exploited to design fast, general algorithms for inference that scale to models over many variables.

Our primary contributions in the first half of this thesis are concerned with gaining greater understanding of the relationships between graph structure and potential function regularity, and exactness guarantees of methods for approximate maximum a posteriori (MAP) inference — see Section 1.2 for definitions.

Throughout the remainder of this thesis, we will refer to discrete undirected graphical models generically as *graphical models*. Of particular interest will be *binary* graphical models, where the random variable domains $\mathcal{X}_v$ ($v \in V$) appearing in Definition 1.1 are each taken to be the set $\{0,1\}$. When discussing binary graphical models specifically, we will often write the tuple specifying the model as $M = (V, E, (\theta_{\mathcal{E}} | \mathcal{E} \in E))$, leaving out explicit specification of the spaces $(\mathcal{X}_v | v \in V)$.

1.2 Inference in graphical models

The graphical models literature defines two broad categories of problems associated with graphical models: those of *learning*, and of *inference* (Koller and Friedman, 2009). It is the latter that we are primarily concerned with in the thesis, but we briefly discuss the former for completeness. Learning consists of specifying a collection of graphical models, observing some kind of empirical data, and performing some kind of selection over the collection of models, with the aim of explaining the observed data well. This selection could be based on maximum likelihood learning (see e.g. Ackley et al., 1985), in which case a single graphical model is selected from the class, or a Bayesian perspective could be taken, in which case beliefs about which models are likely to be responsible for the data generating process are expressed as a distribution over the collection of models (see e.g. Murray and Ghahramani, 2004). Several important subproblems in learning are those of *parameter learning*, in which a hypergraph $G = (V, E)$ is fixed and potential functions $(\theta_{\mathcal{E}} | \mathcal{E} \in E)$ are learnt from data, and *structure learning* (see Dempster, 1972 for early work on structure learning for Gaussian graphical models, and Wainwright et al. (2006) for an example of more recent work on discrete graphical models), in which the hypergraph itself is learnt from data. In many cases, such as when dealing with unobserved latent variables, successful learning requires the solution of many *inference* tasks along the way, such as in the classical EM algorithm (Dempster et al., 1977) and variations thereof. We now describe the problems of inference in greater detail.

Informally, inference may be summarised as the answering of probabilistic queries addressed to a fixed graphical model. In a graphical model designed for medical diagnosis

representing the presence of symptoms and causes in individuals (see e.g. Lauritzen and Spiegelhalter (1988); Shwe et al. (1991)), example queries might be "what is the probability my cough is caused by a chest infection, given that I have no other symptoms?", or "what is the most likely cause of a blocked nose?". Such queries generally either require the extraction of probabilities of particular events (perhaps conditional on other events) from the model, or the extraction of the most likely settings of random variables in the model. Note that in contrast to the learning problem discussed above, inference problems are primarily computational, as opposed to statistical, in nature. We define these two inference tasks formally below.

Definition 1.2 (**MAP inference**)**.** The task of *MAP inference* on a model $M = (V, E, (\mathcal{X}_v | v \in V), (\theta_{\mathcal{E}} | \mathcal{E} \in E))$ consists of finding a most likely configuration of the random variables X_V under M. Using the notation introduced in Equation (1.1), this may be expressed mathematically as finding the maximiser(s) of the score function:

$$\operatorname*{arg\,max}_{x_V \in \mathcal{X}_V} \sum_{\mathcal{E} \in E} \theta_{\mathcal{E}}(x_{\mathcal{E}}) \,. \tag{MAP}$$

The term *MAP* comes from *maximum a posteriori* in Bayesian inference. We emphasise however that MAP inference, as defined here, may be performed on any graphical model, regardless of whether or not it has a Bayesian interpretation.

Definition 1.3 (**Marginal inference**)**.** The task of *marginal inference* on a model M is specified by a subset $U \subseteq V$, and consists of computing the marginal probabilities

$$\mathbb{P}(X_U = x_U) \quad \left[= \sum_{x_{V \setminus U} \in \mathcal{X}_{V \setminus U}} \frac{1}{Z} \exp\left(\sum_{\mathcal{E} \in E} \theta_{\mathcal{E}}(x_{\mathcal{E}}) \right) \right], \qquad x_U \in \mathcal{X}_U \,. \tag{MI}$$

Note that marginal inference is non-trivial since graphical models are specified so that the score function of the model is straightforward to evaluate for a given configuration, but the normalising constant Z is defined only implicitly, as in Equation (1.2). We note that computation of Z, and computation of conditional probabilities, may be straightforwardly reduced to marginal inference as defined in Definition 1.3 via Bayes' rule.

In general both the MAP inference problem (MAP) and the marginal inference problem (MI) are intractable, in the following sense. MAP inference is NP-hard, which may be demonstrated by straightforward reductions from variants of several of Karp's classically NP-complete problems (Karp, 1972). For example, Barahona (1982) exhibits a reduction of the problem of finding a maximum independent set in a planar cubic graph to MAP

inference in binary graphical models over planar graphs G, showing that (MAP) is NP-hard, even on this restricted class of problems. Jerrum and Sinclair (1993) exhibit a reduction of #MAX CUT to marginal inference in binary graphical models over graphs G, demonstrating that the marginal inference problem (MI) is #P-hard in general.

These hardness results first serve to set expectations; unless P = NP, we cannot expect there to exist an efficient (i.e. polynomial time) algorithm which exactly solves either (MAP) or (MI) in generality, even with structural restrictions on the graph G such as maximum hyperedge degree ≤ 2, or planarity as in the comments above. We are thus motivated to study approximate inference algorithms — methods for approximating the solution of (MAP) and (MI). Our original contributions in the chapters that follow are centred around approximate methods for MAP inference, so we focus on such methods for the remainder of this chapter.

Problems equivalent to (MAP) have arisen in many fields in a variety of guises, such as discrete energy minimisation (Kappes et al., 2015), valued constraint satisfaction (Schiex et al., 1995), quadratic pseudo-boolean optimisation (Hammer et al., 1984), and many problems in combinatorial optimisation (Cook et al., 1998). As such, many approaches to approximately solving (MAP) (and special cases thereof) have been studied in the literature, such as (loopy) belief propagation (Pearl, 1988), semi-definite programming relaxations (Erdogdu et al., 2017; Laurent, 2003), graph cuts (Boykov et al., 2001; Kohli and Torr, 2007), branch-and-bound approaches (Land and Doig, 1960), and randomised rounding (Goemans and Williamson, 1995). Further, MAP inference has been studied for specific families of graphical models, such as the Viterbi algorithm for hidden Markov models (Viterbi, 1967) before a unified approach to the problem arose. In Section 1.3, we review approximate MAP inference algorithms based on linear programming relaxations, which form the main focus of Chapters 2 to 4 that follow.

1.3 Linear programming relaxations for MAP inference

One method for MAP inference of particular interest in this thesis is linear programming relaxations via the Sherali–Adams polytope hierarchy, for which we give a brief exposition in Section 1.3.1. The specific relaxations described below were first proposed for 0-1 polynomial optimisation by Sherali and Adams (1990), and have more recently been the subject of much research in the graphical models community, with a range of practical and theoretical developments put forward; see Batra et al. (2011); Sontag (2010); Sontag et al.

(2008); Wainwright and Jordan (2008, 2004) for a representative sample of publications. The Sherali–Adams polytope hierarchy is an attractive object of study, as it also plays a key role in important variational approaches to approximate marginal inference, such as Bethe-Kikuchi approximations (Yedidia et al., 2001), which are themselves intimately related to loopy belief propagation algorithms. For a more detailed background to Sherali–Adams polytopes for MAP inference, the reader may refer to Wainwright and Jordan (2008).

The approach to performing approximate MAP inference via the Sherali–Adams hierarchy is to first express Problem (MAP) as a linear program (LP), and then to relax the feasible region of the LP to a *larger* region described by *fewer* constraints. This relaxed problem is thus generally computationally easier to solve, but its optima may not correspond to a feasible solution of the original problem. In this section, we first give a brief exposition of the derivation of the relaxed LP mentioned above, and then discuss the computational and exactness issues that arise from the relaxation.

1.3.1 From combinatorial optimisation to linear programming

The approach of LP relaxations for combinatorial optimisation problems in general is threefold: firstly, the objective function of the problem is linearised; secondly, the problem is expressed as an equivalent linear program; and finally, the feasible region of the linear program is relaxed — that is, it is replaced with a larger feasible region, typically described by vastly fewer constraints. It is this last step from which gains in computational tractability arise, since linear programs with fewer constraints are generally computationally less intensive to solve. However, since the final step enlarges the feasible region of the LP, it also has the possibility of intro-

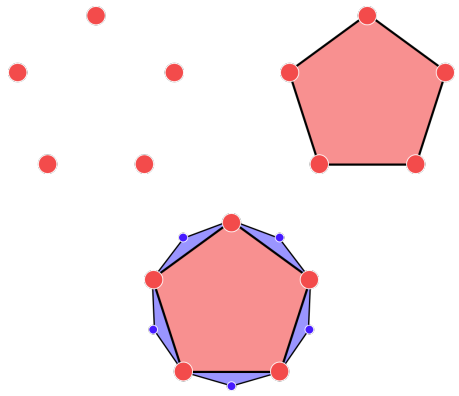


Figure 1.1. *Top left:* schematic representation of the discrete set of configurations of a graphical model. *Top right:* Optimising over the convex hull of these points results in an equivalent LP. *Bottom:* A relaxed LP optimises over a larger feasible region described by fewer constraints and with additional extrema, which do not correspond to valid graphical model configurations. It is not possible to accurately depict the decrease in the number of inequalities of the relaxed polytope in two dimensions. Figure influenced by Figure 2.3 of Sontag (2010).

ducing spurious additional solutions which are not valid for the original problem — in the context of MAP inference, this may mean that solving the relaxed LP does not return a valid configuration for the graphical model. An important theme of combinatorial optimisation is how to choose the relaxed feasible region to trade off computational efficiency against errors introduced by the approximation (Korte and Vygen, 2007).

We now give a brief exposition of Sherali–Adams relaxations for MAP inference, illustrating the general principles of LP relaxations for combinatorial optimisation described above. We treat each of the three general areas above (linearisation, expression as an LP, and relaxation) in turn.

Linearising the problem. We begin by recalling the form of the MAP inference optimisation problem.

$$\operatorname*{arg\,max}_{x_V \in \mathcal{X}_V} \sum_{\mathcal{E} \in E} \theta_{\mathcal{E}}(x_{\mathcal{E}}) \,. \tag{MAP}$$

In general, the input space $\mathcal{X}_V$ need not have any additive structure, so the question of whether the objective is linear is not well posed. To achieve linearisation, we switch perspectives, viewing the objective as the expectation of a fixed function against a varying probability measure supported at a particular configuration. Writing $\delta_{x_V} \in \mathscr{P}(\mathcal{X}_V)$ for the probability distribution over configurations of the graphical model which places all its mass at the configuration $x_V \in \mathcal{X}_V$, we may express the MAP inference problem as:

$$\operatorname*{arg\,max}_{\mu \in \{\delta_{x_V} | x_V \in \mathcal{X}_V\}} \sum_{\mathcal{E} \in E} \mathbb{E}_{X_V \sim \mu} \left[\theta_{\mathcal{E}}(X_{\mathcal{E}})\right] , \tag{1.4}$$

which is *linear* in the input probability distribution. Variants of this linearisation trick are often used generally in the area of 0-1 polynomial optimisation (Padberg, 1989).

Expressing the problem as a linear program. Since Problem (1.4) has a linear objective, with a finite set of feasible points, we may optimise over the convex hull of these points to obtain an equivalent problem. In this vein, recall that the set of extremal points of the convex body $\mathscr{P}(\mathcal{X}_V)$, written $\operatorname{Ext}(\mathscr{P}(\mathcal{X}_V))$, is precisely $\{\delta_{x_V} | x_V \in \mathcal{X}_V\}$. It therefore follows that the following problem has the same optimal value as Problem (1.4):

$$\operatorname*{arg\,max}_{\mu \in \mathscr{P}(\mathcal{X}_V)} \sum_{\mathcal{E} \in E} \mathbb{E}_{X_V \sim \mu} \left[\theta_{\mathcal{E}}(X_{\mathcal{E}})\right] . \tag{1.5}$$

Further, the probability measures attaining the maximal value in Problem (1.5) supported on a single configuration are precisely those attaining the maximal value in Problem (1.4). Thus, Problem (1.5) has a linear objective and a convex feasible region (in the vector space of finite measures over $\mathcal{X}_V$). Moreover, the feasible region is described by a finite number of linear constraints on the measures concerned (in particular, by enforcing $\mu(\{x_V\}) \geq 0$ for all $x_V \in \mathcal{X}_V$, and $\mu(\mathcal{X}_V) = 1$. Thus, we may regard Problem (1.5) as a linear program, although we note that this is a slightly more abstract notion of a linear program than may typically be considered, since we have not identified the feasible region directly with a particular polytope in a Euclidean space (such an identification is typically made before solving an LP computationally, for example). It is however straightforward to recover a concrete LP from Problem (1.5) by selecting an appropriate parametrisation — further details are given in Section 1.4.2. This parametrisation-free perspective is not generally emphasised in expositions of the Sherali–Adams hierarchy, but we will find it particularly useful to discuss the marginal polytope (see Definition 1.4) and Sherali–Adams polytopes (see Definitions 1.6 and 1.15) in this way in describing our original work, particularly in Chapter 2.

The fact that Problem (1.5) depends only on the hyperedge marginals $(\mu_{\mathcal{E}} | \mathcal{E} \in E)$ of μ motivates the following definition.

Definition 1.4 (**Marginal polytope**)**.** For a hypergraph $G = (V, E)$, the marginal polytope is given by

$$\mathbb{M}(G) = \left\{ (\mu_{\mathcal{E}})_{\mathcal{E} \in E} \in \prod_{\mathcal{E} \in E} \mathscr{P}(\mathcal{X}_{\mathcal{E}}) \;\middle|\; \exists \mu_V \in \mathscr{P}(\mathcal{X}_V) \text{ such that } \mu_{V\downarrow\mathcal{E}} = \mu_{\mathcal{E}} \; \forall \mathcal{E} \in E \right\}, \tag{1.6}$$

where for two sets $B \subseteq A \subseteq V$, and a probability distribution $\mu_A \in \mathscr{P}(\mathcal{X}_A)$, the notation $\mu_{A\downarrow B} \in \mathscr{P}(\mathcal{X}_B)$ is the marginalisation of μ_A onto the set $\mathcal{X}_B$. More precisely:

$$\mu_{A\downarrow B}(x_B) = \sum_{x_{A\setminus B} \in \mathcal{X}_{A\setminus B}} \mu_A(x_B, x_{A\setminus B}), \qquad \forall x_B \in \mathcal{X}_B. \tag{1.7}$$

Thus, an element μ of the marginal polytope $\mathbb{M}(G)$ is a collection of probability distributions $(\mu_{\mathcal{E}} | \mathcal{E} \in E)$, one for each hyperedge of G, with the property that all distributions are marginals of a single global distribution over $\mathcal{X}_V$. With this definition established, we may re-express Problem (1.5) as follows:

$$\operatorname*{arg\,max}_{\mu \in \mathbb{M}(G)} \sum_{\mathcal{E} \in E} \mathbb{E}_{X_{\mathcal{E}} \sim \mu_{\mathcal{E}}} \left[\theta_{\mathcal{E}}(X_{\mathcal{E}}) \right]. \tag{1.8}$$

We will refer to the sum in Expression (1.8) as the score of the collection of marginals $(\mu_{\mathcal{E}} | \mathcal{E} \in E)$, because of its similarity with Expression (1.3). This is again an equivalent formulation of the original MAP inference problem. Again, as with our qualifications regarding the use of the term "linear program" to describe Problems (1.5) and (1.8), we note that strictly speaking, $\mathbb{M}(G)$ is not a polytope in the usual sense, as it is not yet identified as a bounded region of a Euclidean space described by finitely many linear constraints. However, $\mathbb{M}(G)$ is indeed the convex hull of the finite set

$$\left\{ (\delta_{x_{\mathcal{E}}} | \mathcal{E} \in E) \;\middle|\; x_V \in \mathcal{X}_V \right\} \subseteq \prod_{\mathcal{E} \in E} \mathscr{P}(\mathcal{X}_{\mathcal{E}}) \,. \tag{1.9}$$

We discuss these issues of parametrisation in Section 1.4.2.

Relaxing the problem. We now take the linear program in Problem (1.8), and replace the feasible region $\mathbb{M}(G)$ with a larger region that may be described with fewer constraints. The approach proposed by Sherali and Adams (1990) is to replace the constraint that a collection of measures $(\mu_{\mathcal{E}} | \mathcal{E} \in E)$ must be globally consistent (in the sense of the existence of a distribution $\mu_V \in \mathscr{P}(\mathcal{X}_V)$ for which each distribution $\mu_{\mathcal{E}}$ is a marginal) with a weaker form of consistency, described below.

Definition 1.5 (**Local consistency**)**.** Given a hypergraph $G = (V, E)$, two subsets $A, B \subseteq V$, and two probability distributions $\mu_A \in \mathscr{P}(\mathcal{X}_A)$, $\mu_B \in \mathscr{P}(\mathcal{X}_B)$, we say that the pair of probability measures μ_A and μ_B is locally consistent if $\mu_{A\downarrow(A\cap B)} = \mu_{B\downarrow(A\cap B)}$ as probability measures in $\mathscr{P}(\mathcal{X}_{A\cap B})$. More generally, given $A_1, \ldots, A_n \subseteq V$, and $\mu_{A_i} \in \mathscr{P}(\mathcal{X}_{A_i})$ for each $i = 1, \ldots, n$, we say that the *collection* of probability measures $\mu_{A_1}, \ldots, \mu_{A_n}$ are locally consistent if each pair of probability measures is locally consistent in the sense described above.

With the definition of local consistency in hand, we may now describe the Sherali–Adams hierarchy of relaxations of $\mathbb{M}(G)$.

Definition 1.6 (**Sherali–Adams polytope**)**.** For a hypergraph $G = (V, E)$, and an integer $r \geq 2$ such that the maximum hyperedge degree of G is less than or equal to r, the r^{th}-level Sherali–Adams polytope is given by

$$\mathbb{L}_r(G) = \left\{ \left(\mu_{\mathcal{E}}\right)_{\mathcal{E}\in E} \in \prod_{\mathcal{E}\in E} \mathscr{P}(\mathcal{X}_{\mathcal{E}}) \middle| \exists (\mu_U)_{\substack{U\subseteq V \\ |U|=r}} \in \prod_{\substack{U\subseteq V \\ |U|=r}} \mathscr{P}(\mathcal{X}_U) \text{ locally consistent,} \right.$$
$$\left. \text{s.t. } \mu_{U\downarrow\mathcal{E}} = \mu_{\mathcal{E}} \quad \forall\, \mathcal{E} \subseteq U \subseteq V,\ |U| = r \right\}. \tag{1.10}$$

Given a collection of measures $\mu = (\mu_{\mathcal{E}})_{\mathcal{E}\in E} \in \mathbb{L}_r(G)$, we call any locally consistent collection of measures $(\mu_U \in \mathscr{P}(\{0,1\}^U) | U \subseteq V, |U| = r)$ which satisfies the marginalisation property described in Equation (1.10) a *consistent marginalising family* for $(\mu_{\mathcal{E}})_{\mathcal{E}\in E}$ with respect to $\mathbb{L}_r(G)$.

In words, $\mathbb{L}_r(G)$ imposes weaker conditions on a collection of measures $(\mu_{\mathcal{E}} | \mathcal{E} \in E) \in \mathbb{L}_r(G)$ than arising as marginals from some global probability distribution $\mu_V \in \mathscr{P}(\mathcal{X}_V)$. For this reason, an element $(\mu_{\mathcal{E}} | \mathcal{E} \in E) \in \mathbb{L}_r(G)$ is referred to as a collection of *pseudomarginals*.

Finally, we may now relax Problem (1.8) by using a Sherali–Adams relaxation $\mathbb{L}_r(G)$ in place of the marginal polytope $\mathbb{M}(G)$ as the feasible region, to obtain the following relaxed linear program:

$$\underset{\mu\in\mathbb{L}_r(G)}{\arg\max} \sum_{\mathcal{E}\in E} \mathbb{E}_{X_{\mathcal{E}}\sim\mu_{\mathcal{E}}} \left[\theta_{\mathcal{E}}(X_{\mathcal{E}})\right] . \tag{1.11}$$

Thus, an approximate method for solving (MAP) is to solve this relaxed problem. Since we have enlarged the feasible region of this maximisation problem, it is immediate that the optimal value for Problem (1.11) is at least as great as that of the exact problem (1.8). Indeed, we note that Problems (1.4), (1.8) and (1.11) correspond to the three schematic diagrams presented earlier in Figure 1.1. A crucial question of theoretical interest is under what circumstances will solving the relaxed Problem (1.11) yield a solution to the original Problem (1.8)? It has been observed that many real-world instances of MAP inference problems can be solved by relaxations of the form given in Problem (1.11) (Batra et al., 2011; Komodakis and Paragios, 2008; Sontag et al., 2008). This is perhaps surprising, given the worst-case hardness results for MAP inference in general. Several of our original contributions in later chapters focus on understanding under what conditions relaxations in Problem (1.11) provide exact answers to the MAP inference problem.

Before addressing the question of exactness in more detail, we first briefly discuss the computational gains that can be made by passing to Problem (1.11). We will see in Section 1.4.2 that a parametrisation of $\mathbb{M}(G)$ yielding a linear program from Problem (1.8) will contain polynomially many variables (in $|V|$), subject to the constraint of fixed maximum hyperedge degree. Now recall that the original MAP inference problem (MAP) is NP-hard, even when restricted to graphs G (i.e. hypergraphs with maximum hyperedge degree 2). Thus, unless P=NP, there must be exponentially many (in $|V|$) linear constraints that describe the feasible region of any linear program derived from Problem (1.8) under such a parametrisation. However, under such a parametrisation, the feasible region derived from Problem (1.11) using the relaxation $\mathbb{L}_r(G)$ requires only $\mathcal{O}(|V|^r)$ linear constraints to

define, as we shall see in Section 1.4.2. Thus, passing from Problem (1.8) to Problem (1.11) represents a move from an NP-hard class of optimisation problems to a set of problems that may be solved in polynomial time.

1.3.2 Tightness of Sherali–Adams relaxations

Given these observations, a crucial concern that remains is when the original optimisation problem (1.8) and the relaxed problem (1.11) yield the same answer. When this is indeed the case, we obtain the correct answer to the MAP inference problem by solving an LP with far fewer constraints than in the original formulation, in many cases representing a speed-up of many orders of magnitude. We say that the relaxation is *tight* in this case.

Here we give a brief summary of some key results from the literature on tightness of Sherali–Adams relaxations for MAP inference. A first result is the following series of inclusions, which follows immediately from the definition of local consistency above.

Proposition 1.7. Let $G = (V, E)$ be a hypergraph, and let r be the maximum degree of a hyperedge in G. Then from Definitions 1.4 and 1.6, the following sequence is immediate:

$$\mathbb{M}(G) = \mathbb{L}_{|V|}(G) \subseteq \mathbb{L}_{|V|-1}(G) \subseteq \cdots \mathbb{L}_{r+1}(G) \subseteq \mathbb{L}_r(G)\,. \tag{1.12}$$

An immediate corollary is that if relaxation onto $\mathbb{L}_r(G)$ is tight for a particular problem, then so too is relaxation onto any $\mathbb{L}_k(G)$ with $k > r$.

A first result as to which kinds of MAP inference problems a Sherali–Adams relaxation is tight for is the following.

Theorem 1.8 (Padberg, 1989)**.** Let G be a tree. Then $\mathbb{L}_2(G) = \mathbb{M}(G)$.

There are strong parallels here with guarantees of exactness for variants of belief propagation, which are also well known to be exact for trees. We present two further important generalisations of this result in Sections 1.3.3 and 1.3.4, which in some sense can be thought of as generalising the “tree” property in two different ways.

1.3.3 Structural conditions for tightness

The first generalisation we review can naturally be thought of as interpreting the "tree" condition on G as a *structural* condition on the underlying (hyper)graph of the model. To state the generalisation, we first require two preliminary definitions.

Definition 1.9 (**Tree decomposition**)**.** Let $G = (V, E)$ be a hypergraph. A tree decomposition of G is a tree T and a collection of subsets $(H_t | t \in T)$ of V, with the property that:

- $\cup_{t \in T} H_t = V$,
- If $\mathcal{E} \in E$, then there exists $t \in T$ such that $\mathcal{E} \subseteq H_t$.
- For each $v \in V$, the subgraph of T spanned by the vertices $\{t \in T | v \in H_t\}$ is itself a tree.

The *width* of a tree decomposition is equal to $\max_{t \in T} |H_t| - 1$. We refer to each subset H_t as a *supervertex*, in order to distinguish from the standard vertices in V.

Definition 1.10 (**Treewidth**)**.** Let G be a hypergraph. The *treewidth* of G is the minimal width across all tree decompositions of G.

The notions of tree decomposition and treewidth are key to the *junction tree algorithm* (Lauritzen and Spiegelhalter, 1988), a method for exact inference. The junction tree algorithm exploits the fact that exact MAP inference and marginal inference can be performed efficiently over tree-structured graphical models via belief propagation (note that Theorem 1.8 shows that MAP inference can also be efficiently performed via LP relaxations), and therefore performs belief propagation over a tree decomposition corresponding to a graphical model of interest. The computation needed to perform the belief propagation is exponential in the size of the largest supervertex in a tree decomposition, and hence the treewidth of a graph G serves as a measure of the cost of exact inference via the junction tree algorithm.

With these definitions in hand, we may now recall the following result, giving sufficient conditions for tightness of Sherali–Adams relaxations in terms of treewidth.

Theorem 1.11 (Wainwright and Jordan, 2004)**.** Let G be a hypergraph with treewidth at most r. Then $\mathbb{L}_{r+1}(G) = \mathbb{M}(G)$.

This generalises Theorem 1.8, since graphs with treewidth one are precisely forests (graphs whose connected components are trees). We observe that in some sense, working with a tree decomposition T of a model turns the problem back into one of MAP inference on a (non-binary) model over the tree T, and that the higher the treewidth of the original graph

G, the larger the domains of the random variables in the model over T, resulting in more costly inference.

1.3.4 Conditions on potentials for tightness

If we restrict ourselves to the consideration of *binary* pairwise graphical models, where each domain $\mathcal{X}_v$ (for in $v \in V$) is given by $\{0,1\}$, Theorem 1.8 may be generalised in quite a different way. The condition of G being a tree is interpreted as a global condition on the structure of potentials that may be present in a model. To present the theorem, we first require a preliminary definition.

Definition 1.12 (**Attractive & repulsive potentials, attractive and balanced models**)**.** Let M be a binary pairwise model on a graph G. A pairwise potential function $\theta_{ij} : \{0,1\}^2 \to \mathbb{R}$ is said to be *attractive* if

$$\theta_{ij}(1,1) + \theta_{ij}(0,0) > \theta_{ij}(1,0) + \theta_{ij}(0,1)\,, \tag{1.13}$$

and *repulsive* if

$$\theta_{ij}(1,1) + \theta_{ij}(0,0) < \theta_{ij}(1,0) + \theta_{ij}(0,1)\,. \tag{1.14}$$

A binary pairwise graphical model is said to be *attractive* if all pairwise potentials are attractive. A cycle of edges with attractive and repulsive potentials is said to be *frustrated* if it contains an odd number of edges with repulsive potentials. A binary pairwise graphical model is said to be *balanced* if it has no frustrated cycles. In particular, attractive graphical models are balanced.

Intuitively, an attractive pairwise potential prefers the two variables concerned to take on the same value, whilst a repulsive potential prefers the two variables concerned to take on different values. Before stating the known connections between attractiveness of potentials and tightness of Sherali-Adams relaxations, we pause to highlight connections with similar concepts in other areas of statistics and optimisation.

We first note that if we allow equality between the two sides of Equation (1.13) as well as strict inequality, we obtain the notion of supermodularity, when the set $\{0,1\}^2$ is equipped with its usual product partial order; supermodular maximisation is a well-studied problem in combinatorial optimisation more generally (Cook et al., 1998).

Secondly, we note that attractiveness and balancedness of graphical models are closely related to the notion of *multivariate total positivity of order* 2 (MTP_2), an idea of increasing importance in the statistical literature. A density $f : \prod_{v\in V} \mathcal{X}_v \to [0,\infty)$ (with each $\mathcal{X}_v$ equipped with a total order, and not necessarily finite, and $\prod_{v\in V} \mathcal{X}_v$ equipped with the standard product lattice structure) is said to be MTP_2 if $f(x_V)f(y_V) \leq f(x_V \wedge y_V)f(x_V \vee y_V)$ for all $x_V, y_V \in \prod_{v\in V} \mathcal{X}_v$. It has long been established that a graphical model being MTP_2 implies statistical association of the random variables concerned (Esary et al., 1967; Fortuin et al., 1971; Holland and Rosenbaum, 1986), whilst at the same time being reasonably straightforward to check. More recently, MTP_2 has been shown to serve as a sparsity-inducing regularisation condition in maximum likelihood estimation for Gaussian graphical models (Lauritzen et al., 2018), and further, there exist efficient maximum likelihood optimisation algorithms under the condition of MTP_2 for binary (Bartolucci and Forcina, 2000) and Gaussian (Lauritzen et al., 2018) graphical models.

If f is positive everywhere, then MTP_2 is equivalent to log-supermodularity (Fallat et al., 2017). The work of Fallat et al. (2017) also shows that if f is positive and represents the probability distribution of a binary pairwise graphical model, then it is MTP_2 if and only if each pairwise potential is supermodular. Thus, the set of binary pairwise graphical models with the MTP_2 property is essentially the set of attractive graphical models; all pairwise potentials in an attractive graphical model are supermodular, and if a binary pairwise graphical model has pairwise potentials that are all supermodular, we note that for any pairwise potential θ_{ij} which is not attractive, the two sides of Expression (1.13) must be equal, and hence the pairwise potential can be removed and replaced by singleton potentials $\theta_i(x_i) = (\theta_{ij}(1,0) - \theta_{ij}(0,0))\mathbb{1}_{x_i=1}$, $\theta_j(x_j) = (\theta_{ij}(0,1) - \theta_{ij}(0,0))\mathbb{1}_{x_j=1}$ without affecting the probability distribution — all remaining pairwise potentials are then attractive.

Just as attractiveness generalises to balancedness, MTP_2 can be generalised to *signed* MTP_2, originally introduced for Gaussian models by Karlin and Rinott (1981). A Gaussian random variable Z taking values in $\mathbb{R}^m$ is said to be signed MTP_2 if there exists a diagonal matrix D with all diagonal entries in $\{\pm 1\}$ such that DZ is MTP_2. It can be shown (Lauritzen et al., 2018; Malioutov et al., 2006) that this characterisation is equivalent to the following condition on the precision matrix $K \in \mathbb{R}^{m\times m}$ associated with Z and the conditional independence graph $G(K) = (V = \{1,\dots,m\}, E)$ (with $ij \in E$ iff $K_{ij} \neq 0$): $(-1)^k K_{i_1 i_2} \cdots K_{i_{k-1} i_k} K_{i_k i_1} > 0$ for all cycles $(i_1,\dots,i_k,i_1)$ in $G(K)$. Thus, signed MTP_2 for multivariate Gaussian random variables has two equivalent characterisations: one based on flipping signs of coordinates, and one based on properties of cycles; the latter condition is

strongly reminiscent of the notion of balancedness described in Definition 1.12. A binary graphical model on $G = (V, E)$ is said to be signed MTP_2 if there is a "relabelling" of a subset of variables such that the resulting model is MTP_2 — we expand on this in Section 1.4.1, where we show that the set of binary pairwise graphical models that are signed MTP_2 is essentially the set of balanced binary pairwise graphical models. We conclude these remarks by reiterating that notions of balancedness, supermodularity, and signed MTP_2 are very closely related. Much of the statistical literature around the MTP_2 condition has focused on *learning* models under these constraints, whilst literature around balanced graphical models has often focused on the *computational* properties of inference problems associated with such models. We thus expect there to be profitable connections drawn between these two areas of the literature in future.

Returning to the connections between potential function structure and tightness of Sherali-Adams relaxations, the first result we give shows that restricting pairwise potentials to be attractive results in tightness of the lowest level of the Sherali–Adams hierarchy.

Theorem 1.13 (Padberg, 1989)**.** Let M be an attractive binary pairwise graphical model. Then $\mathbb{L}_2(G)$ is tight for M.

As remarked in Definition 1.12, any attractive model is balanced. It turns out that a version of Theorem 1.13 holds for balanced models more generally — we discuss the connection between balanced and attractive models in more detail in Section 1.4.1. Finally, observe that any binary graphical model on a tree is balanced, because the tree clearly has no frustrated cycles. With this observation, it is now clear that the following theorem generalises Theorem 1.8.

Theorem 1.14 (Padberg, 1989)**.** Let M be a balanced binary pairwise graphical model. Then $\mathbb{L}_2(G)$ is tight for M.

Note that these results are of a different type to Theorem 1.11; generally, it is not the case that $\mathbb{L}_2(G) = \mathbb{M}(G)$, but that the relaxation is tight for particular models with certain restrictions over their potentials.

1.3.5 Unifying conditions for tightness

In the previous two sections, we have seen two different types of conditions for ensuring tightness of a Sherali–Adams relaxation: one type that relied on structural restrictions on the (hyper)graph G, and another that relied on restrictions on the types of potentials that may be present in a model. We summarise these theoretical results and illustrate some of

the classes of binary pairwise graphical models concerned in a Venn diagram in Figure 1.2. As reviewed above, the relaxation $\mathbb{L}_2(G)$ is known to be tight for all balanced models, which contains the set of all models of treewidth 1 as a subclass. Similarly, $\mathbb{L}_3(G)$ is known to be tight for all models of treewidth ≤ 2, as is $\mathbb{L}_4(G)$ for all models of treewidth ≤ 3.

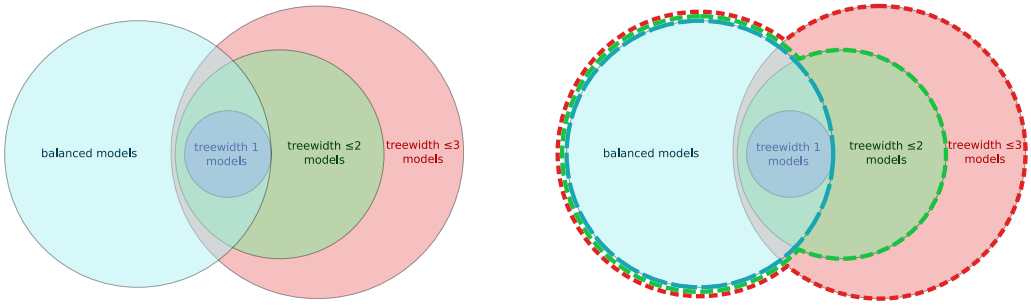


Figure 1.2. Left: Venn diagram of some binary pairwise graphical model classes described in Theorems 1.8, 1.11, and 1.14. Right: Dotted lines indicate classes of binary models guaranteed to be tight for $\mathbb{L}_2(G)$ (teal), $\mathbb{L}_3(G)$ (green), and $\mathbb{L}_4(G)$ (red).

It is natural to ask whether there are broader classes of sufficient conditions for ensuring tightness of Sherali–Adams relaxations. Indeed, on viewing Figure 1.14, natural questions to ask are: (i) whether the treewidth characterisations of tightness may be expanded to larger classes of models?; (ii) in what ways the class of balanced models might be expanded for higher-order Sherali–Adams relaxations?; and (iii) whether there are sufficient conditions for tightness that make use of both structural and potential function restrictions? Several of our main contributions in the chapters that follow concern exactly these questions.

1.3.6 Generalised Sherali–Adams polytopes

The Sherali–Adams relaxations $\mathbb{L}_r(G)$ introduced in Section 1.3.1 enforce local consistency on a collection of measures $(\mu_{\mathcal{E}} | \mathcal{E} \in E)$ across all sets of variables of size r. In practice, the computational cost involved in working with these relaxations grows sharply with the index r, which motivates considering more nuanced consistency constraints; rather than enforcing local consistency *uniformly* across all r-clusters in a graph, it may often be of interest to enforce higher degrees of local consistency in certain regions of the graph G. This leads to the notion of a generalised Sherali–Adams relaxation, defined below.

Definition 1.15 (**Covering antichains, generalised Sherali–Adams polytopes**)**.** Let $G = (V, E)$ be a hypergraph, and consider an antichain $S \subseteq \mathcal{P}(V)$ (that is, a collection of subsets

of V, none of which is contained in another). If S has the property that for each $\mathcal{E} \in E$, there exists $A \in S$ such that $\mathcal{E} \subseteq A$, we say that S *covers* G. If the antichain S covers G, the *generalised Sherali–Adams polytope* for the antichain S is defined by

$$\mathbb{L}_S(G) = \Bigg\{ \big(\mu_{\mathcal{E}}\big)_{\mathcal{E}\in E} \in \prod_{\mathcal{E}\in E} \mathscr{P}(\{0,1\}^{\mathcal{E}}) \Bigg| \exists (\mu_U)_{U\in S} \in \prod_{U\in S} \mathscr{P}(\{0,1\}^U) \text{ locally consistent,} \\ \text{s.t. } \mu_{U\downarrow\mathcal{E}} = \mu_{\mathcal{E}} \quad \forall\, \mathcal{E} \subseteq U \in S \Bigg\}. \tag{1.15}$$

We note that this overloads the notation $\mathbb{L}_r(G)$ used to represent the standard Sherali–Adams polytopes in Equation (1.10). Indeed, this notion generalises the standard Sherali–Adams hierarchy; the relaxation $\mathbb{L}_r(G)$ may be recovered by taking $S = V^{(r)}$, the set of all subsets of V of size r. We shall see in Chapter 3 that this notion leads to more nuanced descriptions of tightness of model classes. Relaxations of this form have been considered under various guises in the literature (Batra et al., 2011; Sontag et al., 2008; Yedidia et al., 2001).

1.4 Additional background material

In this section, we give additional background material that will be useful in the chapters that follow. Firstly, we consider a family of model transformations known as flippings, and show that these transformations have useful properties with regards to the study of tightness of Sherali–Adams relaxations. Secondly, we consider the issues surrounding parametrisation of the marginal polytope and Sherali–Adams relaxations mentioned in Section 1.3.1.

1.4.1 Flipping

Flippings are transformations of binary graphical models and related polytopes. Intuitively, a flipping transformation relabels a specified subset of random variables. This notion is of much use in studying tightness of Sherali–Adams relaxations, and will be used in the chapters that follow. This is made precise in the definitions below; we begin with the notion of a flipped model.

Definition 1.16 (**Flipped models**)**.** Given a binary graphical model specified by a hypergraph $G = (V, E)$ and potentials $(\theta_{\mathcal{E}} | \mathcal{E} \in E)$, the model obtained by flipping a subset $A \subseteq V$

of variables is defined by the same hypergraph G, and a new collection of potentials $(\overline{\theta}_{\mathcal{E}}^{(A)} | \mathcal{E} \in E)$, defined by

$$\overline{\theta}_{\mathcal{E}}^{(A)}(x_{\mathcal{E}}) = \theta_{\mathcal{E}}(x_{\mathcal{E}\setminus A}, \overline{x}_{\mathcal{E}\cap A}) \qquad \forall \mathcal{E} \in E, x_{\mathcal{E}} \in \{0,1\}^{\mathcal{E}}, \tag{1.16}$$

where for a subset $U \subseteq V$, $\overline{x}_U = (1 - x_i | i \in U)$.

Thus, given a model $(G = (V,E), (\theta_{\mathcal{E}} | \mathcal{E} \in E))$, and a collection of random variables $(X_v | v \in V)$ distributed according to this model, if a subset of variables $U \subseteq V$ is flipped to yield the model $(G = (V,E), (\overline{\theta}_{\mathcal{E}}^{(U)} | \mathcal{E} \in E))$, then the collection of random variables $(Y_v | v \in V)$, where $Y_v = 1 - X_v$ for $v \in U$ and $Y_v = X_v$ otherwise, is distributed according to the flipped model. This naturally motivates us to consider flipped variable configurations, defined as follows.

Definition 1.17 (**Flipped configurations**)**.** Given $x_V \in \{0,1\}^V$ and some subset $U \subseteq V$, the configuration given by flipping U is defined by $\overline{x}_V^{(U)} = (x_{V\setminus U}, \overline{x}_U)$, where $\overline{x}_U = (1 - x_i | i \in U)$.

It is immediate that flipping a subset $U \subseteq V$ induces a bijection on the space of configurations $\{0,1\}^V$. The notion of flipping a configuration also generalises to a notion of flipping for collections of pseudomarginals over hyperedges.

Definition 1.18 (**Flipped pseudomarginals**)**.** Given a hypergraph $G = (V,E)$, a (generalised) Sherali–Adams polytope, and a collection of pseudomarginals $(\mu_{\mathcal{E}} | \mathcal{E} \in E) \in \mathbb{L}_S(G)$, the flipped pseudomarginals obtained by flipping a subset $A \subseteq V$ of variables are written $(\overline{\mu}_{\mathcal{E}}^{(A)} | \mathcal{E} \in E)$, and defined by

$$\overline{\mu}_{\mathcal{E}}^{(A)}(x_{\mathcal{E}}) = \mu_{\mathcal{E}}(x_{\mathcal{E}\setminus A}, \overline{x}_{\mathcal{E}\cap A}). \tag{1.17}$$

We now note a key property of flipped models and polytopes below, which will be used several times in the chapters that follow; this is one of the principal reasons for considering flipping transformations.

Proposition 1.19 (**Properties of flippings**)**.** (i) For any hypergraph $G = (V,E)$ and (generalised) Sherali–Adams polytope $\mathbb{L}_S(G)$, flipping a subset of variables $U \subseteq V$ induces an affine bijection on $\mathbb{L}_S(G)$. (ii) A relaxation $\mathbb{L}_S(G)$ is tight for a model $M = (V,E,(\theta_{\mathcal{E}} | \mathcal{E} \in E))$ iff it is tight for all flippings of M.

Proof. For (i), we first simply note that the definition of the flipping mapping for a Sherali–Adams polytope $\mathbb{L}_S(G)$ is affine, and is clearly self-inverse. These two properties together immediately imply that the flipping map must map $\mathbb{L}_S(G)$ onto itself, and so the statement

follows. For (ii), we consider flipping a subset $U \subseteq V$ of variables, and note

$$\begin{aligned}\max_{\mu \in \mathbb{L}_S(G)} \sum_{\mathcal{E} \in E} \mathbb{E}_{X_\mathcal{E} \sim \mu_\mathcal{E}} [\theta_\mathcal{E}(X_\mathcal{E})] &= \max_{\mu \in \mathbb{L}_S(G)} \sum_{\mathcal{E} \in E} \mathbb{E}_{X_\mathcal{E} \sim \mu_\mathcal{E}} \left[\overline{\theta}_\mathcal{E}^{(U)}(\overline{X}_\mathcal{E}^{(U)})\right] \\ &= \max_{\mu \in \mathbb{L}_S(G)} \sum_{\mathcal{E} \in E} \mathbb{E}_{X_\mathcal{E} \sim \overline{\mu}_\mathcal{E}^{(U)}} \left[\overline{\theta}_\mathcal{E}^{(U)}(X_\mathcal{E})\right] \\ &= \max_{\mu \in \mathbb{L}_S(G)} \sum_{\mathcal{E} \in E} \mathbb{E}_{X_\mathcal{E} \sim \mu_\mathcal{E}} \left[\overline{\theta}_\mathcal{E}^{(U)}(X_\mathcal{E})\right], \end{aligned} \tag{1.18}$$

where to go from the penultimate line to the final line, we have used statement (i) of the proposition. Now, since flipped models clearly have the same maximal score, the result follows. □

When we wish to flip all variables in a configuration, we will simply use the bar notation without a superscript, to avoid cluttering the notation. Thus, for $x_V \in \{0,1\}^V$, $\overline{x}_V$ means the same as $\overline{x}_V^{(V)}$, and for $\mu_U \in \mathscr{P}(\{0,1\}^U)$, $\overline{\mu}_U$ means the same as $\overline{\mu}_U^{(U)}$. There is also a group-theoretic interpretation of flipping transformations — we defer discussion of this to Section 4.3. We conclude this section by noting a connection between flipping, balanced models, and signed MTP_2 distributions described in Section 1.3.4. Given a balanced binary graphical model $M = (G = (V,E), (\theta_\mathcal{E} | \mathcal{E} \in E))$, we may partition V into two disjoint sets V_0 and V_1, such that all edge potentials between two vertices both in V_0 or both in V_1 are attractive, and all edge potentials between one vertex in V_0 and one vertex in V_1 are repulsive, by virtue of there being no frustrated cycles in the balanced model. If we now consider the model obtained by flipping the variable set V_0, we see from Definition 1.12 that all pairwise potentials in the model become attractive, and hence the flipped model is attractive. Hence balanced models are precisely those that may be obtained from attractive models via flipping. Since it was shown in Section 1.3.4 that the notion of MTP_2 and attractiveness for binary pairwise graphical models are equivalent once redundant pairwise potentials have been replaced with equivalent singleton potentials, the argument above demonstrates a similar correspondence between signed MTP_2 distributions and balanced binary pairwise graphical models.

1.4.2 Overcomplete and minimal representations

The perspective of the marginal polytope $\mathbb{M}(G)$ and Sherali–Adams polytopes $\mathbb{L}_r(G)$ as abstract convex sets consisting of collections of measures, as in the exposition of Section 1.3, is useful in its generality. However, from the point of view of implementation (and often also analysis) it is useful to be able to view these polytopes explicitly as subsets of

some Euclidean space. This requires a choice of parametrisation, and can be achieved in several ways.

An *overcomplete representation* of the space $\mathscr{P}(\{0,1\}^A)$ of probability measures is an injective affine map $\mathscr{P}(\{0,1\}^A) \to \mathbb{R}^{2^A}$. A common choice of overcomplete representation is the *standard overcomplete representation* (Wainwright and Jordan, 2008), given by

$$\mathscr{P}(\{0,1\}^A) \ni \mu_A \mapsto (\mu_A(x_A) | x_A \in \{0,1\}^A) \in \mathbb{R}^{2^A}. \tag{1.19}$$

In words, the probability measure μ_A is mapped to a coordinate vector in $\mathbb{R}^{2^A}$, with coordinate values given by the probability of each possible configuration in $\{0,1\}^A$.

Given an overcomplete representation of $\mathscr{P}(\{0,1\}^{\mathcal{E}})$ for each $\mathcal{E} \in E$, an overcomplete representation for a polytope of (pseudo)marginals is naturally given by concatenating the individual representations together. For example, the *standard overcomplete representation* of a Sherali–Adams polytope is given by the following map:

$$\mathbb{L}_r(G) \ni (\mu_{\mathcal{E}} | \mathcal{E} \in E) \mapsto \left(\left(\mu_{\mathcal{E}}(x_{\mathcal{E}}) | x_{\mathcal{E}} \in \{0,1\}^{\mathcal{E}} \right) \middle| \mathcal{E} \in E \right) \in \mathbb{R}^{\sum_{\mathcal{E} \in E} 2^{|\mathcal{E}|}}. \tag{1.20}$$

The term *overcomplete* in this context refers to the fact this parametrisation contains redundant information; for example, given the values of $\mu_{\mathcal{E}}(x_{\mathcal{E}})$ for all but one of the values of $x_{\mathcal{E}}$ in $\{0,1\}^{\mathcal{E}}$, the final value is fully determined by the normalisation condition on $\mu_{\mathcal{E}}$. Further, in an overcomplete representation of a Sherali–Adams relaxation, given two hyperedges $\mathcal{E}_1, \mathcal{E}_2 \in E$ with non-empty intersection $\mathcal{E}_1 \cap \mathcal{E}_2$, the representations of $\mu_{\mathcal{E}_1}$ and $\mu_{\mathcal{E}_2}$ will each contain (duplicated) information about the measure $\mu_{\mathcal{E}_1 \cap \mathcal{E}_2}$.

In contrast, a *minimal representation* of a Sherali–Adams polytope $\mathbb{L}_r(G)$ is an injective affine map into a Euclidean space having *full rank*; equivalently, there are no non-trivial affine equations satisfied by all points in the polytope. An example of a minimal representation in the particular case of binary graphical models (where $\mathcal{X}_v = \{0,1\}$ for all $v \in V$), which we will use in Chapter 3, is the *moment representation* (Wainwright and Jordan, 2008), given by

$$\mathbb{L}_r(G) \ni (\mu_{\mathcal{E}} | \mathcal{E} \in E) \mapsto \left(\mathbb{E}_{X_U \sim \mu_U} \left[\prod_{i \in U} X_i \right] \middle| \emptyset \neq U \subset V \text{ such that } \exists \mathcal{E} \in E \text{ with } U \subseteq \mathcal{E} \right). \tag{1.21}$$

Note also that since we are dealing with binary variables, we have $\mathbb{E}_{X_U \sim \mu_U}[\prod_{i \in U} X_i] = \mathbb{P}_{X_U \sim \mu_U}(X_i = 1 \ \forall i \in U)$.

Given a representation of a Sherali–Adams polytope $\mathbb{L}_r(G)$, there is a corresponding affine representation of potentials over G, so that the score of a particular set of pseudomarginals for a particular model can be evaluated as an inner product in Euclidean space. For example, the representation of potentials corresponding to the standard overcomplete representation in Expression (1.20) is given by

$$(\theta_{\mathcal{E}} | \mathcal{E} \in E) \mapsto ((\theta_{\mathcal{E}}(x_{\mathcal{E}}) | x_{\mathcal{E}} \in \{0,1\}^{\mathcal{E}}) | \mathcal{E} \in E), \tag{1.22}$$

so that the score $\sum_{\mathcal{E} \in E} \mathbb{E}_{X_{\mathcal{E}} \sim \mu_{\mathcal{E}}}[\theta_{\mathcal{E}}(X_{\mathcal{E}})]$ is given by the inner product:

$$\left\langle \left(\left(\theta_{\mathcal{E}}(x_{\mathcal{E}}) | x_{\mathcal{E}} \in \{0,1\}^{\mathcal{E}} \right) | \mathcal{E} \in E \right), \left(\left(\mu_{\mathcal{E}}(x_{\mathcal{E}}) | x_{\mathcal{E}} \in \{0,1\}^{\mathcal{E}} \right) | \mathcal{E} \in E \right) \right\rangle . \tag{1.23}$$

1.5 Outline of original work

Having established the relevant background material, we now give a brief outline of our original contributions in this first part of the thesis. A more detailed list of contributions may be found in each of Chapters 2 to 4.

- In Chapter 2, we study *rerooting and uprooting*, a family of transformations of graphical models that may be used to improve the performance of a variety of approximate inference algorithms. A principal theoretical contribution in this chapter is to study the effect of these transformations on the Sherali–Adams hierarchy.
- In Chapter 3, our principal contribution is to generalise Theorem 1.14 to the class of *almost balanced models*, resulting in a new tightness guarantee for a generalised Sherali–Adams polytope (see Definition 1.15). In the proof, we make heavy use of the moment representation described in Section 1.4.2.
- In Chapter 4, we examine the extent to which the treewidth conditions for tightness in Theorem 1.11 can be extended, via graph minor theory and a geometric approach. In doing so, we derive new proofs of Theorems 1.8 and 1.14.

Chapter 2

Rerooting Models and Polytopes

This chapter is based on the following publication:

- Rowland, M.* and Weller, A.* (2017). Uprooting and rerooting higher-order graphical models. In *Neural Information Processing Systems (NIPS)*. [*=equal contribution].

Theoretical work was carried out jointly with Adrian Weller. Experiments were implemented by the author of the thesis, and designed jointly with Adrian Weller. The writing of the original paper was carried out jointly with Adrian Weller, although many parts have been rewritten for the purposes of this thesis.

2.1 Introduction

The focus of this chapter is a family of transformations of graphical models termed *uprootings* and *rerootings*, generalising transforms specifically for binary pairwise models described by Weller (2016b). These transformations have the property that the results of exact inference are easily passed between each of the transformed models, but that approximate inference algorithms may have very different performance. Thus, if one is interested in performing approximate inference on a graphical model, there may be value in first applying a rerooting transformation, performing approximate inference on this transformed model, and translating the results back to the original model. This workflow is illustrated schematically in Figure 2.1.

Weller (2016b) used the notion of rerooting to show that the triplet-consistent polytope, $\mathbb{L}_3(G)$, has the remarkable property of being *universally rooted* (see Section 2.4 for defini-

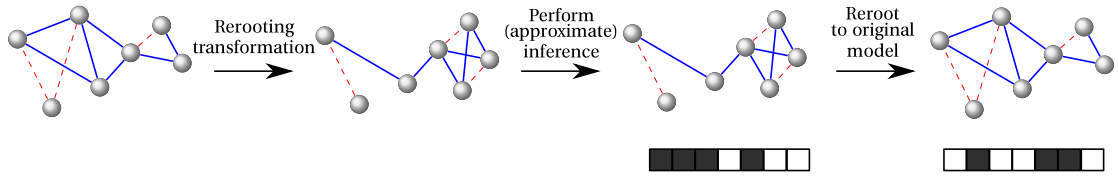


Figure 2.1. Schematic illustration of approximate inference in combination with rerooting transformations. The original model (far left) is first transformed into a rerooted model (centre-left). Approximate MAP inference is then performed on this rerooted model, with computed configuration shown below the model (centre-right), and the results are then translated back to the original model (far right).

tions). It was natural to conjecture that all polytopes $\mathbb{L}_r(G)$ with $r \geq 3$ shared this property, but one of our principal theoretical contributions in this chapter is to show that in fact this is not the case, and that $\mathbb{L}_3(G)$ is unique in being universally rooted. This has important consequences for using Sherali–Adams polytopes for MAP inference in combination with rerooting transformations.

We study these transformations from a theoretical perspective, with particular emphasis on the effects on Sherali–Adams polytopes, and evaluate the effectiveness of the transformations empirically through a variety of approximate inference algorithms. Specifically, we highlight the following contributions:

- In Section 2.2: A description of uprooting and rerooting for general binary graphical models, and equivalence of inference tasks in these models.
- In Section 2.3: The introduction of pure k-potentials, a particular parametrisation of binary graphical models naturally linked to rerooting transformations. These may be of independent interest, and lead to a concise way of computing uprooting and rerooting transformations in general.
- In Section 2.4: A framework for understanding the effects of rerooting transformations on Sherali–Adams polytopes, and in particular, a solution to a question left open by Weller (2016b) in the negative (see Theorem 2.20).
- In Section 2.5: Two heuristics for selecting a variable at which to reroot, and empirical evaluation of these heuristics on MAP inference and marginal inference tasks for a variety of graphical models, demonstrating that significant accuracy gains are sometimes attainable at negligible computational cost.

2.2 Uprooting and rerooting

We begin with precise descriptions of the uprooting and rerooting transformations mentioned above. First, we recall the notion of clamping a variable, a technique often used in performing approximate inference (Eaton and Ghahramani, 2009; Weller and Domke, 2016; Weller and Jebara, 2014).

Definition 2.1 (**Clamping**)**.** Consider a graphical model M given by hypergraph $G = (V, E)$, finite domains $(\mathcal{X}_v | v \in V)$, and potentials $(\theta_\mathcal{E} | \mathcal{E} \in E)$. For a specified vertex $i \in V$, the model $M|_{X_i=a}$ obtained by clamping the variable X_i to the value $a \in \mathcal{X}_i$ is given by the hypergraph (V_i, E_i), where $V_i = V \setminus \{i\}$, $E_i = \{\mathcal{E}_i | \mathcal{E} \in E\}$, and $\mathcal{E}_i = \mathcal{E} \setminus \{i\}$, domains $(\mathcal{X}_v | v \in V_i)$, and potentials $(\theta^{(i)}_{\mathcal{E}_i} | \mathcal{E}_i \in E_i)$, with

$$\theta^{(i)}_{\mathcal{E}_i}(x_{\mathcal{E}_i}) = \begin{cases} \theta_\mathcal{E}(x_\mathcal{E}) & \text{if } i \notin \mathcal{E} \\ \theta_\mathcal{E}(x_{\mathcal{E}_i}, x_i = a) & \text{otherwise .} \end{cases} \tag{2.1}$$

Thus, clamping eliminates a variable from the model, and adjusts the model's potentials over remaining variables correspondingly; indeed, the clamped graphical model represents the conditional distribution of the remaining variables given the event $\{X_i = a\}$. We now specialise to binary graphical models, where $\mathcal{X}_v = \{0, 1\}$ for all $v \in V$. The key to the notion of *uprooting* is to *reverse* this perspective, and view the initial graphical model M as one which has already been clamped. In other words, we consider an augmented graphical model M^+, with an additional random variable, which we denote X_0, with the property that M can be obtained from M^+ by clamping X_0 to some value (without loss of generality, to the value 0); and we may thus write $M = M^+|_{X_0=0}$. Weller (2016b) considered this task algebraically specifically for binary pairwise models, based on an earlier construction that reduces MAP inference on pairwise models to the MAXCUT problem on the suspension graph (Barahona et al., 1988); in contrast, we treat general binary models here. There is considerable freedom in how the potentials of the model M^+ might be specified in terms of those of M, but a key requirement if the transformation is to be of use for approximate inference is that the properties of interest for the model M, such as its partition function, marginal distributions, and MAP configuration, be readily computable from those of M^+. This consideration leads us to the following definition of the uprooting of a binary graphical model, which generalises the definition of Weller (2016b).

Definition 2.2 (**Uprooting, suspension hypergraph**)**.** Given a model M defined by hypergraph $G = (V, E)$ and potentials $(\theta_\mathcal{E} | \mathcal{E} \in E)$, the *uprooted* model M^+ is defined by the

hypergraph ∇G (the suspension hypergraph of G, which has vertex set $V^+ = V \cup \{0\}$ and hyperedge set $E^+ = \{\mathcal{E}^+ = \mathcal{E} \cup \{0\} | \mathcal{E} \in E\}$), and potentials $(\theta^+_{\mathcal{E}^+} | \mathcal{E}^+ \in E^+)$ over the hyperedges of ∇G given by

$$\theta^+_{\mathcal{E}^+}(x_{\mathcal{E}^+}) = \begin{cases} \theta_{\mathcal{E}}(x_{\mathcal{E}}) & \text{if } x_0 = 0 \\ \theta_{\mathcal{E}}(\overline{x}_{\mathcal{E}}) & \text{if } x_0 = 1\,. \end{cases} \tag{2.2}$$

In what follows, we use the notation p^+, Z^+ and score_+ to denote the probability mass function, partition function, and score function, respectively, for the model M^+, to distinguish from those of the original model M. Thus, we have

$$p^+(x_{V^+}) = \frac{1}{Z^+} \exp\left(\text{score}_+(x_{V^+})\right) \qquad \text{for all } x_{V^+} \in \{0,1\}^{V^+}. \tag{2.3}$$

We note the following connection to the notion of flipping discussed in Section 1.4.1: all uprooted potentials are symmetric, in the sense that $\theta^+_{\mathcal{E}^+}(x_{\mathcal{E}^+}) = \theta^+_{\mathcal{E}^+}(\overline{x}_{\mathcal{E}^+})$ for all $x_{\mathcal{E}^+} \in \{0,1\}^{\mathcal{E}^+}$, and for each $\mathcal{E}^+ \in E^+$. Thus, flipping all variables in V^+ leaves the model M^+ invariant. It also follows straightforwardly from this observation that we have the following correspondence between configurations in M and M^+.

Proposition 2.3. We have

$$\text{score}(x_V) = \text{score}_+(x_V, x_0 = 0) = \text{score}_+(\overline{x}_V, x_0 = 1)\,, \tag{2.4}$$

for all $x_V \in \{0,1\}^V$. Thus, for every configuration of the original model M, there are two corresponding configurations in M^+ with the same score, which are related by flipping all variables.

Definition 2.4 (**Score-preserving mapping.**)**.** We refer to the correspondence

$$\{0,1\}^V \ni x_V \longleftrightarrow \left\{ \begin{matrix} (x_V, x_0 = 0) \\ (\overline{x}_V, x_0 = 1) \end{matrix} \right\} \subseteq \{0,1\}^{V^+} \tag{2.5}$$

described in Proposition 2.3 as *score-preserving*, meaning that given a configuration $x_V \in \{0,1\}^V$ for a particular model M, if we pass to the uprooted model M^+, then the configurations given by the correspondence in Expression (2.5) attain the same score for the uprooted model M^+ as the original configuration does for the original model M.

The distribution on M^+ is thus said to be a *palindromic* Bernoulli distribution; Marchetti and Wermuth (2016) recently introduced and studied such distributions in the context

of maximum-likelihood estimation, highlighting a particular application to structure learning via median-dichotomisation of multivariate Gaussian random variables.

Building on Proposition 2.3, the result below reveals the close correspondence between inference on the original model M and the uprooted model M^+. In analogy with the notation p^+ for the probability mass function associated with the model M^+, we use the notation p the probability mass function associated with the model M.

Proposition 2.5. Given a graphical model M and its uprooted model M^+, the following results hold:

- The partition function Z of M is related to the partition function Z^+ of M^+ according to $Z^+ = 2Z$.
- The configuration $x_V \in \{0,1\}^V$ is a MAP configuration for M iff $(x_V, x_0 = 0)$ is a MAP configuration for M^+ iff $(\overline{x}_V, x_0 = 1)$ is a MAP configuration for M^+. The MAP score is the same in all cases.
- The marginal distribution over a subset $U \subseteq V$ of variables in M may be expressed in terms of marginals of variables in M^+ as follows:

$$p(x_U) = p^+(x_U, x_0 = 0) + p^+(\overline{x}_U, x_0 = 1) = 2p^+(x_U, x_0 = 0) = 2p^+(\overline{x}_U, x_0 = 1)\,, \tag{2.6}$$

  for all $x_U \in \{0,1\}^U$. Similarly, the marginal distribution over a subset $U \subseteq V^+$ of variables in M^+ may be expressed in terms of marginals of variables in M as follows:

  - If $0 \notin U$, then

$$p^+(x_U) = \frac{1}{2}p(x_U) + \frac{1}{2}p(\overline{x}_U)\,, \tag{2.7}$$

    for all $x_U \in \{0,1\}^U$.

  - If $0 \in U$, then writing $U_0 = U \setminus \{0\}$, we have

$$p^+(x_{U_0}, x_0 = 0) = \frac{1}{2}p(x_{U_0})\,, \qquad p^+(x_{U_0}, x_0 = 1) = \frac{1}{2}p(\overline{x}_{U_0})\,, \tag{2.8}$$

    for all $x_{U_0} \in \{0,1\}^{U_0}$.

Proposition 2.5 shows that the results of exact inference for either of the two models M or M^+ can be trivially computed from the results of exact inference in the other model. The key insight that leads to the notion of rerooting is to see that the only properties required to establish this correspondence of inference tasks are that M is obtainable as a clamping

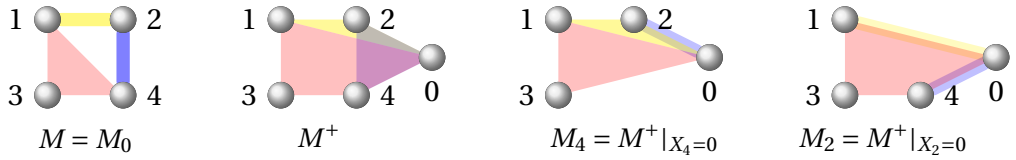


Figure 2.2. Left: The hypergraph G of a graphical model M over 4 variables, with potentials on the hyperedges $\{1,2\}$, $\{1,3,4\}$, and $\{2,4\}$. Centre-left: The suspension hypergraph ∇G of the uprooted model M^+. Centre-right: The hypergraph $\nabla G \setminus \{4\}$ of the rerooted model $M_4 = M^+|_{X_4=0}$, i.e. M^+ with X_4 clamped to 0. Right: The hypergraph $\nabla G \setminus \{2\}$ of the rerooted model $M_2 = M^+|_{X_2=0}$, i.e. M^+ with X_2 clamped to 0.

of M^+, and that the potentials of M^+ are invariant under flipping the underlying input configuration. This reasoning shows that if we clamp *any* variable of M^+, the resulting clamped model will have a similar correspondence to M^+ as that described between M and M^+ in Proposition 2.5. This motivates the notion of rerooting, defined formally below.

Definition 2.6 (**Rerooting**)**.** Given a graphical model M, if we uproot to M^+ then clamp $X_0 = 0$, we recover the original model M. If instead in M^+ we clamp $X_i = 0$ for any $i \in V$, then we obtain the *rerooted* model $M_i := M^+|_{X_i=0}$. The rerooted model M_i has hypergraph with vertex set $V_i^+ = V \cup \{0\} \setminus \{i\}$ and edge set $E_i^+ = \{\mathcal{E}_i^+ = \mathcal{E} \cup \{0\} \setminus \{i\} | \mathcal{E} \in E\}$. In what follows, we use the notation p_i, Z_i and score_i to denote the probability mass function, partition function, and score function, respectively, for the model M_i, to distinguish from those of the original model M. Thus, we have

$$p_i(x_{V_i^+}) = \frac{1}{Z_i} \exp\left(\text{score}_i(x_{V_i^+})\right) \qquad \text{for all } x_{V_i^+} \in \{0,1\}^{V_i^+}. \tag{2.9}$$

See Figure 2.2 for an illustration of uprooting and rerooting transformations on a small graphical model. We record the exact correspondence between results of exact inference for a model M and a rerooting M_i ($i \in V$) in the following proposition. Intuitively, this result may be seen as arising from a score-preserving correspondence between configurations of M and M_i, in analogy with the score-preserving correspondence between M and M^+ discussed around Proposition 2.3.

Proposition 2.7. Given a graphical model M and a rerooting M_i, for some $i \in V$, the following results hold.

- The partition functions of the models M and M_i are equal.
- The configuration $(x_{V\setminus\{i\}}, x_i = 0) \in \{0,1\}^V$ is a MAP configuration for M iff $(x_{V\setminus\{i\}}, x_0 = 0) \in \{0,1\}^{V_i^+}$ is a MAP configuration for M_i. The configuration $(x_{V\setminus\{i\}}, x_i = 1) \in \{0,1\}^V$ is a MAP configuration for M iff $(\overline{x}_{V\setminus\{i\}}, x_0 = 1) \in \{0,1\}^{V_i^+}$ is a MAP configuration for M_i.

- The marginal distribution over a subset $U \subseteq V$ of variables in M may be expressed in terms of marginals of variables in M_i as follows.
  - If $i \in V \setminus U$,

$$p(x_U) = p_i(x_0 = 0, x_U) + p_i(x_0 = 1, \overline{x}_U)\,, \tag{2.10}$$

    for all $x_U \in \{0,1\}^U$.
  - If $i \in U$,

$$p(x_U) = \begin{cases} p_i(x_0 = 0, x_{U\setminus\{i\}}) & \text{if } x_i = 0 \\ p_i(x_0 = 1, \overline{x}_{U\setminus\{i\}}) & \text{if } x_i = 1\,, \end{cases} \tag{2.11}$$

    for all $x_U \in \{0,1\}^U$.

Thus, exact results of inference on any M_i may be efficiently translated back to exact results for M. In practice, a crucial question is how to choose which rerooted model to perform inference on; the aim is to select a rerooted model for which our approximate inference algorithm of choice performs well. We explore the question of how to choose a good variable for rerooting (i.e. how to choose a good variable to clamp in M^+) in Section 2.5. We first study a particular parametrisation of potentials that will be useful in the study and implementation of rerooting in Section 2.3, and then discuss the relationship between rerooting and the Sherali–Adams hierarchy of relaxations of the marginal polytope in Section 2.4.

2.3 Pure k-potentials

In this section, we introduce the notion of *pure k-potentials*. These allow the specification of interactions which act 'purely' over a set of variables of a given size k, without influencing the distribution of any proper subsets of variables. We show that in fact, there is essentially only one pure k-potential. Further, we show that one can express any $\theta_{\mathcal{E}}$ potential in terms of pure potentials over $\mathcal{E}$ and subsets of $\mathcal{E}$, and that pure potentials have appealing properties when uprooted and rerooted, which help our subsequent analysis.

We say that a potential is a *k-potential* if k is the smallest number such that the score of the potential may be determined by considering the configuration of k variables. Usually a potential $\theta_{\mathcal{E}}$ is a k-potential with $k = |\mathcal{E}|$. For example, typically a singleton potential is a

1-potential, and an edge potential is a 2-potential. However, note that $k < |\mathcal{E}|$ is possible if one or more variables in $\mathcal{E}$ are not needed to establish the score (a simple example is $\theta_{12}(x_1, x_2) = x_1$, which clearly is a 1-potential).

In general, a k-potential will affect the marginal distributions of all subsets of the k variables. For example, one popular form of 2-potential is $\theta_{ij}(x_i, x_j) = \phi_{ij} x_i x_j$; when $\phi_{ij} > 0$, it tends to pull X_i and X_j toward the same value, but also tends to increase each of the marginal probabilities $\mathbb{P}(X_i = 1)$ and $\mathbb{P}(X_j = 1)$. For pairwise models, for example, a different reparametrisation of potentials instead writes the score as

$$\text{score}(x_V) = \sum_{i \in V} \phi_i x_i + \frac{1}{2} \sum_{ij \in E} W_{ij} \mathbb{1}[x_i = x_j]\,. \tag{2.12}$$

Expression (2.12) has the desirable feature that in the absence of any other potentials, the $\theta_{ij}(x_i, x_j) = \frac{1}{2} W_{ij} \mathbb{1}[x_i = x_j]$ edge potentials affect only the pairwise marginals, without disturbing singleton marginals; from the point of view of modelling, the parameters W_{ij} are arguably now more interpretable, as their effect is less confounded with that of the parameters ϕ_i, ϕ_j. This motivates the following definition.

Definition 2.8 (**Pure k-potentials**)**.** Let $k \geq 2$, and let U be a set of size k. We say that a k-potential $\theta_U : \{0,1\}^U \to \mathbb{R}$ is a *pure k-potential* if the distribution induced by the potential, $p(x_U) \propto \exp(\theta_U(x_U))$, has the property that for any $\emptyset \neq W \subsetneq U$, the marginal distribution $p(x_W)$ is uniform.

We shall see in Proposition 2.10 that a pure k-potential must essentially be an *even k-potential*, which we define formally below.

Definition 2.9 (**Even k-potentials**)**.** Let $k \geq 1$, and $|U| = k$. An *even k-potential* is a k-potential $\theta_U : \{0,1\}^U \to \mathbb{R}$ of the form $\theta_U(x_U) = a \mathbb{1}[\,|\{i \in U | x_i = 1\}| \text{ is even}]$, for some $a \in \mathbb{R} \setminus \{0\}$ which is its *coefficient*. In other words, $\theta_U(x_U)$ takes value a if x_U has an even number of 1s, else it takes value 0.

As an example, the 2-potential $\theta_{ij}(x_i, x_j) = \frac{1}{2} W_{ij} \mathbb{1}[x_i = x_j]$ in Expression (2.12) is an even 2-potential with $U = \{i, j\}$ and coefficient $W_{ij}/2$. We now provide the following characterisation of pure potentials in terms of even potentials.

Proposition 2.10 (**All pure potentials are essentially even potentials**)**.** Let $k \geq 2$, and $|U| = k$. If $\theta_U : \{0,1\}^U \to \mathbb{R}$ is a pure k-potential then θ_U must be an affine function of the even k-potential, i.e. there exist $a \in \mathbb{R} \setminus \{0\}, b \in \mathbb{R}$ such that $\theta_U(x_U) = a \mathbb{1}[\,|\{i \in U | x_i = 1\}| \text{ is even}] + b$.

Note that any constant b in Proposition 2.10 will be absorbed into the partition function Z, and does not affect the probability distribution (see Expression (1.1)) associated with the graphical model. The next proposition establishes the expressive power of even potentials.

Proposition 2.11 (**Even k-potentials form a basis**)**.** For a finite set U, the set of even k-potentials $\left(\mathbb{1}[\,|\{i \in W | x_i = 1\}| \text{ is even}]\right)_{W \subseteq U}$, indexed by subsets $W \subseteq U$, forms a basis for the vector space of all potential functions $\theta : \{0,1\}^U \to \mathbb{R}$.

Thus, the results collected above show that any binary graphical model may always be expressed in terms of pure potentials. The final result of this section illustrates why thinking in terms of pure potentials is particularly useful for uprooting and rerooting transformations; we have the following direct characterisation of the effect of uprooting.

Lemma 2.12 (**Uprooting an even k-potential**)**.** When an even k-potential $\theta_{\mathcal{E}} : \{0,1\}^U \to \mathbb{R}$ with $|\mathcal{E}| = k$ is uprooted:

- if k is an even number, then the uprooted potential is exactly the same even k-potential. That is, $\theta_{\mathcal{E}^+}(x_{\mathcal{E}^+}) = \theta_{\mathcal{E}}(x_{\mathcal{E}})$;
- if k is odd, then we obtain the even $(k+1)$-potential over $\mathcal{E} \cup \{0\}$ with the same coefficient as the original $\theta_{\mathcal{E}}$.

Notice that although according to Definition 2.2 a k-potential $\theta_{\mathcal{E}}$ is uprooted to become a $(k+1)$-potential in general, Lemma 2.12 states that if $\theta_{\mathcal{E}}$ is an even k-potential with k an even number, then the uprooted potential $\theta_{\mathcal{E}^+}$ is in fact a pure k-potential, and so does not depend on the configuration of the additional variable x_0 in the uprooted model. Thus the arity of the uprooted potential does not increase in this case. For completeness, we also give the corresponding result for *clamping* even potentials; taken together with Lemma 2.12, this characterises the effect of rerooting even k-potentials too.

Lemma 2.13 (**Clamping an even k-potential**)**.** When an even k-potential $\theta_{\mathcal{E}} : \{0,1\}^U \to \mathbb{R}$ with $|\mathcal{E}| = k$ is clamped at $i \in V^+$:

- if $i \notin \mathcal{E}$, then the clamped potential is exactly the same even k-potential.
- if $i \in \mathcal{E}$, then the clamped potential is the even potential over $\mathcal{E} \setminus \{i\}$ with the same coefficient as the original $\theta_{\mathcal{E}}$.

These are important observations from the point of view of implementation, since they provide an automatic mechanism for keeping track of the true arity of potentials, which can lead to vastly increased efficiency when using approximate inference packages that do not necessarily automatically check for such redundancies in potential functions. As

a concrete example, a computer vision model may employ 4-ary potentials (Kohli and Rother, 2012), and since 4 is even, the remarks above guarantee that these 4-ary potentials *remain* 4-ary potentials in the corresponding uprooted model, and do not become 5-ary potentials (which may complicate inference significantly). Indeed, together with Proposition 2.11, this suggests an efficient way of performing uprooting and rerooting transformations in practice, as well as a useful perspective for more theoretical questions: express the potentials of a graphical model in terms of even k-potentials, and then read off the uprooted/rerooted potentials using Lemmas 2.12 and 2.13.

We conclude this section by remarking that even k-potentials may be viewed as log-linear parametrisations of binary graphical models, up to additive constants. Indeed, given a graphical model parametrised in the following ways:

$$p(x_V) = \exp\left(\sum_{x'_V \in \{0,1\}^V} \log p(x'_V)\, \mathbb{1}[x_V = x'_V]\right) = \frac{1}{Z} \exp\left(\sum_{U \subseteq V} \theta_U \mathbb{1}[\,|\{i \in U | x_i = 1\}| \text{ is even}]\right),$$

there is a particular affine transformation relating the log-probabilities $(\log p(x_V) | x_V \in \{0,1\}^V)$ and the coefficients $(\theta_U | U \subseteq V)$ of the even k-potentials (Marchetti and Wermuth, 2016). Such contrast-based log-linear parametrisations arise frequently in the statistics literature, for example in the modelling of count data (Haberman, 1973).

2.4 The effect of rerooting on Sherali–Adams relaxations

We saw in Proposition 2.3 that there is a score-preserving one-to-two correspondence between configurations of M and M^+, and a bijection between configurations of M and any M_i. Here we examine the extent to which these score-preserving mappings extend to (pseudo)marginal probability distributions over variables by considering the Sherali–Adams relaxations of the respective marginal polytopes, as described in Section 1.3; this leads to insights in understanding the effects of uprooting and rerooting on approximate inference algorithms making use of Sherali–Adams relaxations.

2.4.1 Symmetrised polytopes

We introduce two variants of the Sherali–Adams polytopes which will be helpful in analysing uprooted and rerooted models. We start with a preliminary definition. Recall from Section 1.4.1 that given a measure $\mu_U \in \mathscr{P}(\{0,1\}^U)$, the flipped measure

$\overline{\mu}_U \in \mathscr{P}(\{0,1\}^U)$ is given by $\overline{\mu}_U(x_U) = \mu_U(\overline{x}_U)$ for all $x_U \in \{0,1\}^U$, where $\overline{x}_U = (1 - x_i | i \in U)$ is the flipping of the configuration x_U.

Definition 2.14 (**Flipping invariance**)**.** Given a probability measure $\mu_U \in \mathscr{P}(\{0,1\}^U)$, we say that μ_U is *flipping-invariant* if $\mu_U = \overline{\mu}_U$.

We may now proceed with the definition of our two variants of the Sherali–Adams polytopes.

Definition 2.15 (**Symmetrised Sherali–Adams polytopes**)**.** The symmetrised Sherali–Adams polytopes for an uprooted hypergraph $\nabla G = (V^+, E^+)$ (as given in Definition 2.2) are given by:

$$\widetilde{\mathbb{L}}_S(\nabla G) = \left\{ (\mu_{\mathcal{E}} | \mathcal{E} \in E^+) \in \mathbb{L}_S(\nabla G) \middle| \mu_{\mathcal{E}} = \overline{\mu}_{\mathcal{E}} \ \forall \mathcal{E} \in E^+ \right\}, \tag{2.13}$$

for all antichains S covering ∇G, as in Definition 1.15.

Definition 2.16 (**Uprooted symmetrised Sherali–Adams polytopes**)**.** For any $i \in V^+$, and any integer $r \geq 2$ such that $\max_{\mathcal{E} \in E} |\mathcal{E}| \leq r \leq |V^+|$, we define the symmetrised Sherali–Adams polytope of order r *uprooted at* i, denoted by $\widetilde{\mathbb{L}}^i_{r+1}(\nabla G)$, to be the symmetrised Sherali–Adams polytope $\widetilde{\mathbb{L}}_S(\nabla G)$, where S is the antichain given by all subsets of V^+ of size $r+1$ containing i, that is:

$$S = \{U \subseteq V^+ \mid |U| = r+1, \ \ i \in U\}. \tag{2.14}$$

Thus, for each collection of measures over hyperedges in $\widetilde{\mathbb{L}}^i_{r+1}(\nabla G)$, there exist corresponding flipping-invariant, locally consistent measures on sets of size $r+1$ which contain i. Note that for any hypergraph $G = (V, E)$ and any $i \in V^+$, we have $\widetilde{\mathbb{L}}_{r+1}(\nabla G) \subseteq \widetilde{\mathbb{L}}^i_{r+1}(\nabla G) \subseteq \widetilde{\mathbb{L}}_r(\nabla G)$. We next extend the correspondence of Proposition 2.3 to collections of locally consistent probability distributions on the hyperedges of G, using the symmetrised polytopes defined in Definitions 2.15 and 2.16.

Theorem 2.17. For a hypergraph $G = (V, E)$, and integer r such that $\max_{\mathcal{E} \in E} |\mathcal{E}| \leq r \leq |V|$, there is an affine score-preserving bijection

$$\mathbb{L}_r(G) \underset{\textbf{RootAt0}}{\overset{\textbf{Uproot}}{\rightleftarrows}} \widetilde{\mathbb{L}}^0_{r+1}(\nabla G). \tag{2.15}$$

Theorem 2.17 establishes the following diagram of polytope inclusions and affine bijections:

$$\begin{array}{lccccc} \text{For } M = M_0: & \mathbb{L}_{r+1}(G) & \subseteq & \text{Unnamed} & \subseteq & \mathbb{L}_r(G) \\ & \Big\downarrow \text{Uproot} \;\; \Big\uparrow \text{RootAt0} & & \Big\downarrow \text{Uproot} \;\; \Big\uparrow \text{RootAt0} & & \Big\downarrow \text{Uproot} \;\; \Big\uparrow \text{RootAt0} \\ \text{For } M^+: & \widetilde{\mathbb{L}}^0_{r+2}(\nabla G) & \subseteq & \widetilde{\mathbb{L}}_{r+1}(\nabla G) & \subseteq & \widetilde{\mathbb{L}}^0_{r+1}(\nabla G)\,. \end{array} \tag{2.16}$$

A question of theoretical interest and practical importance is which of the inclusions in Diagram (2.16) are strict. Our perspective here generalises earlier work. Using different language, Deza and Laurent (1997) identified $\mathbb{L}_2(G)$ with $\widetilde{\mathbb{L}}^0_3(\nabla G)$, which was termed RMET, the *rooted semimetric polytope*; and $\widetilde{\mathbb{L}}_3(\nabla G)$ with MET, the *semimetric polytope*. Building on this, Weller (2016b) considered $\mathbb{L}_3(G)$, the triplet-consistent polytope, though only in the context of pairwise models, and showed that $\mathbb{L}_3(G)$ has the remarkable property that if it is used to optimise an LP for a model M on G, the exact same optimum is achieved for $\mathbb{L}_3(G_i)$ for any rerooting M_i. This leads to an important insight: if using $\mathbb{L}_3(G)$ for approximate MAP inference, then there is *no benefit* to rerooting. It was natural to conjecture that $\mathbb{L}_r(G)$ might have this same property for all $r > 3$, yet this was left as an open question.

2.4.2 $\mathbb{L}_3$ is unique in being universally rooted

We shall first strengthen the result of Weller (2016b) to show that $\mathbb{L}_3$ is *universally rooted* in the following stronger sense.

Definition 2.18. We say that the r^{th}-order Sherali–Adams relaxation is *universally rooted* (and write "$\mathbb{L}_r$ is universally rooted" for short) if for all hypergraphs $G = (V, E)$ with maximal hyperedge degree $\leq r$, the following conditions are satisfied:

- If r is even, then we have $\widetilde{\mathbb{L}}^0_{r+1}(\nabla G) = \widetilde{\mathbb{L}}^i_{r+1}(\nabla G)$ for all $i \in V$.
- If r is odd, then we have $\widetilde{\mathbb{L}}^0_{r+1}(\nabla G) = \widetilde{\mathbb{L}}^i_{r+1}(\nabla G)$ for all $i \in V$ such that all hyperedges in G of degree r include i.

The restriction on $i \in V$ in the case of r odd is precisely so that a model M defined on G can be rerooted at vertex i and written in terms of r-ary potentials, in light of the observations on rerootings of even potentials in Section 2.3.

If $\mathbb{L}_r$ is universally rooted, this applies for potentials on up to r variables (the maximum which makes sense in this context), and clearly it implies that optimising score over any rerooting (as in MAP inference) will attain the same objective.

Extending arguments used by Weller (2016b), we are able to demonstrate the following result.

Theorem 2.19. $\mathbb{L}_3$ is universally rooted.

We next provide a striking and rather surprising result.

Theorem 2.20. $\mathbb{L}_3$ is unique in being universally rooted. Specifically, for any integer $r > 1$ other than $r = 3$, we constructively demonstrate a hypergraph $G = (V, E)$ with $|V| = r + 1$ variables for which $\widetilde{\mathbb{L}}^0_{r+1}(\nabla G) \neq \widetilde{\mathbb{L}}^i_{r+1}(\nabla G)$ for any $i \in V$.

Theorem 2.20 shows that we may hope for benefits from rerooting for any inference method based on a Sherali–Adams relaxed polytope $\mathbb{L}_r$, unless $r = 3$.

2.5 Experiments

Here we show empirically the benefits of uprooting and rerooting for approximate inference methods in models with higher-order potentials. We introduce an efficient heuristic which can be used in practice to select a variable for rerooting, and demonstrate its effectiveness.

We compared performance after different rerootings of marginal inference (to guarantee convergence we used the double-loop method of Heskes et al. (2003), which relates to generalised belief propagation (Yedidia et al., 2005)) and MAP inference (using loopy belief propagation, LBP (Pearl, 1988)). For ground truth exact inference, we used the junction tree algorithm. All methods were implemented using libDAI (Mooij, 2010). We ran experiments on complete hypergraphs (with 8 variables) and toroidal grid models (5×5 variables). Potentials up to order 4 were selected randomly, by drawing even k-potential coefficients from Unif($[-W_{\max}, W_{\max}]$) distributions for a variety of $W_{\max}$ parameters. For each regime of maximum potential values, we plot results averaged over 20 runs. We display average error of the inference method applied to: the original model M; the uprooted model M^+; then rerootings at: the *worst* variable, the *best* variable, the K *heuristic* variable, and the G *heuristic* variable (see Section 2.5.1 for an explanation of these heuristics). *Best* and *worst* always refer to the variable at which rerooting gave,

with hindsight, the best and worst error for the partition function (even in plots for other inference tasks).

2.5.1 Heuristics to pick a good variable for rerooting

From Definition 2.6, a rerooted model M_i is obtained by clamping the uprooted model M^+ at variable X_i. Hence, selecting a good variable for rerooting is exactly the choice of a good variable to clamp in M^+. Considering pairwise models, Weller (2016b) refined the *maxW* method (Weller and Domke, 2016; Weller and Jebara, 2014) to introduce the *maxtW* heuristic, and showed that it was very effective empirically. *maxtW* selects the variable X_i with the property that $i \in V^+$ maximises the objective

$$\sum_{j \in \mathcal{N}(i)} \tanh\left(\left|\frac{W_{ij}}{4}\right|\right), \tag{2.17}$$

where $\mathcal{N}(i)$ is the set of neighbours of i in the model graph, and W_{ij} is coefficient of the even 2-potential between i and j.

The intuition for *maxtW* is as follows. Approximate inference methods for pairwise models, such as the Bethe approximation, are exact for models with no cycles. If we could, we would like to 'break' tight cycles with strong edge weights, since these lead to errors in such methods of approximate inference. When a variable is clamped, it is effectively removed from the model. Hence, we would like to reroot at a variable that sits on many cycles with strong edge weights. Identifying such cycles is NP-hard (Sontag et al., 2012), but the *maxtW* heuristic attempts to do this by looking only locally around each variable. Further, the effect of a strong edge weight saturates (Weller and Jebara, 2014); a sufficiently strong edge weight W_{ij} effectively 'locks' its incident variables (either to the same value, or to opposite values, depending on the sign of W_{ij}), and this effect cannot be significantly increased even by an extremely strong edge. Hence the tanh function was introduced to the earlier *maxW* method, leading to the *maxtW* heuristic in Expression (2.17).

As observed in Section 2.3, if we express our model potentials in terms of pure k-potentials, then the uprooted model will only have pure k-potentials for various values of k which are even numbers. Intuitively, the higher the coefficients on these potentials, the more tightly connected the model is, leading to more challenging inference. Hence, a natural way to generalise the *maxtW* approach to handle higher-order potentials is to pick a variable X_i

in M^+ such that the index $i \in V^+$ maximises the following objective:

$$\text{clamp-heuristic-measure}(i) = \sum_{\mathcal{E} \ni i : |\mathcal{E}|=2} c_2 \tanh |t_2 a_{\mathcal{E}}| + \sum_{\mathcal{E} \ni i : |\mathcal{E}|=4} c_4 \tanh |t_4 a_{\mathcal{E}}|, \qquad (2.18)$$

where $a_{\mathcal{E}}$ is the coefficient (weight) of the relevant pure k-potential (see Definition 2.9) and the $\{c_2, t_2\}, \{c_4, t_4\}$ terms are constants for pure 2-potentials and for pure 4-potentials respectively. This approach extends to potentials of higher orders by adding similar further terms. Since our goal is to rank the measures for each $i \in V^+$, without loss of generality we take $c_2 = 1$. We fit the t_2, c_4 and t_4 constants to the data from our experimental runs. Our *K heuristic* was fit based only on performance for complete hypergraphs, whilst the *G heuristic* was fit only based on performance for grid models. To fit the heuristic, we used gradient-free optimisation. For the K heuristic, we generated a collection of graphical models on K_8, and constructed a fitness function over the remaining parameters t_2, c_4, t_4, given by the average ranking of the rerooting selected by the heuristic for $\log Z$ estimation across our collection of complete graphs. We then initialised the parameters $t_2 = 0, c_4 = 1, t_4 = 0$, and performed a local exploration of the parameter space dictated by a Gaussian random walk, updating our parameter settings when they led to an improvement in the value of the fitness function. The G heuristic was constructed similarly, instead using a collection of grids to define the fitness function.

The precise values of the fitted heuristics are given below:

$$\begin{aligned}
\text{K-heuristic-measure}(i) &= \sum_{\mathcal{E} \ni i : |\mathcal{E}|=2} \tanh |0.051 a_{\mathcal{E}}| + \sum_{\mathcal{E} \ni i : |\mathcal{E}|=4} 0.091 \tanh |1.482 a_{\mathcal{E}}|, \qquad (2.19)\\
\text{G-heuristic-measure}(i) &= \sum_{\mathcal{E} \ni i : |\mathcal{E}|=2} \tanh |0.019539 a_{\mathcal{E}}| + \sum_{\mathcal{E} \ni i : |\mathcal{E}|=4} 0.3788 \tanh |0.033997 a_{\mathcal{E}}|.
\end{aligned}$$

Recognising that the benefits of our heuristics appeared somewhat robust to exact parameter choice, when we extended analysis to 6-potentials in Section 2.B.4, we extended our K heuristic by eye (without fitting to any data, and before examining the results for higher-order models), and explore a variant on the G heuristic. We used the following measures:

$$\begin{aligned}
\text{K-heuristic-measure}(i) &= \sum_{\substack{\mathcal{E} \ni i \\ |\mathcal{E}|=2}} \tanh |0.2 a_{\mathcal{E}}| + \frac{1}{3} \sum_{\substack{\mathcal{E} \ni i \\ |\mathcal{E}|=4}} \tanh |1.2 a_{\mathcal{E}}| + \frac{1}{5} \sum_{\substack{\mathcal{E} \ni i \\ |\mathcal{E}|=6}} \tanh |3 a_{\mathcal{E}}|,\\
\text{G-heuristic-measure}(i) &= \sum_{\substack{\mathcal{E} \ni i \\ |\mathcal{E}|=4}} |a_{\mathcal{E}}|. \qquad (2.20)
\end{aligned}$$

2.5.2 Model structures and parameters used for libDAI

In this section we give further information about the model structures used in our experiments, as well as the methods of approximate inference used.

Complete graphs. For complete graph experiments, there is a pure k-potential for each subset of k variables, for $k = 1,2,3,4$.

Grids. All grids are square and toroidal. There is a 1-potential for each variable, and a 2-potential for each edge of the graph. There is a 3-potential for each possible "L-shaped" connected subgraph of size 3 (any of the four possible orientations), and a 4-potential for each cycle of size 4.

Potentials. In our experiments, unless otherwise specified, the default is that all pure 2- and 4-potential coefficients are drawn independently from Unif($[-8,8]$) distributions, while all pure 1- and 3-potential coefficients are set to 0. Using the notation introduced above, in each experiment a parameter $W_{\max}$ is varied, and the default distribution of one class of pure potentials (either 1-, 2-, 3-, or 4-potentials) is overridden from the default specification to be replaced by coefficients from a Unif($[-W_{\max}, W_{\max}]$) distribution.

LibDAI settings. In all cases, we use the junction tree algorithm with Hugin updates for exact inference. For partition function estimation, we use the LibDAI `HAK` implementation of Heskes et al. (2003), with precise parameters passed to MATLAB given by:

```
'[doubleloop=1,clusters=BETHE,init=UNIFORM,tol=1e-9,maxiter=10000]'.
```

For approximate MAP inference, we use the libDAI `BP` loopy belief propagation implementation, with precise parameters passed to MATLAB given by:

```
'[inference=MAXPROD,updates=SEQFIX,logdomain=1,tol=1e-9,maxiter=10000,damping=0.0]'.
```

The `HAK` method finds a stationary point of the related Kikuchi free energy by minimising a sequence of convex upper bounds on the Kikuchi objective. These stationary points are known to be in correspondence with fixed points of generalised belief propagation (Yedidia et al., 2001). In contrast, loopy max-product belief propagation is a fast, simple method for approximate MAP inference, but does not always exactly solve the associated Sherali–Adams relaxation (see e.g. Wainwright et al. (2002)).

2.5.3 Overview of experimental results

Before giving in-depth commentary on our experimental results, we first summarise the experiments that were run, together with a summary of high-level findings. We provide results on partition function inference in Section 2.5.4, and the following additional experiments results in the Appendix:

- Section 2.B.1: Timings;
- Section 2.B.2: MAP inference;
- Section 2.B.3: Marginal inference;
- Section 2.B.4: Higher-order potentials over clusters of 5 and 6 variables;
- Section 2.B.5: Comparison of our heuristics to the *maxtW* heuristic used in Weller (2016b);
- Section 2.B.6: Experiments on larger models.

Considering all results across models and approximate methods for estimating $\log Z$, marginals and MAP inference (see Figure 2.3 and Sections 2.B.1-2.B.6), we make the following general observations. Both K and G heuristics perform well (in and out of sample): they never hurt materially and often significantly improve accuracy, attaining results close to the best possible rerooting. Since our two heuristics achieve similar performance, sensitivity to the exact constants in Expression (2.18) appears low. We verified this by comparing to *maxtW* for pairwise models as in Weller (2016b) — both K and G heuristics performed slightly better than *maxtW*. For all runs, inference on rerooted models and on original models took similar times (time required to reroot and later to map back inference results is negligible relative to the cost of performing approximate inference) — see Section 2.B.1.

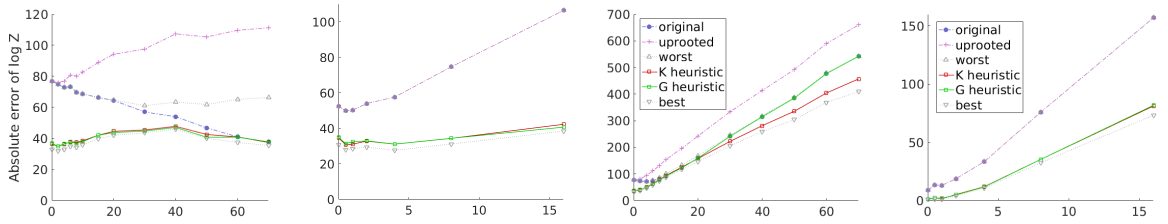
Observe that stronger 1-potentials tend to make inference *easier*, pulling each variable towards a specific setting, effectively concentrating the probability distribution onto a small collection of configurations, and reducing the benefits from rerooting (see for example the left column of Figure 2.3). Stronger pure k-potentials for $k > 1$ intertwine variables more tightly; this typically makes inference harder and increases the gains in accuracy from rerooting. The pure k-potential perspective facilitates this analysis.

When we examine larger models, or models with still higher order potentials, we observe qualitatively similar results — see Sections 2.B.4 and 2.B.6.

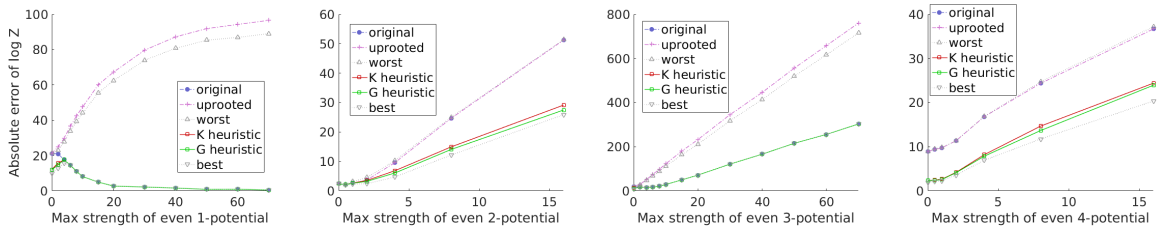
2.5.4 Partition function inference

Our results are displayed in Figure 2.3. In general, both the G heuristic and K heuristic perform well, in all cases providing approximation errors close to that of the optimal choice of rerooting. We observe in particular that for the complete graphs with varied pure 1- and 3-potential strengths, the original model transitions from being a poor choice of rerooting to a better a choice of rerooting as the strength of the potentials concerned increases, and the heuristics are able to distinguish when it is worth swapping to a different rerooting.

Average abs(error) in $\log Z$ for K_8 complete hypergraphs (fully connected) on 8 variables.



Average abs(error) in $\log Z$ for Grids on 5×5 variables (toroidal). Legends are consistent across all plots.



vary $W_{\max}$ for 1-pots | vary $W_{\max}$ for 2-pots | vary $W_{\max}$ for 3-pots | vary $W_{\max}$ for 4-pots

Figure 2.3. Error in estimating $\log Z$ for random models with various pure k-potentials over 20 runs. If not shown, the $W_{\max}$ coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. Where the red K heuristic curve is not visible, it coincides with the green G heuristic. Both K and G heuristics for selecting a rerooting work well: they never hurt and often yield large benefits. See Section 2.5 for details.

2.6 Discussion

We introduced methods which broaden the application of the uprooting and rerooting approach to binary models with higher-order potentials of any order. We demonstrated several theoretical insights, including Theorems 2.19 and 2.20 which show that $\mathbb{L}_3$ is unique in being universally rooted. We developed the helpful tool of even k-potentials in Section 2.3, which may be of independent interest. We empirically demonstrated significant benefits for rerooting in higher-order models — particularly for the hard case of strong

cluster potentials and weak 1-potentials — and provided an efficient heuristic to select a variable for rerooting. This heuristic is also useful to indicate when rerooting is unlikely to be helpful for a given model (if Expression (2.18) is maximised by taking $i = 0$).

It is natural to compare the effect of rerooting M to M_i, against simply clamping X_i in the original model M. A key difference is that rerooting achieves the clamping at X_i for negligible computational cost. In contrast, if X_i is clamped in the original model then the inference method will have to be run twice: once clamping $X_i = 0$, and once clamping $X_i = 1$, then results must be combined. This is avoided with rerooting given the symmetry of M^+. Rerooting effectively replaces what may be a poor initial implicit choice of clamping at X_0 with a carefully selected choice of clamping variable almost for free. This is true even for large models where it may be advantageous to clamp a series of variables: by rerooting, one of the series is obtained for free, potentially gaining significant benefit with little work required. Note that each separate connected component may be handled independently, with its own added variable. This could be useful for (repeatedly) composing clamping and then rerooting each separated component to obtain an almost free clamping in each.

Appendix 2.A Proofs

2.A.1 Proofs of results from Section 2.2

Proposition 2.3. We have

$$\operatorname{score}(x_V) = \operatorname{score}_+(x_V, x_0 = 0) = \operatorname{score}_+(\overline{x}_V, x_0 = 1)\,, \tag{2.4}$$

for all $x_V \in \{0,1\}^V$. Thus, for every configuration of the original model M, there are two corresponding configurations in M^+ with the same score, which are related by flipping all variables.

Proof. We observe directly from the definition of the potentials of the uprooted model M^+ that for any $x \in \{0,1\}^V$, we have

$$\operatorname{score}(x_V) = \sum_{\mathcal{E}\in E} \theta_{\mathcal{E}}(x_{\mathcal{E}}) = \sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_{\mathcal{E}}, x_0 = 0) = \sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(\overline{x}_{\mathcal{E}}, x_0 = 1)\,. \tag{2.21}$$

We note that the penultimate expression is equal to $\text{score}_+(x_V, x_0 = 0)$, and the final expression is equal to $\text{score}_+(\overline{x}_V, x_0 = 1)$, completing the proof. □

Proposition 2.5. Given a graphical model M and its uprooted model M^+, the following results hold:

- The partition function Z of M is related to the partition function Z^+ of M^+ according to $Z^+ = 2Z$.
- The configuration $x_V \in \{0,1\}^V$ is a MAP configuration for M iff $(x_V, x_0 = 0)$ is a MAP configuration for M^+ iff $(\overline{x}_V, x_0 = 1)$ is a MAP configuration for M^+. The MAP score is the same in all cases.
- The marginal distribution over a subset $U \subseteq V$ of variables in M may be expressed in terms of marginals of variables in M^+ as follows:

$$p(x_U) = p^+(x_U, x_0 = 0) + p^+(\overline{x}_U, x_0 = 1) = 2p^+(x_U, x_0 = 0) = 2p^+(\overline{x}_U, x_0 = 1)\,, \tag{2.6}$$

  for all $x_U \in \{0,1\}^U$. Similarly, the marginal distribution over a subset $U \subseteq V^+$ of variables in M^+ may be expressed in terms of marginals of variables in M as follows:

  - If $0 \notin U$, then

$$p^+(x_U) = \frac{1}{2}p(x_U) + \frac{1}{2}p(\overline{x}_U)\,, \tag{2.7}$$

    for all $x_U \in \{0,1\}^U$.

  - If $0 \in U$, then writing $U_0 = U \setminus \{0\}$, we have

$$p^+(x_{U_0}, x_0 = 0) = \frac{1}{2}p(x_{U_0})\,, \qquad p^+(x_{U_0}, x_0 = 1) = \frac{1}{2}p(\overline{x}_{U_0})\,, \tag{2.8}$$

    for all $x_{U_0} \in \{0,1\}^{U_0}$.

Proof. **Partition functions.** The partition function claim holds by the following calculation:

$$\begin{aligned} Z^+ &= \sum_{x_{V^+}\in\{0,1\}^{V^+}} \exp\left(\sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_{\mathcal{E}^+})\right) \\ &= \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_0=0,x_{\mathcal{E}})\right) + \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_0=1,x_{\mathcal{E}})\right) \\ &= \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}\in E} \theta_{\mathcal{E}}(x_{\mathcal{E}})\right) + \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}\in E} \theta_{\mathcal{E}}(\overline{x}_{\mathcal{E}})\right) \\ &= 2Z\,. \end{aligned} \tag{2.22}$$

MAP configurations. Equality of optimal score follows immediately from the result of Proposition 2.3 that $\text{score}(x_V) = \text{score}_+(x_V, x_0 = 0) = \text{score}_+(\overline{x}_V, x_0 = 1)$ for all $x_V \in \{0,1\}^V$, from which the correspondence between optimal configurations also follows.

Marginal distributions. Let $x_U \in \{0,1\}^U$. Observe that

$$\begin{aligned} p(x_U) &= \frac{1}{Z} \sum_{x_{V\setminus U}\in\{0,1\}^{V\setminus U}} \exp\left(\sum_{\mathcal{E}\in E} \theta_{\mathcal{E}}(x_{\mathcal{E}})\right) \\ &= \frac{1}{Z} \sum_{x_{V\setminus U}\in\{0,1\}^{V\setminus U}} \exp\left(\sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_0=0,x_{\mathcal{E}})\right) \\ &= \frac{1}{2Z}\left(\sum_{x_{V\setminus U}\in\{0,1\}^{V\setminus U}} \exp\left(\sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_0=0,x_{\mathcal{E}})\right) + \sum_{x_{V\setminus U}\in\{0,1\}^{V\setminus U}} \exp\left(\sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_0=1,\overline{x}_{\mathcal{E}})\right)\right) \\ &= p^+(x_0=0,x_U) + p^+(x_0=1,\overline{x}_U) = 2p^+(x_0=0,x_U) = 2p^+(x_0=1,\overline{x}_U)\,, \end{aligned} \tag{2.23}$$

as required. □

2.A.2 Proofs of results from Section 2.3

Proposition 2.10 (**All pure potentials are essentially even potentials**)**.** Let $k \geq 2$, and $|U| = k$. If $\theta_U : \{0,1\}^U \to \mathbb{R}$ is a pure k-potential then θ_U must be an affine function of the even k-potential, i.e. there exist $a \in \mathbb{R}\setminus\{0\}, b \in \mathbb{R}$ such that $\theta_U(x_U) = a\mathbb{1}[\,|\{i \in U | x_i = 1\}| \text{ is even}] + b$.

Proof. It is sufficient to demonstrate that if, for two configurations $x_U, y_U \in \{0,1\}^U$, we have $\sum_{i\in U} x_i = \sum_{i\in U} y_i \bmod 2$, then $\theta_U(x_U) = \theta_U(y_U)$, since this shows that θ_U depends

on its input argument only through the quantity $\mathbb{1}_{\#\{i\in U|x_i=1\} \text{ is even}}$, and since this only takes on two possible values, θ_U may be expressed as an affine function of this indicator.

To demonstrate the claim above, it is sufficient to show that if $x_U \in \{0,1\}^U$, and $i, j \in V$ are two distinct indices, and $F_{ij}(x_U) \in \{0,1\}^U$ denotes the configuration obtained from x_U by flipping coordinates i and j, then $\theta_U(x_U) = \theta_U(F_{ij}(x_U))$. This is sufficient since given $x_U, y_U \in \{0,1\}^U$ with $\sum_{i\in U} x_i = \sum_{i\in U} y_i \bmod 2$, it is possible to obtain y_U from x_U by iteratively flipping pairs of distinct variables.

Let $F_i(x_U)$ denote the configuration obtained from x_U by flipping x_i. By the uniform marginalisation property, we have

$$p(x_U) + p(F_i(x_U)) = p(F_j(x_U)) + p(F_{ij}(x_U)) \tag{2.24}$$

and

$$p(F_i(x_U)) + p(F_{ij}(x_U)) = p(x_U) + p(F_j(x_U))\,. \tag{2.25}$$

Subtracting these equations from one another yields

$$p(x_U) = p(F_{ij}(x_U))\,. \tag{2.26}$$

Taking logarithms of this equation yields $\theta_U(x_U) = \theta_U(F_{ij}(x_U))$, as required. □

Proposition 2.11 (**Even k-potentials form a basis**)**.** For a finite set U, the set of even k-potentials $\left(\mathbb{1}[\,|\{i \in W|x_i = 1\}| \text{ is even}]\right)_{W\subseteq U}$, indexed by subsets $W \subseteq U$, forms a basis for the vector space of all potential functions $\theta : \{0,1\}^U \to \mathbb{R}$.

Proof. We show that the indicators $(\mathbb{1}[|\{i \in W|x_i = 1\}| \text{ is even}])_{W\subseteq U}$ form a basis for the vector space of functions $\mathbb{R}^{\{0,1\}^U}$; we interpret the indicator corresponding to the empty set as being the constant function equal to 1. Given this, we then note that $\mathscr{P}(\{0,1\}^U)$ may be viewed as a convex subset of an affine subspace of $\mathbb{R}^{\{0,1\}^U}$ of co-dimension 1; more precisely, we identify the measure $\mu \in \mathscr{P}(\{0,1\}^U)$ with the vector $(\mu(x_U)|x_U \in \{0,1\}^U)$. We then note that the indicator corresponding to the empty set is orthogonal to this affine subspace. This is then sufficient to show that for any probability distribution $\mu \in \mathscr{P}(\{0,1\}^U)$, there is a unique set of parameters $(\eta_W)_{\emptyset\neq W\subseteq U}$ such that

$$\mu(x_U) = 1 + \sum_{\emptyset\neq W\subseteq U} \eta_W \mathbb{1}[|\{i \in W|x_i = 1\}| \text{ is even}]\,, \tag{2.27}$$

as required.

To demonstrate that $(\mathbb{1}[|\{i \in W | x_i = 1\}| \text{ is even}])_{W \subseteq U}$ forms a basis for the vector space of functions $\mathbb{R}^{\{0,1\}^U}$, we first note that it has the correct number of elements to form a basis, and it is therefore sufficient to either demonstrate that it is a spanning set, or that it is a linearly independent set; we take the latter approach. Suppose we have a collection of coefficients $(\alpha_W)_{W \subseteq U}$ such that

$$\sum_{W \subseteq U} \alpha_W \mathbb{1}[|\{i \in W | x_i = 1\}| \text{ is even}] = 0\,, \tag{2.28}$$

for all $x_U \in \{0,1\}^U$. Given a subset $X \subseteq U$, note that we have

$$\begin{aligned}
&\Big(\mathbb{1}[|\{i \in X | x_i = 1\}| \text{ is even}] - \mathbb{1}[|\{i \in X | x_i = 1\}| \text{ is odd}]\Big) \cdot \\
&\qquad\qquad \Big(\sum_{W \subseteq U} \alpha_W \mathbb{1}[|\{i \in W | x_i = 1\}| \text{ is even}]\Big) = 0 \\
\implies &\sum_{W \subseteq U} \alpha_W \sum_{x_U \in \{0,1\}^U} \Big(\mathbb{1}[|\{i \in W | x_i = 1\}| \text{ is even}, |\{i \in X | x_i = 1\}| \text{ is even}] \\
&\qquad\qquad - \mathbb{1}[|\{i \in W | x_i = 1\}| \text{ is even}, |\{i \in X | x_i = 1\}| \text{ is odd}]\Big) = 0\,.
\end{aligned} \tag{2.29}$$

Considering the summand above for a fixed subset $W \subseteq U$, note that if $W = X$, then the result of summing over all configurations $x_U \in \{0,1\}^U$ is $2^{|U|-1}$. However, if $W \neq X$, the result of the sum is 0. From this it immediately follows that $\alpha_X = 0$, and the proof of linear independence is complete. An elegant perspective which demonstrates that the sum concerned above evaluates to 0 is to view $\{0,1\}^U$ as a vector space over the finite field with 2 elements $\mathbb{F}_2$, with addition defined component-wise. In this case, the set $\{x \in \{0,1\}^U | \#\{i \in W | x_i = 1\} \text{ is even}\}$ is exactly the kernel of the linear form $\{0,1\}^U \ni x \mapsto \sum_{i \in W} x_i \in \mathbb{F}_2$ (where the addition is to be interpreted modulo 2). Considering the linear form $\{x \in \{0,1\}^U | \#\{i \in W | x_i = 1\} \text{ is even}\} \ni x \mapsto \sum_{i \in X} x_i \in \mathbb{F}_2$, we observe that the two sets

$$\begin{aligned}
&\{x \in \{0,1\}^U | \#\{i \in W | x_i = 1\} \text{ is even}, \#\{i \in X | x_i = 1\} \text{ is even}\} \text{ and} \\
&\{x \in \{0,1\}^U | \#\{i \in W | x_i = 1\} \text{ is even}, \#\{i \in X | x_i = 1\} \text{ is odd}\}\,,
\end{aligned} \tag{2.30}$$

are the preimage of $0 \in \mathbb{F}_2$ and $1 \in \mathbb{F}_2$ under this linear form, respectively. Therefore, if the linear form is surjective, the two sets have the same size, and since they are clearly disjoint, the relevant term of Equation (2.29) evaluates to 0. To see that the form is surjective, recall that by assumption $X \neq W$. If $X \setminus W$ is non-empty, then surjectivity is demonstrated by changing a single coordinate corresponding to an index in $X \setminus W$. If $X \setminus W$ is empty,

then $W \setminus X$ is non-empty, and by simultaneously changing a coordinate in $W \setminus X$ and X, surjectivity is demonstrated. □

2.A.3 Proofs of results from Section 2.4

Theorem 2.17. For a hypergraph $G = (V, E)$, and integer r such that $\max_{\mathcal{E} \in E} |\mathcal{E}| \leq r \leq |V|$, there is an affine score-preserving bijection

$$\mathbb{L}_r(G) \underset{\textbf{RootAt0}}{\overset{\textbf{Uproot}}{\rightleftarrows}} \widetilde{\mathbb{L}}^0_{r+1}(\nabla G). \tag{2.15}$$

Proof. The structure of the proof is as follows. We first construct the uprooting map **Uproot**, which we will denote by $\Psi : \mathbb{L}_r(G) \to \widetilde{\mathbb{L}}^0_{r+1}(\nabla G)$ for notational convenience, and show that it is bijective by exhibiting its double-sided inverse, **RootAt**0, which we will denote by $\Phi : \widetilde{\mathbb{L}}^0_{r+1}(\nabla G) \to \mathbb{L}_r(G)$. We then directly show that this bijection is affine and score-preserving.

To construct Ψ, let $\mu = (\mu_{\mathcal{E}} | \mathcal{E} \in E) \in \mathbb{L}_r(G)$, and define

$$\Psi(\mu) = \mu^+ = \left(\mu^+_{\mathcal{E}^+} \in \mathscr{P}(\{0,1\}^{\mathcal{E}^+}) \middle| \mathcal{E}^+ \in E^+ \right) \in \widetilde{\mathbb{L}}^0_{r+1}(\nabla G) \tag{2.31}$$

as follows. For each uprooted hyperedge $\mathcal{E}^+ \in E^+$, we define:

$$\mu^+_{\mathcal{E}^+}(x_{\mathcal{E}^+}) = \begin{cases} \frac{1}{2}\mu_{\mathcal{E}}(x_{\mathcal{E}}) & \text{if } x_0 = 0 \\ \frac{1}{2}\mu_{\mathcal{E}}(\overline{x}_{\mathcal{E}}) & \text{if } x_0 = 1 \end{cases} \qquad \forall\, x_{\mathcal{E}^+} \in \{0,1\}^{\mathcal{E}^+}. \tag{2.32}$$

To show $\mu^+ \in \widetilde{\mathbb{L}}^0_{r+1}(\nabla G)$, we exhibit a consistent marginalising family $(\mu^+_{U^+} | U \subseteq V, |U| = r+1)$ for μ^+ with respect to $\widetilde{\mathbb{L}}^0_{r+1}(G)$ (see Definition 1.10). Since $\mu \in \mathbb{L}_r(G)$, there exists a consistent marginalising family $(\mu_U | U \subseteq V, |U| = r)$ for μ with respect to $\mathbb{L}_r(G)$. From this, we define

$$\mu^+_{U^+}(x_{U^+}) = \begin{cases} \frac{1}{2}\mu_U(x_U) & \text{if } x_0 = 0 \\ \frac{1}{2}\mu_U(\overline{x}_U) & \text{if } x_0 = 1 \end{cases} \qquad \forall\, x_{U^+} \in \{0,1\}^{U^+}. \tag{2.33}$$

The family $(\mu^+_{U^+} | U \subseteq V, |U| = r)$ is clearly locally consistent, and marginalises down to the family $(\mu_{\mathcal{E}^+} | \mathcal{E}^+ \in E^+)$, with each measure flipping invariant, as required.

Thus, we have defined the map $\Psi : \mathbb{L}_r(G) \to \widetilde{\mathbb{L}}^0_{r+1}(\nabla G)$. We now exhibit its inverse.

Given $\eta \in \widetilde{\mathbb{L}}^0_{r+1}(\nabla G)$, we define $\Phi(\eta) = \mu = \left(\mu_{\mathcal{E}} \in \mathscr{P}(\{0,1\}^{\mathcal{E}}) \middle| \mathcal{E} \in E\right) \in \mathbb{L}_r(G)$ as follows. Given $\mathcal{E} \in E$, define

$$\mu_{\mathcal{E}}(x_{\mathcal{E}}) = \eta_{\mathcal{E}^+}(x_0 = 0, x_{\mathcal{E}}) + \eta_{\mathcal{E}^+}(x_0 = 1, \overline{x}_{\mathcal{E}}), \qquad \forall x_{\mathcal{E}} \in \{0,1\}^{\mathcal{E}}. \tag{2.34}$$

A consistent marginalising family for $(\mu_{\mathcal{E}} | \mathcal{E} \in E)$, demonstrating that $(\mu_{\mathcal{E}} | \mathcal{E} \in E) \in \mathbb{L}_r(G)$, is given as follows. For each $U \subseteq V$ with $|U| = r$, we define

$$\mu_U(x_U) = \eta_{U^+}(x_0 = 0, x_U) + \eta_{U^+}(x_0 = 1, \overline{x}_U), \qquad \forall x_U \in \{0,1\}^U. \tag{2.35}$$

Local consistency of these measures follows from that of η, as does the fact that they marginalise down to $(\mu_{\mathcal{E}} | \mathcal{E} \in E)$.

We now directly show that for $\mu \in \mathbb{L}_r(G)$, we have $\Phi(\Psi(\mu)) = \mu$, and for all $\eta \in \widetilde{\mathbb{L}}^0_{r+1}(\nabla G)$, we have $\Psi(\Phi(\eta)) = \eta$. This demonstrates that Ψ and Φ are two-sided inverses of one another, and hence are bijective maps on their domains. First, let $\mu \in \mathbb{L}_r(G)$. We take $\mathcal{E} \in E$, and note that from our definitions of Ψ and Φ, we have for all $x_{\mathcal{E}} \in \{0,1\}^{\mathcal{E}}$ that

$$\Phi(\Psi(\mu))_{\mathcal{E}}(x_{\mathcal{E}}) = \mu^+_{\mathcal{E}}(x_{\mathcal{E}}, x_0 = 0) + \mu^+_{\mathcal{E}}(\overline{x}_{\mathcal{E}}, x_0 = 1) = \frac{1}{2}\mu_{\mathcal{E}}(x_{\mathcal{E}}) + \frac{1}{2}\mu_{\mathcal{E}}(\overline{\overline{x}}_{\mathcal{E}}) = \mu_{\mathcal{E}}(x_{\mathcal{E}}). \tag{2.36}$$

Now let $\eta \in \widetilde{\mathbb{L}}^0_{k+1}(G)$, and write $\mu = \Phi(\eta)$. Taking $\mathcal{E} \in E$, we note

$$\begin{aligned}
\Psi(\Phi(\eta))_{\mathcal{E}^+}(x_{\mathcal{E}^+}) &= \frac{1}{2}\mu_{\mathcal{E}}(x_{\mathcal{E}})\mathbb{1}[x_0 = 0] + \frac{1}{2}\mu_{\mathcal{E}}(\overline{x}_{\mathcal{E}})\mathbb{1}[x_0 = 1] \\
&= \frac{1}{2}\left(\eta_{\mathcal{E}^+}(x_{\mathcal{E}}, x_0 = 0) + \eta_{\mathcal{E}^+}(\overline{x}_{\mathcal{E}}, x_0 = 1)\right)\mathbb{1}[x_0 = 0] \\
&\quad + \frac{1}{2}\left(\eta_{\mathcal{E}^+}(\overline{x}_{\mathcal{E}}, x_0 = 0) + \eta_{\mathcal{E}^+}(\overline{\overline{x}}_{\mathcal{E}}, x_0 = 1)\right)\mathbb{1}[x_0 = 1] \\
&= \frac{1}{2}\left(\eta_{\mathcal{E}^+}(x_{\mathcal{E}^+}) + \eta_{\mathcal{E}^+}(\overline{x}_{\mathcal{E}^+})\right)\mathbb{1}[x_0 = 0] \\
&\quad + \frac{1}{2}\left(\eta_{\mathcal{E}^+}(\overline{x}_{\mathcal{E}^+}) + \eta_{\mathcal{E}^+}(x_{\mathcal{E}^+})\right)\mathbb{1}[x_0 = 1] \\
&= \eta_{\mathcal{E}^+}(x_{\mathcal{E}^+}),
\end{aligned} \tag{2.37}$$

where in the final equality we have used the flipping-invariance of $\eta_{\mathcal{E}^+}$.

Finally, to see that the map is score-preserving, let $(\theta_{\mathcal{E}} | \mathcal{E} \in E)$ be a collection of potentials defining a model on $G = (V, E)$, and let $(\theta^+_{\mathcal{E}^+} | \mathcal{E}^+ \in E^+)$ be the corresponding set of potentials defining the uprooted model on $\nabla G = (V^+, E^+)$. Then for any $\mu^+ \in \widetilde{\mathbb{L}}^0_{k+1}(G)$, note that we

have

$$\begin{aligned}
&\sum_{\mathcal{E}\in E} \mathbb{E}_{X_{\mathcal{E}^+}\sim\mu^+_{\mathcal{E}^+}}\left[\theta^+_{\mathcal{E}^+}(X_{\mathcal{E}^+})\right] \\
&= \sum_{\mathcal{E}\in E} \sum_{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+}} \theta^+_{\mathcal{E}^+}(x_{\mathcal{E}^+})\mu^+_{\mathcal{E}^+}(x_{\mathcal{E}^+}) \\
&= \sum_{\mathcal{E}\in E}\left[\sum_{\substack{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+}\\ x_0=0}} \theta^+_{\mathcal{E}^+}(x_{\mathcal{E}^+})\mu^+_{\mathcal{E}^+}(x_{\mathcal{E}^+}) + \sum_{\substack{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+}\\ x_0=1}} \theta^+_{\mathcal{E}^+}(x_{\mathcal{E}^+})\mu^+_{\mathcal{E}^+}(x_{\mathcal{E}^+})\right] \\
&= \sum_{\mathcal{E}\in E}\left[\sum_{\substack{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+}\\ x_0=0}} \theta_{\mathcal{E}}(x_{\mathcal{E}})\frac{1}{2}\mu_{\mathcal{E}}(x_{\mathcal{E}}) + \sum_{\substack{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+}\\ x_0=1}} \theta_{\mathcal{E}}(\overline{x}_{\mathcal{E}})\frac{1}{2}\mu_{\mathcal{E}}(\overline{x}_{\mathcal{E}})\right] \\
&= \sum_{\mathcal{E}\in E} \sum_{x_{\mathcal{E}}\in\{0,1\}^{\mathcal{E}}} \theta_{\mathcal{E}}(x_{\mathcal{E}})\mu_{\mathcal{E}}(x_{\mathcal{E}}) \\
&= \sum_{\mathcal{E}\in E} \mathbb{E}_{X_{\mathcal{E}}\sim\mu_{\mathcal{E}}}\left[\theta_{\mathcal{E}}(X_{\mathcal{E}})\right], \qquad (2.38)
\end{aligned}$$

as required. □

Theorem 2.19. $\mathbb{L}_3$ is universally rooted.

Proof. The analysis of Weller (2016b) shows that $\widetilde{\mathbb{L}}^0_4(\nabla G) = \widetilde{\mathbb{L}}^i_4(\nabla G)$ for all hypergraphs $G = (V, E)$ of maximum hyperedge degree ≤ 2, and for all $i \in V$. We now simply need to show that given $i \in V$, and augmenting the hyperedge set of G with an additional collection of hyperedges of degree 3 all containing i, that we still have $\widetilde{\mathbb{L}}^0_4(\nabla G) = \widetilde{\mathbb{L}}^i_4(\nabla G)$. We argue this directly, letting $H = (V, F)$ denote this augmented hypergraph. Let $(\mu_{\mathcal{E}^+}|\mathcal{E}^+ \in F^+) \in \widetilde{\mathbb{L}}^0_4(\nabla H)$, and let $(\mu_U|U \subseteq V^+,\ |U| = 4,\ 0 \in U)$ be a consistent marginalising family. Then note that the restricted set of hyperedge marginals $(\mu_{\mathcal{E}^+}|\mathcal{E}^+ \in E^+)$ lies in $\widetilde{\mathbb{L}}^0_4(\nabla G)$, and hence in $\widetilde{\mathbb{L}}^i_4(\nabla G)$; let $(\nu_U|U \subseteq V^+,\ |U| = 4,\ i \in V)$ by a consistent marginalising family in the latter polytope. We then claim that $(\mu_U|U \subseteq V^+,\ |U| = 4,\ 0 \in U,\ i \in U) \cup (\nu_U|U \subseteq V^+,\ |U| = 4,\ i \in U,\ 0 \notin U)$ is a consistent marginalising family for $(\mu^+_{\mathcal{E}}|\mathcal{E}^+ \in F^+)$ in $\widetilde{\mathbb{L}}^i_4(\nabla H)$. This is readily verified; consistency of these measures over four variables onto a subset of two variables is immediate, since $\widetilde{\mathbb{L}}^0_4(\nabla G) = \widetilde{\mathbb{L}}^i_4(\nabla G)$. Consistency on subsets of three variables follows due to the flipping-invariance of the polytopes concerned; flipping-invariant distributions over three variables are determined by their marginals on pairs of variables. This demonstrates that $\widetilde{\mathbb{L}}^0_4(\nabla H) \subseteq \widetilde{\mathbb{L}}^i_4(\nabla H)$; the reverse inclusion is entirely analogous. □

Theorem 2.20. $\mathbb{L}_3$ is unique in being universally rooted. Specifically, for any integer $r > 1$ other than $r = 3$, we constructively demonstrate a hypergraph $G = (V, E)$ with $|V| = r + 1$ variables for which $\widetilde{\mathbb{L}}^0_{r+1}(\nabla G) \neq \widetilde{\mathbb{L}}^i_{r+1}(\nabla G)$ for any $i \in V$.

Proof. For each $r \neq 3$, we shall constructively demonstrate a model M on hypergraph G as stated such that the LP relaxation over $\mathbb{L}_r(G)$ is not tight for M but the LP relaxation over $\mathbb{L}_r(\nabla G \setminus \{i\})$ is tight for every rerooted model $M_i, i \in V$.

Case 1: r is even. Let $G = (V, E)$, with $V = \{1, \ldots, r+1\}$, and E the set of all subsets of V of size r. Consider a model with the following set of potentials on this hypergraph:

$$\theta_{\mathcal{E}}(x_{\mathcal{E}}) = -\mathbb{1}[\#\{i \in \mathcal{E} | x_i = 1\} \text{ is even}] \qquad \forall \mathcal{E} \in E. \tag{2.39}$$

The optimum score for a configuration $x_V \in \{0,1\}^V$ is -1. We show this by demonstrating (i) that the optimum is at most -1 (which is all we need here), then (ii) that the optimum is at least -1. For (i): Toward contradiction, assume that there exists a configuration $x_V \in \{0,1\}^V$ that does not activate any of the $\theta_{\mathcal{E}}$ potentials, i.e. all r-clusters have an odd number of variables set to 1. Pick one of the r-clusters and call it S. Since $r \geq 2$ is even, S contains at least one index $a \in V$ such that $x_a = 0$, and at least one index $b \in V$ such that $x_b = 1$. Now V has $r + 1$ elements consisting of those in S, together with one more, which we denote by z. If $z = 0$ then consider the r-cluster $T = S \setminus \{b\} \cup \{z\}$. If $z = 1$ then let $T = S \setminus \{a\} \cup \{z\}$. In either case, we have that $\#\{i \in T | x_i = 1\}$ is even, and hence the configuration x_V activates the potential θ_T, incurring a score of -1. For (ii): Consider the setting $x_1 = 1$ with all other variables set to 0. All r-clusters including x_1 are inactive. There is just one r-cluster not including x_1, and this r-cluster has no 1s thus its potential is active. Hence, this configuration achieves a score of -1.

However, the set of pseudomarginal distributions in $\mathbb{L}_r(G)$ below attains a score of 0:

$$\mu_{\mathcal{E}}(x_{\mathcal{E}}) = \frac{1}{r} \sum_{i \in \mathcal{E}} \delta_{x_i = 1, x_{\mathcal{E} \setminus \{i\}} = 0} \qquad \forall \mathcal{E} \in E. \tag{2.40}$$

Now observe that when this model is uprooted, we have the hypergraph $\nabla G = (V^+, E)$, where $V^+ = V \cup \{0\}$, and the hyperedge set $E^+ = E$ as before with the same set of potentials as in Equation (2.39), by Lemma 2.12. Therefore, rerooting at a variable $i \in \{1, \ldots, r+1\}$ will result in a graphical model on the graph $\nabla G \setminus \{i\}$ with vertices $\{0, 1, \ldots, r+1\} \setminus \{i\}$, and hyperedges given by one hyperedge of size r (the original hyperedge which did not include i), which is $\{1, \ldots, r+1\} \setminus \{i\}$, along with all subsets of $\{1, \ldots, r+1\} \setminus \{i\}$ of size $r - 1$. In

particular, the model consists of potentials over the set of r variables $\{1,\dots,r+1\}\setminus\{i\}$, and the variable X_0 is independent of the rest of the variables, with symmetric distribution on its state space $\{0,1\}$. Therefore, the polytope $\mathbb{L}_r(\nabla G\setminus\{i\})$ is tight for this potential since it is effectively a model over r variables, proving the claim.

Case 2: $r\geq 5$ is odd. Let $r\geq 5$ be odd, and again let $G=(V,E)$, with $V=\{1,\dots,r+1\}$, this time letting E be the set of all subsets of V of size $r-1$ (an even number). Consider the following set of potentials on this hypergraph

$$\theta_{\mathcal{E}}(x_{\mathcal{E}}) = -\mathbb{1}[\#\{i\in\mathcal{E}\,|\,x_i=1\}\text{ is even}] \qquad \forall\mathcal{E}\in E. \tag{2.41}$$

We note that the polytope $\mathbb{L}_r(G)$ is not tight for this potential, by considering the following set of pseudomarginals over hyperedges of G:

$$\mu_{\mathcal{E}}(x_{\mathcal{E}}) = \frac{1}{r}\delta_{x_{\mathcal{E}}=0} + \frac{1}{r}\sum_{i\in\mathcal{E}}\delta_{x_i=1,x_{\mathcal{E}\setminus\{i\}}=0} \qquad \forall\mathcal{E}\in E. \tag{2.42}$$

These are valid pseudomarginals in $\mathbb{L}_r(G)$, as the following distributions over r-clusters are consistent and marginalise down to the distributions over hyperedges:

$$\mu_U(x_U) = \frac{1}{r}\sum_{i\in U}\delta_{x_i=1,x_{U\setminus\{i\}}=0} \qquad \forall U\subseteq V, |U|=r. \tag{2.43}$$

Note that the score of this set of pseudomarginals is

$$\sum_{\mathcal{E}\in E} -\mu_{\mathcal{E}}(\#\{i\in\mathcal{E}\,|\,x_i=1\}\text{ is even}) = -\binom{r+1}{r-1}\frac{1}{r} = -\frac{r+1}{2}. \tag{2.44}$$

We now argue that this exceeds the maximum score obtainable by a configuration $x_V\in\{0,1\}^V$, demonstrating non-tightness of $\mathbb{L}_r(G)$ for this model. To see this, let $\ell\in\{0,\dots,r+1\}$ be the number of non-zero coordinates of x_V. We count the number of subsets U of $\{1,\dots,r+1\}$ of size $r-1$ for which x_U has an even number of non-zero coordinates, and show that this is greater than $(r+1)/2$, leading to a score of less than $-(r+1)/2$. The number of such subsets is given by

$$\sum_{p=0}^{\lfloor \ell/2\rfloor}\binom{\ell}{2p}\binom{r+1-l}{r-1-2p} = \begin{cases}\binom{\ell}{\ell-2}\binom{r+1-l}{r+1-l}+\binom{\ell}{\ell}\binom{r+1-l}{r-1-\ell} = \frac{\ell(\ell-1)}{2}+\frac{(r+1-\ell)(r-\ell)}{2} & \ell\text{ even}\\ \binom{\ell}{\ell-1}\binom{r+1-\ell}{r-\ell} = \ell(r+1-\ell) & \ell\text{ odd}.\end{cases} \tag{2.45}$$

For ℓ even, we observe that the quadratic expression in ℓ above is minimised at $\ell = (r+1)/2$ (which is an integer, as r is odd), and takes the value $(r^2-1)/4$, which is greater than $(r+1)/2$ for all odd $r \geq 5$ (though the two values are equal for $r = 3$). For ℓ odd, we observe that the minimal value of the quadratic expression above is r, which is greater than $(r+1)/2$ for all odd $r \geq 5$.

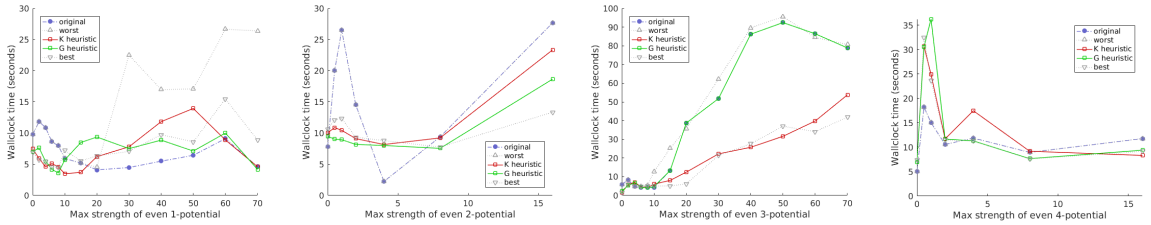
Now observe that when this model is uprooted and subsequently rerooted at a new variable $i \in V$, we obtain a model on $r+1$ variables, but with the variable X_0, introduced by uprooting, independent of the rest. Therefore, the model is effectively over only r variables, and hence it follows that $\mathbb{L}_r(\nabla G \setminus \{i\})$ is tight for this rerooting, proving the claim. □

Appendix 2.B Additional experimental results

2.B.1 Timings

Times in seconds to run marginal inference (i.e. estimating $\log Z$ and marginals) using libDAI are shown in Figure 2.4. Inference for rerooted models took a similar amount of time as for the original model. We caution against relying heavily on the accuracy of these timings since we made no attempt to optimise our code for speed, and we ran our inference algorithms in a cluster environment beyond our control.

Time/sec to run marginal inference for K_8 complete hypergraphs (fully connected) on 8 variables.



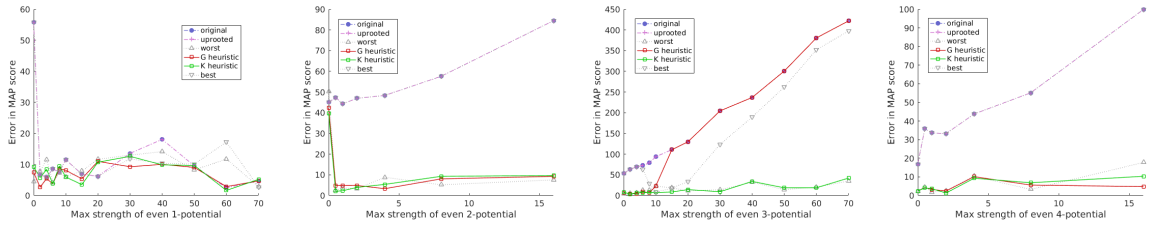
vary $W_{\max}$ for 1-pots | vary $W_{\max}$ for 2-pots | vary $W_{\max}$ for 3-pots | vary $W_{\max}$ for 4-pots

Figure 2.4. Average time to perform marginal inference using libDAI over 20 runs. If not shown, $W_{\max}$ max coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. *Best* and *worst* refer to the rerootings which *ex post* gave the lowest error in estimating $\log Z$.

2.B.2 MAP inference

Results are shown in Figure 2.5. We observe here that rerooting does not help much when 1-potentials are varied, but can provide great benefit for the other cases shown. The K heuristic (which was trained on complete graphs like the one we analyse here) performs well in all settings. Curiously, the G heuristic (which was trained only on grids) performs well when 2-potentials or 4-potentials are varied, but not when 3-potentials are varied (though even here it does no worse than the original rooting). We aim to explore this further in future work.

Error in estimating MAP score for K_8 complete hypergraphs (fully connected) on 8 variables.



vary $W_{\max}$ for 1-pots | vary $W_{\max}$ for 2-pots | vary $W_{\max}$ for 3-pots | vary $W_{\max}$ for 4-pots

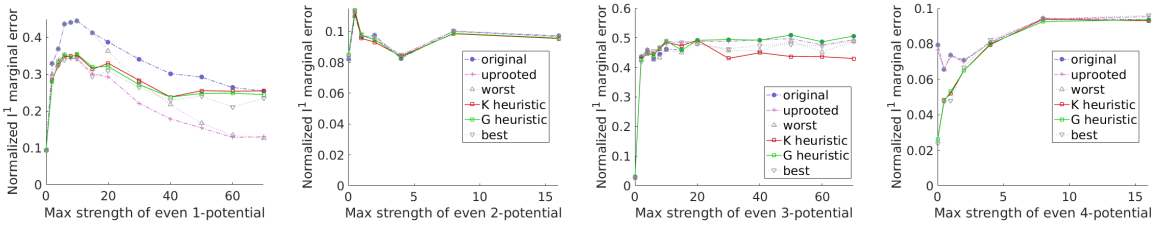
Figure 2.5. Average error in estimating MAP score using libDAI over 20 runs. If not shown, the $W_{\max}$ coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. *Best* and *worst* refer to the rerootings which *ex post* gave the lowest error in estimating $\log Z$.

2.B.3 Marginal inference

Results are shown in Figure 2.6. Our models were selected to present an interesting range of problems for partition function estimation, which led to marginals often being challenging to estimate. However, results for marginal inference were often improved by rerooting.

We note that another natural way to estimate marginals is as the ratio of a clamped partition function to the original partition function. Since we have seen good evidence that rerooting can help significantly with partition function estimation, it is reasonable to hope that in future work, we may observe benefits to marginal inference via this approach by using rerooting.

Error in estimating 1-marginals for K_8 complete hypergraphs (fully connected) on 8 variables.



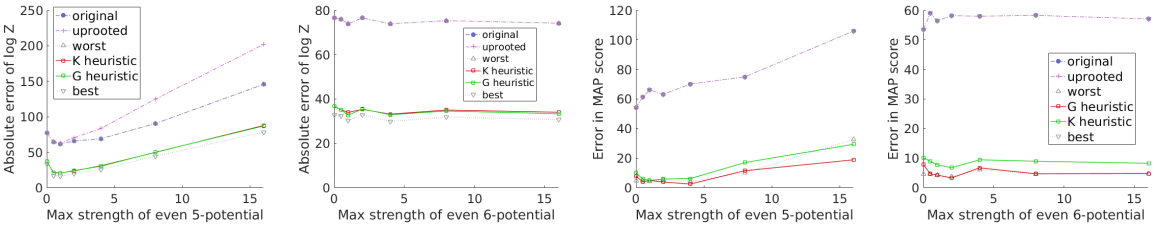
vary $W_{\max}$ for 1-pots | vary $W_{\max}$ for 2-pots | vary $W_{\max}$ for 3-pots | vary $W_{\max}$ for 4-pots

Figure 2.6. Average ℓ_1 error in estimating marginals (minimal representation corresponding to pure k-potentials — see Section 2.3) using libDAI over 20 runs. If not shown, the $W_{\max}$ coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. *Best* and *worst* refer to the rerootings which *ex post* gave the lowest error in estimating $\log Z$.

2.B.4 Higher-order potentials over clusters of 5 and 6 variables

Results for partition function estimation over a complete hypergraph K_8 on 8 variables, this time with potentials up to order 6, are shown in Figure 2.7. In all cases, as in the case with lower-order potentials, rerooting using our heuristics is very helpful.

Error in estimating $\log Z$ (left) and MAP score (right) for K_8 hypergraphs on 8 variables with **potentials up to order 6**.



vary $W_{\max}$ for 5-pots | vary $W_{\max}$ for 6-pots | vary $W_{\max}$ for 5-pots | vary $W_{\max}$ for 6-pots

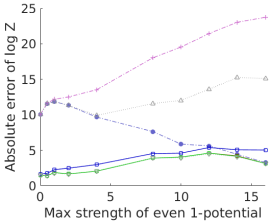
Figure 2.7. Average error in estimating $\log Z$ (left) and MAP score (right) using libDAI over 20 runs. If not shown, the $W_{\max}$ coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. *Best* and *worst* refer to the rerootings which *ex post* gave the lowest error in estimating $\log Z$.

2.B.5 Comparison of our heuristics to the earlier *maxtW* heuristic

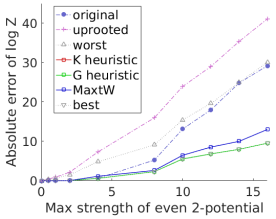
Results for a complete graph K_8 on 8 variables, this time with potentials only up to order 2, are shown in Figure 2.8. We have added the earlier *maxtW* heuristic used in Weller (2016b), which using our notation corresponds to setting $t_2 = \frac{1}{2}$ in Expression (2.18). Note that for the pairwise models considered here, the clamp heuristic constants for potentials of order higher than 2 are irrelevant. We observe that our K and G heuristics (fitted on different models with potentials up to order 4, so here are out of sample) achieve similar

performance to the earlier *maxtW* heuristic, in fact yielding slightly better results. This is encouraging evidence for robustness of the simple form of heuristic score in Expression (2.18).

Average abs(error) in $\log Z$ for K_8 complete **pairwise** graphs (fully connected) on 8 variables:
adding earlier *maxtW* heuristic for comparison (our K and G heuristics coincide on these runs).



vary $W_{\max}$ for 1-pots



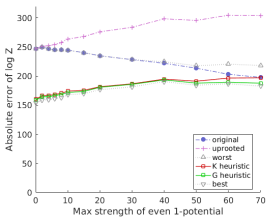
vary $W_{\max}$ for 2-pots

Figure 2.8. Error in estimating $\log Z$ for random pairwise models with various pure k-potentials over 20 runs. If not shown, the $W_{\max}$ coefficients for pure k-potentials are 8 for $k = 1$, and 8 for $k = 2$. K and G heuristics coincide. See Section 2.B.5.

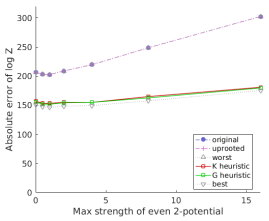
2.B.6 Larger models

Results for a complete hypergraph K_{10} on 10 variables (potentials up to order 4) are shown in Figure 2.9. Results are qualitatively similar to those for smaller models in Section 2.5.

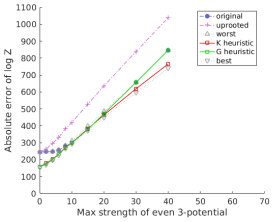
Average abs(error) in $\log Z$ for K_{10} complete hypergraphs (fully connected) on 10 variables.



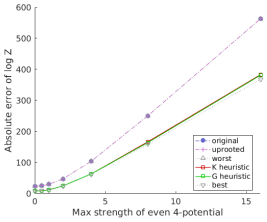
vary $W_{\max}$ for 1-pots



vary $W_{\max}$ for 2-pots



vary $W_{\max}$ for 3-pots



vary $W_{\max}$ for 4-pots

Figure 2.9. Error in estimating $\log Z$ for random models with various pure k-potentials over 20 runs. If not shown, the $W_{\max}$ coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. See Section 2.B.6.

Chapter 3

Tightness of LP Relaxations for Almost Balanced Models

This chapter is based on the following publications:

- Weller, A., Rowland, M., and Sontag, D. (2016). Tightness of LP relaxations for almost balanced models. In *Artificial Intelligence and Statistics (AISTATS)*.
- Rowland, M., Pacchiano, A., and Weller, A. (2017). Conditions beyond treewidth for tightness of higher-order LP relaxations. In *Artifical Intelligence and Statistics (AISTATS)*.

The proof of Theorem 3.2 is an original contribution by the author of the thesis, the original version of which appears in Rowland et al. (2017). The proof is based on that of a weaker result appearing in Weller et al. (2016), which was a joint contribution by the authors of the paper. Writing was undertaken jointly for both of these papers, although the presentation here has been significantly modified for the purposes of this thesis.

3.1 Introduction

Tractability of MAP inference for balanced pairwise graph models has long been understood; indeed, in Theorem 1.14, we recall that for any graph G, the pairwise Sherali–Adams relaxation $\mathbb{L}_2(G)$ is guaranteed to be tight for all balanced pairwise models on G. Many real-world instances of MAP inference problems are over pairwise graphical models which are *not* balanced, and often for such models, the pairwise relaxation $\mathbb{L}_2(G)$ is not tight. In this case, the polytope relaxation is required to be tightened in order to obtain the

correct solution to the MAP inference problem. That is, we are required to move up the Sherali–Adams hierarchy.

It has been demonstrated (see e.g. Batra et al. (2011); Komodakis and Paragios (2008); Sontag et al. (2008)) that using higher-order cluster constraints to tighten the polytope $\mathbb{L}_2(G)$ to a more constrained (yet still tractable) polytope enables many real-world instances of the MAP inference problem to be exactly solved efficiently. This is perhaps surprising, given the worst-case hardness results for MAP inference discussed in Section 1.2, and suggests that there are some properties of real-world MAP inference instances that allow tractable inference to be performed. A large open problem, which this chapter takes steps towards, is understanding why many graphical models of interest can be solved exactly using Sherali–Adams polytopes with relatively few constraints.

However, aside from purely topological conditions bounding treewidth (see Theorem 1.11), to date there has been little theoretical understanding of when these higher-order cluster methods will be effective. In this chapter, we focus on binary pairwise models, and provide an important contribution by proving that adding certain triplet constraints to $\mathbb{L}_2(G)$ yields a polytope that is guaranteed to be tight for the significant class of models that satisfy the hybrid condition (combining restrictions on graph topology and potentials) of being *almost balanced*, first considered in Weller (2015), and defined precisely below.

3.2 Almost balanced models

We now introduce the concepts required to state our main result. The first notion is that of an almost balanced model; given the tightness guarantees that exist for balanced models (see Theorem 1.14), this model class is natural to consider.

Definition 3.1 (**Almost balanced models**)**.** A binary graphical model M given by a graph $G = (V, E)$ is said to be almost balanced if there exists a vertex $s \in V$ such that the model that results from deleting the vertex s is balanced. Any vertex $s \in V$ which may be deleted to yield a balanced model is referred to as a *distinguished vertex* in the model.

For an almost balanced model on a graph $G = (V, E)$ with distinguished vertex $s \in V$, we define the pairwise Sherali–Adams polytope with localised triplet clusters at s as the generalised Sherali–Adams polytope (see Definition 1.15) for the covering antichain S,

given by

$$S = \{\{s, i, j\} | i, j \in V \setminus \{s\}, i \neq j\}. \tag{3.1}$$

Note then that

$$\mathbb{L}_3(G) \subseteq \mathbb{L}_S(G) \subseteq \mathbb{L}_2(G), \tag{3.2}$$

that is, the polytope $\mathbb{L}_S(G)$ is at least as stringent as $\mathbb{L}_2(G)$ in enforcing consistency amongst pseudomarginals, but no more stringent than $\mathbb{L}_3(G)$. We remark also that the number of linear constraints that may be used to represent the polytope $\mathbb{L}_S(G)$ scales quadratically with $|V|$, rather than cubically, as with $\mathbb{L}_3(G)$.

For the remainder of this chapter, we refer to the polytope $\mathbb{L}_S(G)$ with the notation $\mathbb{L}_3^s(G)$, to make explicit that it enforces triplet-consistency for triplets involving s, and to avoid the reader having to keep the meaning of the set S in mind.

We now state the main result of this chapter.

Theorem 3.2. Let M be an almost balanced model with graph $G = (V, E)$, and distinguished vertex $s \in V$. Then (i) the polytope $\mathbb{L}_3^s(G)$ is tight for the model M, and (ii) there are no linear constraints of $\mathbb{L}_3^s(G)$ that may be removed and yield a polytope which is tight for all almost balanced models on G with distinguished vertex s.

This provides an important step into hybrid characterisations of tightness (Cooper and Živný, 2011), which remains an exciting uncharted field following success in characterisations of tractability using only topological constraints (Chandrasekaran et al., 2008), or only families of potentials (Kolmogorov et al., 2015; Thapper and Živný, 2016). We remark also that MAP inference in an almost balanced model is straightforward; a distinguished vertex may be identified by considering the $|V|$ models given by deleting a vertex from G, and checking whether each is balanced. Once the distinguished vertex $s \in V$ has been found, we can obtain two balanced models by conditioning the variable corresponding to the distinguished vertex to take the values 0 and 1. MAP inference on these balanced models can then be solved (e.g. using the pairwise Sherali–Adams relaxation), and the MAP configuration for the almost balanced model can be straightforwardly derived from these results. However, we emphasise that even given this algorithm for MAP inference on almost balanced models, it is far from trivial to understand the tightness of Sherali–Adams relaxations for such models. The remainder of this chapter focuses on proving Theorem

3.2, with additional commentary on the result and possibilities for further work given in Section 3.6.

3.3 Preliminaries

For the purposes of analysis, we will work with a particular minimal representation of the Sherali–Adams polytope $\mathbb{L}_3^s(G)$ of interest — see Section 1.4.2 for a review of minimal representations. We consider the minimal representation of the form $((q_i)_{i\in V},(q_{ij})_{ij\in E}) \in \mathbb{R}^{V\cup E}$ for elements of $\mathbb{L}_3^s(G)$, where for a collection of abstract measures $((\mu_i|i\in V),(\mu_{ij}|ij\in E)) \in \mathbb{L}_3^s(G)$, the corresponding parameters under the minimal representation are $q_i = \mathbb{P}_{\mu_i}(X_i = 1)$ for all $i \in V$ and $q_{ij} = \mathbb{P}_{\mu_{ij}}(X_i = 1, X_j = 1)$ for all $ij \in E$. We describe the explicit form of the local inequalities (the pairwise inequalities each involving two variables sharing an edge, describing the polytope $\mathbb{L}_2(G)$) and the triplet inequalities (which together with the local inequalities describe the polytope $\mathbb{L}_3(G)$) under this minimal representation below. In this way, we obtain the precise set of linear equalities that describe the polytope $\mathbb{L}_3^s(G)$ under this parametrisation.

From the definition of the polytope $\mathbb{L}_2(G)$, we need to ensure that for each pair of distinct vertices $i, j \in V$, there exists a valid joint distribution over the variables X_i and X_j. Recalling that q_{ij} represents the probability $\mathbb{P}(X_i = 1, X_j = 1)$, and q_i (respectively q_j) represents $\mathbb{P}(X_i = 1)$ (respectively $\mathbb{P}(X_j = 1)$), we have the following parametrisation for the four atomic events associated with X_i and X_j:

$$\begin{aligned}
\mathbb{P}(X_i = 1, X_j = 1) &= q_{ij}\,,\\
\mathbb{P}(X_i = 0, X_j = 1) &= q_j - q_{ij}\,,\\
\mathbb{P}(X_i = 1, X_j = 0) &= q_i - q_{ij}\,,\\
\mathbb{P}(X_i = 0, X_j = 0) &= 1 - q_i - q_j + q_{ij}\,.
\end{aligned} \tag{3.3}$$

By the form of our parametrisation these quantities sum to 1, so it is sufficient to enforce that they are all non-negative to ensure that the parameters q_i, q_j, and q_{ij} correspond to a valid distribution. Enforcing these non-negativity constraints yields the linear constraints

of the polytope $\mathbb{L}_2(G)$ under this minimal representation:

$$\begin{aligned} 0 &\leq q_{ij} \\ q_i + q_j - 1 &\leq q_{ij} \\ q_{ij} &\leq q_i \\ q_{ij} &\leq q_j \qquad \forall ij \in E. \end{aligned} \tag{3.4}$$

We reiterate that these will be referred to as local inequalities, and we will occasionally refer to them as LOC inequalities for short. Note that these inequalities are also sometimes referred to as Fréchet bounds in the literature. The additional constraints needed to describe the polytope $\mathbb{L}_3(G)$ can be derived in the same way; we now need to ensure that for every triplet of distinct indices $i, j, k \in V$ for which we are asserting local consistency, there exists a joint distribution over X_i, X_j, X_k such that $q_i, q_j, q_k, q_{ij}, q_{ik}, q_{jk}$ are marginals of this distribution. The atomic events associated with this distribution over three variables can be expressed in terms of these six parameters and one additional parameter α, which will represent the probability of all three variables taking the value 1. We then have

$$\begin{aligned} \mathbb{P}(X_i = 1, X_j = 1, X_k = 1) &= \alpha, \\ \mathbb{P}(X_i = 0, X_j = 1, X_k = 1) &= q_{jk} - \alpha, \\ \mathbb{P}(X_i = 1, X_j = 0, X_k = 1) &= q_{ik} - \alpha, \\ \mathbb{P}(X_i = 1, X_j = 1, X_k = 0) &= q_{ij} - \alpha, \\ \mathbb{P}(X_i = 1, X_j = 0, X_k = 0) &= q_i - q_{ij} - q_{ik} + \alpha, \\ \mathbb{P}(X_i = 0, X_j = 1, X_k = 0) &= q_j - q_{ij} - q_{jk} + \alpha, \\ \mathbb{P}(X_i = 0, X_j = 0, X_k = 1) &= q_k - q_{ik} - q_{jk} + \alpha, \\ \mathbb{P}(X_i = 0, X_j = 0, X_k = 0) &= 1 - q_i - q_j - q_k + q_{ij} + q_{ik} + q_{jk} - \alpha. \end{aligned} \tag{3.5}$$

Again, by construction of the parametrisation these eight quantities sum to 1, so we just need to enforce non-negativity of each entry to ensure that they describe a valid joint distribution. This gives rise to eight inequalities; performing Fourier-Motzkin elimination of the variable α in these inequalities then yields the local inequalities derived above for each pair of vertices (i, j and i, k and j, k), as well as four additional triplet inequalities

involving all three vertices:

$$\begin{aligned} q_i + q_{jk} &\geq q_{ij} + q_{ik} \\ q_j + q_{ik} &\geq q_{ij} + q_{jk} \\ q_k + q_{ij} &\geq q_{ik} + q_{jk} \\ q_{ij} + q_{ik} + q_{jk} &\geq q_i + q_j + q_k - 1\,. \end{aligned} \tag{3.6}$$

We refer to these inequalities as the triplet (or TRI) inequalities. The polytope $\mathbb{L}_3(G)$ is obtained by enforcing the local constraints in Expression (3.4) for each pair of distinct vertices in V, and the TRI inequalities in Expression (3.6) for each triplet of distinct vertices in V. To obtain the polytope $\mathbb{L}_3^s(G)$, we enforce the local constraints for each pair of distinct vertices in V, and the TRI inequalities for each triplet of distinct vertices in V which includes the variable s.

Note also that with this minimal representation, any binary pairwise graphical model over G may be specified by parameters $\phi = ((\phi_v | v \in V), (\phi_{ij} | ij \in E)) \in \mathbb{R}^{V \cup E}$, with distribution given by

$$\mathbb{P}(X_V = x_V) = \frac{1}{Z} \exp\left(\sum_{v \in V} \phi_v x_v + \sum_{ij \in E} \phi_{ij} x_i x_j \right). \tag{3.7}$$

With these parametrisations of (pseudo)marginals and potential functions, the linear program over the Sherali–Adams relaxation $\mathbb{L}_3^s(G)$ that concerns us in this chapter may be expressed as

$$\arg\max_{q \in \mathbb{L}_3^s(G)} \langle \phi, q \rangle\,. \tag{3.8}$$

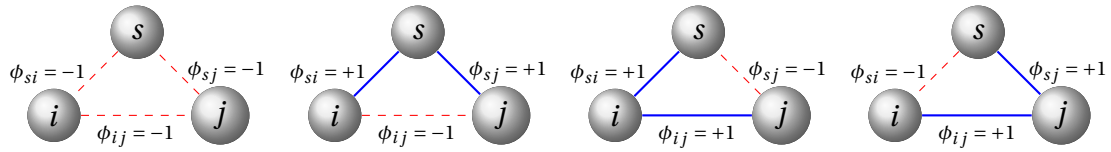
With this preliminary notation established, we may now proceed with the proof of Theorem 3.2.

3.4 Proof of the second claim of Theorem 3.2

We first give a short proof for the second claim of Theorem 3.2, and then discuss the structure of the proof for the first claim.

Consider the polytope given by enforcing all linear constraints of $\mathbb{L}_3^s(G)$ except one TRI constraint, say on a triangle sij. There are four possible inequalities to remove: a) $q_{si} +$

$q_{sj}+q_{ij} \geq q_s+q_i+q_j-1$; b) $q_s+q_{ij} \geq q_{si}+q_{sj}$; c) $q_i+q_{sj} \geq q_{si}+q_{ij}$; d) $q_j+q_{si} \geq q_{sj}+q_{ij}$. In each case, we consider a model with all potentials apart from those corresponding to the singletons s, i and j and those corresponding to the edges si, sj, ij set to 0. We set the singleton potentials ϕ_s, ϕ_i, ϕ_j and edge potentials $\phi_{si}, \phi_{sj}, \phi_{ij}$ as in the corresponding Figures 3.1a to 3.1d. We now show that these potentials are non-tight.



(a) $\phi_s = \phi_i = \phi_j = 1/2$ (b) $\phi_s = \phi_i = \phi_j = -1/2$ (c) $\phi_s = \phi_i = \phi_j = -1/2$ (d) $\phi_s = \phi_i = \phi_j = -1/2$

Figure 3.1. Solid blue edges are attractive, dashed red edges are repulsive.

- *Inequality a).* In this case the MAP score is $1/2$, attained by the configuration $(X_s, X_i, X_j) = (1,0,0)$. A valid configuration in the polytope given by $\mathbb{L}_3^s(G)$ without inequality a) is $(q_s, q_i, q_j, q_{si}, q_{sj}, q_{ij}) = (1/2, 1/2, 1/2, 0, 0, 0)$, which has score $3/4$, demonstrating non-tightness of this polytope.
- *Inequality b).* In this case the MAP score is 0, attained by the configuration $(X_s, X_i, X_j) = (0,0,0)$. A valid configuration in the polytope given by $\mathbb{L}_3^s(G)$ without inequality b) is $(q_s, q_i, q_j, q_{si}, q_{sj}, q_{ij}) = (1/2, 1/2, 1/2, 1/2, 1/2, 0)$, which has score $1/4$, demonstrating non-tightness of this polytope.
- *Inequality c).* In this case the MAP score is 0, attained by the configuration $(X_s, X_i, X_j) = (0,0,0)$. A valid configuration in the polytope given by $\mathbb{L}_3^s(G)$ without inequality c) is $(q_s, q_i, q_j, q_{si}, q_{sj}, q_{ij}) = (1/2, 1/2, 1/2, 1/2, 0, 1/2)$, which has score $1/4$, demonstrating non-tightness of this polytope.
- *Inequality d).* In this case the MAP score is 0, attained by the configuration $(X_s, X_i, X_j) = (0,0,0)$. A valid configuration in the polytope given by $\mathbb{L}_3^s(G)$ without inequality d) is $(q_s, q_i, q_j, q_{si}, q_{sj}, q_{ij}) = (1/2, 1/2, 1/2, 0, 1/2, 1/2)$, which has score $1/4$, demonstrating non-tightness of this polytope.

3.5 Proof of the first claim of Theorem 3.2

The proof of the first claim of Theorem 3.2 is considerably more substantial than that of the second claim. For clarity, we begin by setting out the high-level structure of the proof.

3.5.1 Structure of the proof for the first claim of Theorem 3.2

We begin by noting that it is sufficient to prove the result for almost attractive models, since given an almost balanced model, there exists a flipping (see Section 1.4.1) that renders the model almost *attractive*, and preserves properties of (non-)tightness of $\mathbb{L}_3^s(G)$, by Proposition 1.19.

Next, consider an almost attractive graphical model M, with underlying graph $G = (V, E)$ and distinguished variable $s \in V$. Without loss of generality, we may take G to be complete by taking potentials ϕ_{ij} on any edges ij that need to be added to be equal to 0. In fact, it suffices to prove the claim for models in which any edge potentials ϕ_{ij} which are equal to 0 are set to a small positive constant, since the set of potentials for which a Sherali–Adams relaxation is tight is closed. Let $\phi \in \mathbb{R}^{V \cup E}$ be a representation of the potentials of this model under the minimal representation discussed in Section 3.3, so that the probability of a given configuration $x_V \in \{0,1\}^V$ under the model is proportional to

$$\exp\left(\left\langle \phi, ((\mathbb{1}_{\{x_i=1\}})_{i\in V}, (\mathbb{1}_{\{x_i=x_j=1\}})_{ij\in E}) \right\rangle\right), \tag{3.9}$$

and the MAP inference problem is given by

$$\max_{x_V \in \{0,1\}^V} \left[\sum_{v\in V} \phi_v \mathbb{1}_{\{x_i=1\}} + \sum_{ij\in E} \phi_{ij} \mathbb{1}_{\{x_i=x_j=1\}} \right]. \tag{3.10}$$

We then define the function $F^s_{\mathbb{L}_3^s(G)} : [0,1] \to \mathbb{R}$ by

$$F^s_{\mathbb{L}_3^s(G)}(t) = \sup_{q \in \mathbb{L}_3^s(G)|_{q_s=t}} \langle \phi, q \rangle, \tag{3.11}$$

where the notation $\mathbb{L}_3^s(G)|_{q_s=t}$ denotes the polytope $\mathbb{L}_3^s(G)$ with the additional constraint that the marginal probability of the event $\{X_s = 1\}$ is equal to t. We will show the mildly stronger result that $F^s_{\mathbb{L}_3^s(G)}$ is linear, and from this it will follow that its maximal value is attained for $t = 0$ or $t = 1$. This shows that a global maximiser of ϕ is achieved over $\mathbb{L}_3^s(G)$ when $q_s = 0$ or $q_s = 1$. But we know that $\mathbb{L}_3^s(G)|_{q_s=0}$ (respectively, $\mathbb{L}_3^s(G)|_{q_s=1}$) is tight with respect to $M|_{X_s=0}$ (respectively, $M|_{X_s=1}$), since these problems are equivalent to MAP inference problems for balanced models over $G \setminus \{s\}$, and regarding the polytopes $\mathbb{L}_3^s(G)|_{q_s=0}$ and $\mathbb{L}_3^s(G)|_{q_s=1}$ as polytopes over this reduced model, they are both equal to $\mathbb{L}_2(G \setminus \{s\})$, which is tight on balanced models. The result then follows.

Thus, it remains to show that $F^s_{\mathbb{L}^s_3(G)}$ is linear. To do this, we will show that it is both convex and concave. The proof of concavity is straightforward, and is given below. The proof of convexity is more involved, and requires several structural notions for optimal pseudomarginal distributions in $\mathbb{L}^s_3(G)$; these are given in Section 3.5.2.

Lemma 3.3. $F^s_{\mathbb{L}^s_3(G)}$ is concave.

Proof. Let $t_1, t_2 \in [0,1]$, and let $\lambda \in [0,1]$. Let $q^{(i)} \in \arg\max_{q \in \mathbb{L}^s_3(G)|_{q_s = t_i}} \langle \phi, q \rangle$ for $i = 1,2$, so that we have

$$F^s_{\mathbb{L}^s_3(G)}(t_i) = \left\langle \phi, q^{(i)} \right\rangle \tag{3.12}$$

for $i = 1,2$. Let $\widetilde{t} = \lambda t_1 + (1-\lambda) t_2$. Then consider $\widetilde{q} = \lambda q^{(1)} + (1-\lambda) q^{(2)}$. By convexity of $\mathbb{L}^s_3(G)$, $\widetilde{q} \in \mathbb{L}^s_3(G)$, and since $\widetilde{q}_s = \lambda q^{(1)}_s + (1-\lambda) q^{(2)}_s = \lambda t_1 + (1-\lambda) t_2 = \widetilde{t}$, we have $\widetilde{q} \in \mathbb{L}^s_3(G)|_{q_s = \widetilde{t}}$. By linearity of the score, we have

$$F^s_{\mathbb{L}^s_3(G)}(\widetilde{t}) \geq \langle \phi, \widetilde{q} \rangle = \lambda \langle \phi, q^{(1)} \rangle + (1-\lambda) \langle \phi, q^{(2)} \rangle = \lambda F^s_{\mathbb{L}^s_3(G)}(t_1) + (1-\lambda) F^s_{\mathbb{L}^s_3(G)}(t_2), \tag{3.13}$$

which demonstrates concavity. □

3.5.2 Strong and locking edges

Given $q = ((q_i)_{i \in V}, (q_{ij})_{ij \in E}) \in \mathbb{L}^s_3(G)$, we say that edge ij is *strong* if a local constraint is tight for vertices i and j, so that $q_{ij} = \min(q_i, q_j)$ or $q_{ij} = \max(0, q_i + q_j - 1)$; in the case of the former, we say that edge ij is *strong up*, and in the case of the latter, *strong down*. We say that a cycle $i_1 i_2, \ldots, i_m i_1$ in G is *strong frustrated* if all edges are strong, and there are an odd number of strong down edges. We say that variables i and j are *locking up* (or *locked up*) if $q_i = q_j = q_{ij}$, and that variables i and j are *locking down* (or *locked down*) if $q_i = 1 - q_j$, and $q_{ij} = 0$. If an edge is not strong, we say it is *weak*.

We conclude this section with several preliminary results that will be of use in the main structural and perturbative arguments that follow.

Lemma 3.4. If for a triplet of vertices $\{s, i, j\} \subset V$, there are two tight triplet constraints, there must be a locking edge in this triplet.

Proof. By flipping and permutation symmetries, it is sufficient to prove the lemma in the case that the following two triplet constraints are in fact tight:

$$q_i + q_{sj} \geq q_{si} + q_{ij}\,, \quad q_j + q_{si} \geq q_{sj} + q_{ij}\,. \tag{3.14}$$

As we assume that equality holds for both of the inequalities above, adding them together yields

$$q_i + q_j = 2q_{ij}\,. \tag{3.15}$$

But together with the local constraint $q_{ij} \leq \min(q_i, q_j)$, this implies that $q_{ij} = q_i = q_j$, so the edge ij is indeed locking up. □

Lemma 3.5. Suppose $q^* \in \mathbb{L}_3^s(G)|_{q_s=t}$ is optimal for a potential $\phi \in \mathbb{R}^{V\cup E}$ and suppose there is an edge $ij \in E$ away from s such that $\phi_{ij} > 0$ (that is, the edge is attractive). Then if ij is not strong up, then one of the TRI constraints $q_i^* + q_{sj}^* \geq q_{si}^* + q_{ij}^*$ and $q_j^* + q_{si}^* \geq q_{sj}^* + q_{ij}^*$ must be tight.

Proof. Since $\phi_{ij} > 0$, if it were possible to increase q_{ij}^* whilst keeping all other pseudo-marginals fixed, this would increase the score of q^*. Thus this must not be possible, and so there must be an inequality in $\mathbb{L}_3^s(G)|_{q_s=t}$ preventing q_{ij}^* from increasing any higher. Considering the inequalities described in Section 3.3, the only inequalities that upper-bound q_{ij}^* in $\mathbb{L}_3^s(G)|_{q_s=t}$ are the LOC constraints for the ij, and the two TRI constraints in the statement of the lemma. Since the edge is assumed not to be strong up, the LOC constraints cannot be tight, and so one of the TRI constraints must be. □

3.5.3 Main structural result for optimal vertices of the polytope $\mathbb{L}_3^s(G)|_{q_s=t}$

With the definitions of Section 3.5.2 in hand, we will demonstrate the following:

Theorem 3.6. Let $q^* \in \arg\max_{q\in\mathbb{L}_3^s(G)|_{q_s=t}} \langle \phi, q\rangle$. Then q^* has all singleton marginals in the set $\{0, t, 1-t, 1\}$, any edges between variables with singleton marginals in the set $\{t, 1-t\}$ are locking, and there are no strong frustrated cycles in the configuration.

With Theorem 3.6, we can then deduce convexity of $F^s_{\mathbb{L}^s_3(G)}$ immediately by writing

$$F^s_{\mathbb{L}^s_3(G)}(t) = \sup_{y\in[0,1]} \left\langle \phi, q(t|y) \right\rangle, \tag{3.16}$$

where $q(\cdot|y) : [0,1] \to \mathbb{L}^s_3(G)$ is defined as follows. Let q^* be optimal for ϕ in $\mathbb{L}^s_3(G)|_{q_s=y}$. By Theorem 3.6, the sets $A_y = \{j : q^*_j = 0\}, B_y = \{j : q^*_j = y\}$, $C_y = \{j : q^*_j = 1-y\}$ and $D_y = \{j : q^*_j = 1\}$ exhaustively partition the set V. Now define the components of $q(\cdot|y)$ as follows:

$$q_j(t|y) = \begin{cases} 0 & j \in A_y \\ t & j \in B_y \\ 1-t & j \in C_y \\ 1 & j \in D_y \end{cases}, \quad q_{ij}(t|y) = \begin{cases} 0 & i \in A_y \text{ or } j \in A_y \\ q_i(t|y) & j \in D_y \\ q_j(t|y) & i \in D_j \\ t & i,j \in B_y \\ 1-t & i,j \in C_y \\ 0 & i \in B_y \text{ and } j \in C_y\text{; or } i \in C_y \text{ and } j \in B_y. \end{cases}$$

Note that $q(t|y) \in \mathbb{L}^s_3(G)$ for all $t \in [0,1]$ since all edges of $q(t|y)$ are locking and there are no frustrated cycles. From Equation (3.16), note that $F^s_{\mathbb{L}^s_3(G)}$ is a supremum of affine functions, and is therefore convex. Thus, by the discussion in Section 3.5.1, Theorem 3.6 is sufficient to demonstrate linearity of the function $F^s_{\mathbb{L}^s_3(G)}$, and so the first claim of Theorem 3.2 follows.

In fact, we also remark that the structure of an element $q^* \in \operatorname{argmax}_{q\in\mathbb{L}^s_3(G)|_{q_s=y}} \langle \phi, q \rangle$ established by Theorem 3.6 reveals that the pseudomarginals of q^* actually arise as the mixture of one deterministic configuration $x^{(0)}_V \in \{0,1\}^V$ with $x^{(0)}_s = 0$, $x_i = 1$ for $i \in C_y \cup D_y$ and $x^{(0)}_i = 0$ otherwise, and another $x^{(1)}_V \in \{0,1\}^V$ with $x^{(1)}_s = 1$, $x^{(1)}_i = 1$ for $i \in B_y \cup D_y$, and $x^{(1)}_i = 0$ otherwise. Thus, another interpretation of this result is that we have shown that a collection of pseudomarginals that is optimal in the constrained polytope $\mathbb{L}^s_3(G)|_{q_s=y}$ is simply a mixture of pseudomarginals that are optimal for the polytopes $\mathbb{L}^s_3(G)|_{q_s=0}$ and $\mathbb{L}^s_3(G)|_{q_s=1}$.

We thus turn our attention to proving Theorem 3.6. We work up to proving Theorem 3.6 via several intermediate results. We first prove a series of structural results on the ways in which locking edges may be present in an optimal configuration in $\mathbb{L}^s_3(G)|_{q_s=t}$ for the potential ϕ in Section 3.5.4. We then show that with these structural results in hand, any configuration in $\mathbb{L}^s_3(G)|_{q_s=t}$ *not* of the form described in Theorem 3.6 admits a symmetric perturbation within the polytope $\mathbb{L}^s_3(G)|_{q_s=t}$, and thus cannot be extremal. This proves

Theorem 3.6, and thus by the remarks following the statement of Theorem 3.6, our main result Theorem 3.2 follows.

3.5.4 Structural results concerning locking edges

We now consider an optimal configuration $q^* \in \mathbb{L}_3^s(G)|_{q_s=t}$ for our almost attractive model, and establish a variety of results on the ways in which locking edges may be present in q^*. The assumption that q^* is optimal is implicit in all results in the remainder of this section. We first demonstrate that the relation on the set of variables defined by $i \sim j$ if the edge ij is locking (and prescribing that $i \sim i \ \forall i \in V$) is an equivalence relation, and use this to show that these *locking components* (that is, the equivalence classes under $\sim$) can, in some sense, be shrunk down to a single variable for the purposes of understanding the behaviour of the q^*. Reflexivity of $\sim$ is included in the definition, and symmetry is immediate. It remains to show transitivity. We provide a full proof of this result as Lemma 3.9, first giving intermediate results in Lemmas 3.7 and 3.8.

Our first result deals specifically with locked up edges.

Lemma 3.7. Locking up is a transitive relation on the set of vertices of the graph.

Proof. First consider vertices a, b, c distinct from s, and suppose a and b are locked up, and b and c are locked up. Consider the triangle sab. TRI inequalities imply that $q^*_{sa} = q^*_{sb}$. Similarly, TRI inequalities in sbc imply that $q^*_{sb} = q^*_{sc}$. Note that ac is locked up iff $q^*_{ac} = q^*_a$. Suppose this does not hold. Since the potential associated with edge ac is attractive, either sa or sc is holding ac down in a TRI constraint. By symmetry, suppose sa is holding down ac. Then

$$\begin{aligned} & q^*_c + q^*_{sa} = q^*_{sc} + q^*_{ac} \\ \implies & q^*_a + q^*_{sa} = q^*_{sa} + q^*_{ac} \\ \implies & q^*_a = q^*_{ac}, \end{aligned} \tag{3.17}$$

and so we deduce that a and c are locked up. Now consider vertices s, a, b, and suppose two pairs are locked up, so that the triangle sab has two locked up edges. TRI constraints immediately imply that the third edge is also locked up. □

We then use the following lemma to deal with locking down edges.

Lemma 3.8. i) If an edge ab away from s is locked down, then the edges sa and sb are also locked. ii) For any variables u, v such that su and sv are both locked, then uv is locked too.

Proof. For i), let $q_b^* = 1 - q_a^*$, and $q_{ab}^* = 0$. Since ab is attractive, this edge is being held down by some TRI inequality in the triangle sab. Without loss of generality, $q_a^* + q_{sb}^* = q_{sa}^* + q_{ab}^*$, so $q_a^* + q_{sb}^* = q_{sa}^*$, so $q_a^* = q_{sa}^*$ and $q_{sb}^* = 0$. Now consider the following TRI inequality:

$$\begin{aligned} & q_{sa}^* + q_{sb}^* + q_{ab}^* \geq q_s^* + q_a^* + q_b^* - 1 \\ \implies & q_a^* + 0 + 0 \geq q_s^* + q_a^* + (1 - q_a^*) - 1 \\ \implies & q_a^* \geq q_s^* . \end{aligned} \tag{3.18}$$

But $q_{sa}^* = q_a^*$, so $q_a^* \leq q_s^*$, so $q_s^* = q_a^*$, and so sa is locked up, and sb is locked down.

For ii), we show that any two variables locking to s lock with one another. Suppose u, v are locked to s. By flipping s if necessary, we need not deal with the case where u and v are both locked up with s.

Consider the case where su is locked up, and sv is locked down. Then consider the TRI inequality $q_u^* + q_{sv}^* \geq q_{su}^* + q_{uv}^*$. Substituting our values in , this gives $q_u^* + 0 \geq q_u^* + q_{uv}^*$. So $q_{uv}^* = 0$, and so uv is locking down.

Now consider the case where su and sv are locked down. Consider the following TRI inequality:

$$\begin{aligned} & q_{su}^* + q_{sv}^* + q_{uv}^* \geq q_s^* + q_u^* + q_v^* - 1 \\ \implies & q_{uv}^* \geq q_s^* + (1 - q_s^*) + (1 - q_s^*) - 1 \\ \implies & q_{uv}^* \geq 1 - q_s^* . \end{aligned} \tag{3.19}$$

But $q_{uv}^* \leq \min(q_u^*, q_v^*) = 1 - q_s^*$, so $q_{uv}^* = 1 - q_s^*$ and uv is locked up. □

Finally, the results above are combined to provide a full proof of transitivity.

Lemma 3.9. Locking is an equivalence relation on the set of vertices of the graph.

Proof. Let $a \sim b$, $b \sim c$; we aim to show that $a \sim c$. If ab is locked up and bc is locked up, then Lemma 3.7 applies immediately to give $a \sim c$. If both ab and ac are locked down, then we have from Lemma 3.8 i) that sa, sb, sc are all locked, and by applying Lemma

3.8 ii) to sa and sc, we deduce that $a \sim c$. Finally, if ab is locked down and bc is locked up, Lemma 3.8 i) applies to give either that sa is locked down and sb is locked up, or that sa is locked up and sb is locked down. In the case of the former, note that sb and bc are locked up, so sc is locked up by Lemma 3.7, and so ac is locked up by Lemma 3.8 ii). In the case of the latter, we consider the TRI inequality $q^*_c + q^*_{sb} \geq q^*_{sc} + q^*_{bc}$, and deduce that sc is locked, and so by Lemma 3.8 ii), ac is locked. This completes the proof. □

We summarise the results of this section in the following proposition, and provide an illustration in Figure 3.2; the large coloured circles represent distinct locking components.

Proposition 3.10. For a configuration $q^* \in \arg\max_{q \in \mathbb{L}^s_3(G)|_{q_s=t}} \langle \phi, q \rangle$, the vertices of the underlying graph G may be partitioned into locking components, such that any two variables which are locked are in the same component, and all variables in the same component lock with one another. Any variables which are locked down are in the same locking component as s, and there are no frustrated locking cycles within a locking component.

Proof. The partitioning of the vertex set is given immediately by the equivalence classes of the locking relation, as shown in Lemma 3.9. Lemma 3.8 i), shows that any two vertices that are locked down are also locked to s, and hence are in the same locking component. Finally, since locked down edges are in the same locking component as s, this is the only locking component in which a strong frustrated cycle could be. Noting also that the existence of a frustrated locking cycle is equivalent to the existence of a frustrated locking triplet, we suppose for a contradiction that we have a frustrated locking triplet in the same locking component of s. Note that s cannot lie in this locking frustrated triplet, due to the presence of triplet-consistency constraints for all triplets of variables involving s. Suppose then that abc is a locking frustrated triangle in the same locking component as s. Note that to have a frustrated locking triangle, we must have $q^*_a = q^*_b = q^*_c = 1/2$, and since s is in the same locking component, $q^*_s = 1/2$ as well. Next, at least one edge must be locking down for the triangle to be frustrated; without loss of generality take this to be ab, so that $q^*_{ab} = 0$. By the proof of Lemma 3.8 i), we have (without loss of generality) that sa is locking up and sb is locking down. Now if ac is locking up, TRI constraints over the triplet sac imply that sc is locking up too, and then TRI constraints over the triplet sbc imply that sb is locking down, so that the triangle abc is not locking frustrated after all. Similarly, if ac is locking down, TRI constraints imply first that sc is locking down, and then that bc is locking up, so again abc is not frustrated locking, yielding the contradiction required for the statement of the proposition. □

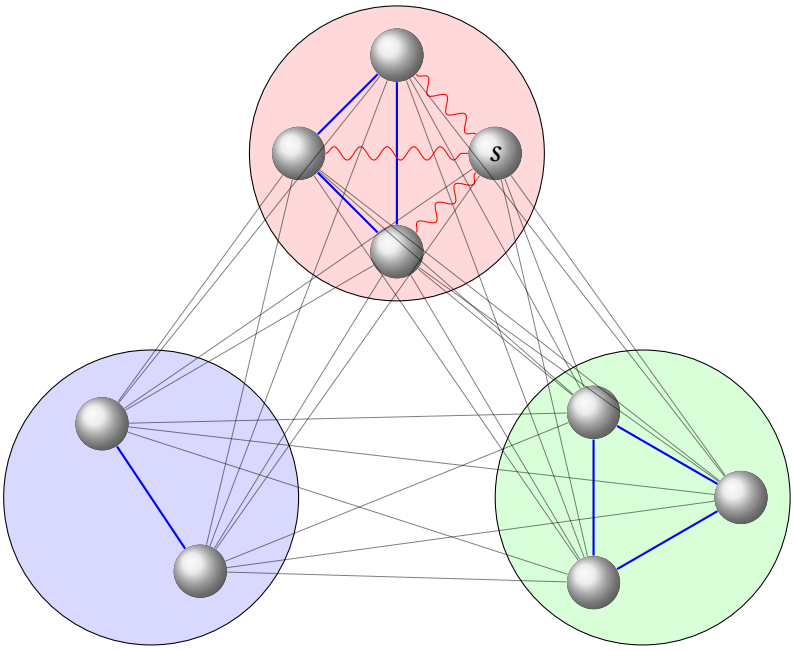


Figure 3.2. Example of the locking edge structure given by Proposition 3.10, with three locking components. The s-locking component is highlighted in red, and the 0-1-locking component is highlighted in blue.

Note that if the only locking components present are the one containing s, and the one containing any variables with singleton marginals in the set $\{0,1\}$, then the statement of Theorem 3.6 holds. In the next section, we use a perturbative argument to show that this indeed must be the case.

3.5.5 Primal perturbation

Given an optimal configuration $q^* \in \mathbb{L}_3^s(G)|_{q_s=t}$ for an almost attractive model, Proposition 3.10 establishes a partition of the variables into locking components. As noted already, if the only locking components present are the one containing s (which we refer to as the s-locking component) and the one containing any variables with singleton marginals in the set $\{0,1\}$ (which we refer to as the 0-1-locking component), then the statement of Theorem 3.6 holds. In this section, we follow a perturbative argument to show that for all almost attractive models, we do indeed have just the s-locking component and the 0-1-locking component.

We shall specify a non-zero perturbation $\delta \in \{-1,0,1\}^{V\cup E}$, and demonstrate that for a sufficiently small $\varepsilon > 0$, we have that $q^* \pm \varepsilon\delta \in \mathbb{L}_3^s(G)|_{q_s=t}$, by verifying that all constraints

defining the polytope $\mathbb{L}_3^s(G)|_{q_s=t}$ are still satisfied by $q^* \pm \varepsilon\delta$. The number ε is chosen so that any constraints in $\mathbb{L}_3^s(G)|_{q_s=t}$ which are not tight at q^* remain non-tight at $q^* \pm \varepsilon\delta$, and we will explicitly demonstrate that any constraints which *are* tight at q^* remain so at $q^* \pm \varepsilon\delta$. When δ is non-zero, this demonstrates that q^* is not extremal in $\mathbb{L}_3^s(G)|_{q_s=t}$. We shall see that the only configurations that do not admit a non-zero perturbation are those with only an s-locking component 0-1-locking component.

We begin by specifying the perturbation for singleton marginals.

Singleton marginal perturbations. The perturbation δ_v for the singleton marginal of any variable $v \in V$ in the s-locking component, or in the 0-1-locking component, is 0. For all other variables $v \in V$, the singleton perturbation δ_v depends on the edge marginal q_{sv} to s, according to the following exhaustive options:

$$\delta_v = \begin{cases} +1 & \text{if } v \text{ is strong up to } s \text{ and } q_v^* > q_s^*, \quad \text{or } v \text{ is strong down to } s \text{ and } q_v^* + q_s^* < 1 \\ -1 & \text{if } v \text{ is strong up to } s \text{ and } q_v^* < q_s^*, \quad \text{or } v \text{ is strong down to } s \text{ and } q_v^* + q_s^* > 1 \\ 0 & \text{if } v \text{ has a weak edge to } s. \end{cases} \tag{3.20}$$

Note that this specification is indeed exhaustive; by assumption, s and v are not locking, so we cannot have v strong up to s and $q_v^* = q_s^*$, nor can we have v strong down to s and $q_v^* + q_s^* = 1$.

As a side note, we remark that this perturbation has the appealing property that it maps to -1 times itself under a flipping of s (if a perturbation works for all almost attractive models, then the version obtained from it by flipping s must also work for all almost attractive models, since flipping s is a bijection from the set of all almost attractive models to itself).

Given the specified perturbation in Equation (3.20) for singleton marginals, we now exhibit the perturbation for edge marginals. We consider several different sub-cases, depending on the type of edge marginal present under q^*.

Strong edge perturbations. If an edge $uv \in E$ is strong (i.e. a LOC constraint is tight), we may immediately determine the perturbation on this edge required in order that this LOC constraint remains tight for both $q^* + \varepsilon\delta$ and $q^* - \varepsilon\delta$. Specifically, δ_{uv} is defined implicitly

by

$$q^*_{uv} + \varepsilon\delta_{uv} = \begin{cases} \min(q^*_u + \varepsilon\delta_u, q^*_v + \varepsilon\delta_v) & \text{if } uv \text{ is strong up} \\ \max(0, q^*_u + \varepsilon\delta_u + q^*_v + \varepsilon\delta_v - 1) & \text{if } uv \text{ is strong down}. \end{cases} \tag{3.21}$$

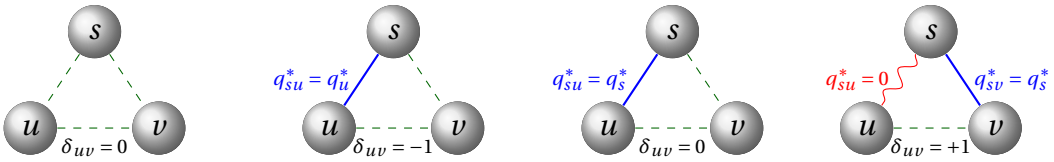
We observe that the relevant LOC constraint is indeed tight for $q^* \pm \varepsilon\delta$ by considering several subcases. First, if the edge is not locking and strong up, then the conclusion is immediate. Next, if the edge is not locking and strong down, then by considering which TRI constraints may be holding this edge down, we deduce $\delta_{uv} \in \{-1, 0, 1\}$, and again the conclusion is immediate. If the edge is locking down, then by Proposition 3.10, both u, v are locking with s, and hence $\delta_u = \delta_v = 0$, and so $\delta_{uv} = 0$ and the conclusion follows. If the edge is locking up, then considering tight TRI constraints for the triplet suv leads to $q^*_{su} = q^*_{sv}$, so that $\delta_u = \delta_v$ by the specification of singleton perturbations, and again the conclusion follows.

Weak edges incident to the locking component of s. If the edge uv is incident to the locking component of s (meaning that one of u and v lies in the same locking component of s), then the perturbation δ_{uv} is specified as follows, taking u to be locked to s without loss of generality:

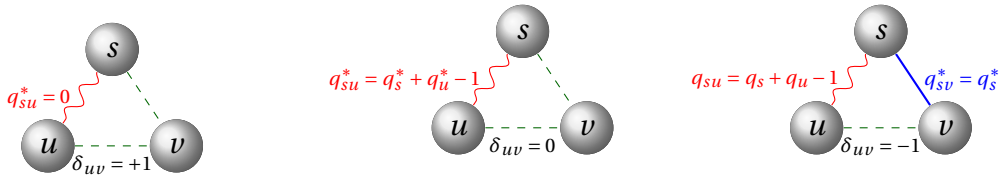
$$\delta_{uv} = \begin{cases} -1 & \text{if } u \text{ is locked up with } s, \text{ or } u = s \\ +1 & \text{if } u \text{ is locked down with } s. \end{cases} \tag{3.22}$$

Weak edges not incident to the locking component of s. Finally, given a weak edge uv such that neither u nor v is in the same locking component as s, this defines a unique triplet involving s: suv. We note that since uv is weak, there must be a tight TRI constraint arising from this triplet, preventing the edge from becoming strong, and it is this tight constraint which we must ensure remains tight under the specified perturbation; note that there cannot be more than one tight TRI constraint for the triplet suv, since then by Lemma 3.4, there would be a locking edge in the triangle.

We consider exhaustively the types of edges that may be present in the triangle suv such that uv is weak, and in each case deduce the TRI constraint that must be tight, and the required perturbation of the marginal on the weak edge uv. There are exactly 7 possible cases to consider, as shown in Figure 3.3.



(a) su, sv weak (b) $q^*_{su} = q^*_u$, sv weak (c) $q^*_{su} = q^*_s$, sv weak (d) $q^*_{su} = 0$ and $q^*_{sv} = q_s$



(e) $q^*_{su} = 0$, sv weak (f) $q^*_{su} = q^*_s + q^*_u - 1$, sv weak (g) $q^*_{su} = q^*_s + q^*_u - 1$ and $q^*_{sv} = q^*_v$

Figure 3.3. Cases where a weak edge uv is not incident to s. Weak edges are illustrated by black dashed lines, strong up edges are illustrated by solid blue lines, and strong down edges are illustrated by wavy red lines.

(a) $q^*_{su} = 0$ and $q^*_{sv} = q^*_v$ (b) $q^*_{su} = q^*_s + q^*_u - 1$ and $q^*_{sv} = q^*_s$

Figure 3.4. Cases which are not possible when suv has a tight TRI constraint, since each implies that uv is strong down.

Note that it is not possible for an edge uv away from s to be weak and be in a triangle of the form shown in Figure 3.4. For Figure 3.4a, note that the constraint $q_v + q_{su} \geq q_{sv} + q_{uv}$ forces $q^*_{uv} = 0$, so that the edge uv is strong down, not weak. Similarly, consider Figure 3.4b and note that the constraint $q_v + q_{su} \geq q_{sv} + q_{uv}$ forces $q^*_{uv} = q^*_u + q^*_v - 1$, so uv is strong down, not weak. Thus, these cases are omitted from Figure 3.3 and may be excluded from further analysis. Observe that a configuration of the form given in Figure 3.3g may be obtained by flipping the variable s in Figure 3.3d, and the configurations shown in Figures 3.3e and 3.3f may similarly be obtained from those in Figures 3.3b and 3.3c by flipping s. We may therefore exclude these cases from our analysis too, and need only show here that the perturbations defined for the weak edges in Figures 3.3a, 3.3b, 3.3c, and 3.3d are consistent.

The perturbations for the weak edge uv that are indicated in the various configurations of Figure 3.3 may be derived straightforwardly by considering the tight TRI constraint in

each case, using the prescribed perturbation for the other edges and singletons as defined above, and observing what perturbation of the weak edge is implied in order to maintain tightness of the relevant TRI constraint. We treat the cases one by one:

- In Figure 3.3a, the tight TRI constraint must be either $q^*_u + q^*_{sv} = q^*_{su} + q^*_{uv}$ or $q^*_v + q^*_{su} = q^*_{uv} + q^*_{sv}$. In either case, by noting that $\delta_u = \delta_v = 0$ and $\delta_{su} = \delta_{sv} = 0$, it follows that to maintain tightness of the TRI constraint, we must have $\delta_{uv} = 0$.
- In Figure 3.3b, the tight TRI constraint must be $q^*_u + q^*_{sv} = q^*_{su} + q^*_{uv}$. Noting that $\delta_u = -1$, $\delta_{sv} = 0$, and $\delta_{su} = \delta_u = -1$, it follows that to maintain tightness of the TRI constraint, we must have $\delta_{uv} = -1$.
- In Figure 3.3c, the tight TRI constraint must be $q^*_u + q^*_{sv} = q^*_{su} + q^*_{uv}$. Noting that $\delta_u = 1$, $\delta_{sv} = 0$ and $\delta_{su} = \delta_s = 0$, we must have $\delta_{uv} = 0$.
- In Figure 3.3d, the tight TRI constraint must be $q^*_v + q^*_{su} = q^*_{sv} + q^*_{uv}$. Noting that $\delta_v = 1$, $\delta_{su} = 0$, and $\delta_{sv} = \delta_s = 0$, we must have $\delta_{uv} = +1$.

This completes the specification of the perturbation δ. Having now fully specified the perturbed pseudomarginals $q^* \pm \varepsilon\delta$, we check that we do indeed have $q^* \pm \varepsilon\delta \in \mathbb{L}^s_3(G)|_{q_s=t}$. By construction, the singleton marginal for variable s in the perturbed pseudomarginals is still equal to x, so it remains to check that all the constraints of $\mathbb{L}^s_3(G)$ are satisfied for $q^* \pm \varepsilon\delta$.

3.5.6 Checking consistency

We remark that consistency of TRI constraints for triplets suv such that uv is weak and not incident to the locking component with s is guaranteed by construction of the perturbation δ_{uv}. Therefore, to deduce that $q^* \pm \varepsilon\delta \in \mathbb{L}^s_3(G)$, it only remains to verify the following three propositions, which deal with all other triplet constraints present in $\mathbb{L}^s_3(G)$. We first deal with triplets of variables incident to the 0-1-locking component in Proposition 3.11, since the rule for singleton marginal perturbations for variables corresponding to this locking component is different than for other locking components. We then deal with the remaining cases in Propositions 3.12 and 3.13.

Proposition 3.11 (**Consistency of TRI constraints for triplets incident to the** 0-1-**locking component**)**.** All TRI constraints of $\mathbb{L}^s_3(G)$ which are tight at q^* and arise from triplets suv for which at least one of u or v is in the 0-1-locking component, remain tight at $q^* \pm \varepsilon\delta$.

Proof. Without loss of generality, let u be in the 0-1-locking component, first with singleton marginal $q^*_u = 1$. We deal with the four possible tight TRI constraints in turn, noting that in all cases, we have $\delta_u = \delta_s = \delta_{su} = 0$ and $\delta_v = \delta_{uv}$.

- If $q^*_s + q^*_{uv} = q^*_{su} + q^*_{sv}$, then we note that $q^*_{sv} = q^*_v$, so $\delta_v = \delta_{sv}$, and so the TRI constraint is satisfied for $q^* \pm \varepsilon\delta$.
- If $q^*_u + q^*_{sv} = q^*_{su} + q^*_{uv}$, then we note that $q^*_{sv} = q^*_s + q^*_v - 1$, so $\delta_v = \delta_{sv}$, and so the TRI constraint is satisfied for $q^* \pm \varepsilon\delta$.
- If $q^*_v + q^*_{su} = q^*_{sv} + q^*_{uv}$, then we note that $q^*_{sv} = q^*_s$, so $\delta_{sv} = 0$, and so the TRI constraint is satisfied for $q^* \pm \varepsilon\delta$.
- Finally, if $q^*_{su} + q^*_{sv} + q^*_{uv} = q^*_s + q^*_u + q^*_v - 1$, then we note that $q^*_{sv} = 0$, so $\delta_{sv} = 0$, and so the TRI constraint is satisfied for $q^* \pm \varepsilon\delta$.

Next, considering $q^*_u = 0$, similarly we note that in all cases, we have $\delta_s = \delta_u = \delta_{su} = \delta_{uv} = 0$. Now, considering possible tight TRI constraints:

- If $q^*_s + q^*_{uv} = q^*_{su} + q^*_{sv}$, then we note that $q^*_{sv} = q^*_s$, so $\delta_{sv} = 0$, and so the TRI constraint is satisfied for $q^* \pm \varepsilon\delta$.
- If $q^*_u + q^*_{sv} = q^*_{su} + q^*_{uv}$, then we note that $q^*_{sv} = 0$, so $\delta_{sv} = 0$, and so the TRI constraint is satisfied for $q^* \pm \varepsilon\delta$.
- If $q^*_v + q^*_{su} = q^*_{sv} + q^*_{uv}$, then we note that $q^*_{sv} = q^*_v$, so $\delta_{sv} = \delta_v$, and so the TRI constraint is satisfied for $q^* \pm \varepsilon\delta$.
- Finally, if $q^*_{su} + q^*_{sv} + q^*_{uv} = q^*_s + q^*_u + q^*_v - 1$, then we note that $q^*_{sv} = q^*_s + q^*_v - 1$, so $\delta_{sv} = \delta_v$, and so the TRI constraint is satisfied for $q^* \pm \varepsilon\delta$. □

Proposition 3.12 (**Consistency of TRI constraints for triplets with weak edge incident to the locking component containing** s)**.** All TRI constraints of $\mathbb{L}^s_3(G)$ which are tight at q^* and arise from triplets suv for which (i) uv is weak in q^*, and (ii) one of u and v is contained in the locking component containing s, remain tight at $q^* \pm \varepsilon\delta$.

Proof. As uv is incident to the locking component containing s, we may take u to be locked with s, without loss of generality. Since uv is weak, v must lie in a different locking component. We consider the two cases where u is locked up to s and locked down to s separately.

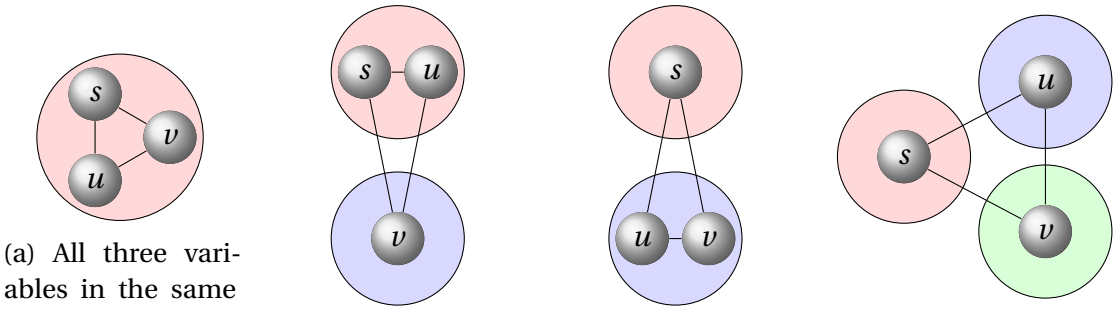
Firstly, when su is locked up (and hence $q^*_s = q^*_u = q^*_{su}$), we note that the TRI inequalities $q^*_u + q^*_{sv} \geq q^*_{su} + q^*_{uv}$ and $q^*_s + q^*_{uv} \geq q^*_{su} + q^*_{sv}$ imply that $q^*_{sv} = q^*_{uv}$. Thus, from the

specification of the perturbation in Section 3.5.5, we have $\delta_s = \delta_u = \delta_v = \delta_{su} = 0$, and $\delta_{sv} = \delta_{uv} = -1$, and so any tight TRI constraint for the triplet suv at q^* remains tight for $q^* \pm \varepsilon\delta$.

Secondly, when su is locked down (and hence $q^*_u = 1 - q^*_s$ and $q^*_{su} = 0$), we note that the TRI inequalities $q^*_v + q^*_{su} \geq q^*_{sv} + q^*_{uv}$ and $q^*_{su} + q^*_{sv} + q^*_{uv} \geq q^*_s + q^*_u + q^*_v - 1$ imply that $q^*_v = q^*_{uv} + q^*_{sv}$. From this, we deduce that uv weak implies sv weak, and so again we have $\delta_s = \delta_u = \delta_v = \delta_{su} = 0$, $\delta_{sv} = -1$, $\delta_{uv} = +1$, and so any tight TRI constraint for the triplet suv at q^* remains tight for $q^* \pm \varepsilon\delta$. □

Proposition 3.13 (Consistency of TRI constraints for triplets with strong edge away from s). All TRI constraints of $\mathbb{L}^s_3(G)$ which are tight at q^* and arise from triplets suv for which uv is strong in q^* remain tight at $q^* \pm \varepsilon\delta$.

Proof. We first enumerate several possibilities as to which locking components the variables u and v belong; these are summarised in Figure 3.5, considering each case in turn.



(a) All three variables in the same locking component.

(b) One variable in a locking component with s.

(c) Two variables in a locking component away from s.

(d) All variables in separate locking components.

Figure 3.5. Different cases of a strong edge uv away from s to be considered.

Case 3.5a. Since all variables are locking with s, we have $\delta_s = \delta_u = \delta_v = \delta_{su} = \delta_{sv} = \delta_{uv} = 0$, so any tight TRI constraint at q^* remains tight at $q^* \pm \varepsilon\delta$.

Case 3.5b. We consider the case where su is locked up and locked down separately. Firstly, if su is locked up, then the two tight TRI constraints must be $q^*_s + q^*_{uv} \geq q^*_{su} + q^*_{sv}$ and $q^*_u + q^*_{sv} \geq q^*_{su} + q^*_{uv}$, which imply that $q^*_{sv} = q^*_{uv}$. Hence, we have $\delta_{sv} = \delta_{uv}$, and so these two tight TRI constraints remain tight at $q^* \pm \varepsilon\delta$. Secondly, if su is locked down, then the two tight TRI constraints must be $q^*_v + q^*_{su} \geq q^*_{sv} + q^*_{uv}$ and $q^*_{su} + q^*_{sv} + q^*_{uv} \geq q^*_s + q^*_u + q^*_v - 1$, which imply that $q^*_v = q^*_{sv} + q^*_{uv}$. This yields the following possibilities for q^*_{sv} when uv is

strong:

$$q^*_{uv} = q^*_u \implies q^*_{sv} = q^*_s + q^*_v - 1, \quad q^*_{uv} = q^*_v \implies q^*_{sv} = 0,$$
$$q^*_{uv} = 0 \implies q^*_{sv} = q^*_v, \quad q^*_{uv} = q^*_u + q^*_v - 1 \implies q^*_{sv} = q^*_s. \tag{3.23}$$

In all four possible settings of the strong edges uv and sv, we recover consistency for the tight TRI constraints at $q^* \pm \varepsilon\delta$.

Case 3.5c. As u and v are in a locking component away from s, they must be locked up (by Proposition 3.10). Hence, the two tight TRI constraints are $q^*_u + q^*_{sv} \geq q^*_{su} + q^*_{uv}$ and $q^*_v + q^*_{su} \geq q^*_{sv} + q^*_{uv}$, which imply $q^*_{su} = q^*_{sv}$. Hence, we have $\delta_u = \delta_v$, and $\delta_{su} = \delta_{sv}$, so these TRI constraints remain tight at $q^* \pm \varepsilon\delta$.

Case 3.5d. Recall that Lemma 3.4 states that if two TRI constraints are tight in one triplet of variables, then there must be a locking edge. Since in this case there are no locking edges present in the triple suv, we may assume that there is *exactly* one tight TRI constraint in the triplet, and so need not check consistency for any other TRI constraints (since they are not tight).

First, we consider the cases where uv is strong down. Since it is an attractive edge, there must be a tight TRI constraint holding this pseudomarginal down, so we must have either $q^*_u + q^*_{sv} = q^*_{su} + q^*_{uv}$, or $q^*_v + q^*_{su} = q^*_{sv} + q^*_{uv}$, and by symmetry, it is sufficient to consider the case where the former holds. We thus examine the two possibilities for the pseudomarginal q^*_{uv}:

- If $q^*_{uv} = 0$, then from the tight TRI constraint $q^*_u + q^*_{sv} = q^*_{su} + q^*_{uv}$, we have $q^*_u + q^*_{sv} = q^*_{su}$. Since $q^*_{su} \leq q^*_u$, we deduce that $q^*_u = q^*_{su}$ and $q^*_{sv} = 0$, so su is strong up, and sv is strong down. Thus, from the perturbations specified above, we have $\delta_u = -1$, $\delta_{sv} = 0$, $\delta_{su} = -1$, and $\delta_{uv} = 0$. It is therefore the case that the tight TRI constraint remains tight at the perturbed pseudomarginals $q^* \pm \varepsilon\delta$.
- If $q^*_{uv} = q^*_u + q^*_v - 1$, then from the tight TRI constraint $q^*_u + q^*_{sv} = q^*_{su} + q^*_{uv}$, we have $q^*_{sv} = q^*_{su} + q^*_v - 1$. But since $q^*_{sv} \geq q^*_s + q^*_v - 1 \geq q^*_{su} + q^*_v - 1$, we deduce that $q^*_{su} = q^*_s$ and $q^*_{sv} = q^*_s + q^*_v - 1$. Thus, from the perturbations specified above, we have $\delta_u = +1$, $\delta_{sv} = -1$, $\delta_{su} = 0$, and $\delta_{uv} = 0$. It is therefore the case that the tight TRI constraint remains tight at the perturbed pseudomarginals $q^* \pm \varepsilon\delta$.

Having dealt with the cases where uv is strong down, we now deal with the cases where uv is strong up. By symmetry, it is sufficient to consider the case $q^*_{uv} = q^*_u$. We treat the four possible TRI constraints that could be tight in turn:

1. If the tight TRI constraint is $q^*_s + q^*_{uv} = q^*_{su} + q^*_{sv}$, then substituting in $q^*_{uv} = q^*_u$, we obtain $q^*_s + q^*_u = q^*_{su} + q^*_{sv}$. But since $q^*_s \geq q^*_{sv}$ and $q^*_u \geq q^*_{su}$, we must have $q^*_s = q^*_{sv}$ and $q^*_u = q^*_{su}$, so sv and su are strong up. Thus, from the perturbations specified above, we have $\delta_s = 0$, $\delta_{uv} = -1$, $\delta_{su} = -1$, and $\delta_{sv} = 0$. It is therefore the case that the tight TRI constraint remains tight at the perturbed pseudomarginals $q^* \pm \varepsilon\delta$.

2. If the tight TRI constraint is $q^*_{su} + q^*_{sv} + q^*_{uv} = q^*_s + q^*_u + q^*_v - 1$, then substituting in $q^*_{uv} = q^*_u$, we obtain $q^*_{su} = q^*_{sv} = q^*_s + q^*_v - 1$. But since $q^*_{sv} \geq q^*_s + q^*_v - 1$, we must have $q^*_{sv} = q^*_s + q^*_v - 1$ and $q^*_{su} = 0$, so su and sv are strong down. Thus, from the perturbations specified above, we have $\delta_s = 0$, $\delta_u = +1$, $\delta_v = -1$, $\delta_{su} = 0$, $\delta_{sv} = -1$ and $\delta_{uv} = +1$. It is therefore the case that the tight TRI constraint remains tight at the perturbed pseudomarginals $q^* \pm \varepsilon\delta$.

3. If the tight TRI constraint is $q^*_u + q^*_{sv} = q^*_{su} + q^*_{uv}$, we consider the possibilities for the edge su. Firstly, if it is strong up and $q^*_{su} = q^*_s$, then from the tight TRI inequality, we obtain $q^*_{sv} = q^*_s$ too. In this case, from the perturbation specified above, we have $\delta_u = +1$, $\delta_{sv} = 0$, $\delta_{su} = 0$ and $\delta_{uv} = +1$, so the tight TRI constraint remains tight at the perturbed pseudomarginals $q^* \pm \varepsilon\delta$. Secondly, if su is strong up and $q^*_{su} = q^*_u$, the tight TRI constraint yields $q^*_{sv} = q^*_u$. The only possibilities now for sv are that it is weak, or is strong down with $q^*_{sv} = q^*_s + q^*_v - 1$. In either case, the perturbations specified are $\delta_u = -1$,$\delta_{sv} = -1$, $\delta_{su} = -1$ and $\delta_{uv} = -1$, so the tight TRI constraint remains tight at the perturbed pseudomarginals $q^* \pm \varepsilon\delta$. Next, if su is weak, then the tight TRI constraint yields $q^*_{sv} = q^*_{su}$. Either sv is weak too, or is strong down with $q^*_{sv} = q^*_s + q^*_v - 1$. In either case, the perturbations specified are $\delta_u = -1$,$\delta_{sv} = -1$, $\delta_{su} = -1$ and $\delta_{uv} = -1$, so the tight TRI constraint remains tight at the perturbed pseudomarginals $q^* \pm \varepsilon\delta$. Next, if su is strong down with $q^*_{su} = q^*_s + q^*_u - 1$, then the tight TRI constraint yields $q^*_{sv} = q^*_s + q^*_u - 1$, contradicting the LOC constraint $q^*_{sv} \geq q^*_s + q^*_v - 1$, so we need not consider this case. Finally, if su is strong down with $q^*_{su} = 0$, the tight TRI constraint yields $q^*_{sv} = 0$. In this case, the perturbations specified are $\delta_u = -1$,$\delta_{sv} = 0$, $\delta_{su} = 0$ and $\delta_{uv} = -1$, so the tight TRI constraint remains tight at the perturbed pseudomarginals $q^* \pm \varepsilon\delta$.

4. If the tight TRI constraint is $q^*_v + q^*_{su} = q^*_{sv} + q^*_{uv}$, we consider the different possibilities for q^*_{sv}. Firstly, if $q^*_{sv} = 0$, then $q^*_{su} = 0 + u - v < 0$, a contradiction. If $q^*_{sv} = s + v - 1$, then from the tight TRI inequality, $q^*_{su} = q^*_s + q^*_u - 1$. Thus $\delta_v = \delta_{su} = \delta_{uv} = \delta_{sv} = -1$, and tightness of the TRI constraint is preserved under the perturbation. If $q^*_{sv} = q^*_v$, then the tight TRI inequality implies $q^*_{su} = q^*_u$, thus $\delta_v = \delta_{su} = \delta_{uv} = \delta_{sv} = -1$, and the tightness of the TRI constraint is preserved under the perturbation. If

$q^*_{sv} = q^*_s$, then the tight TRI inequality implies that either su is weak, and thus $\delta_v = +1$, $\delta_{su} = -1$, and $\delta_{uv} = \delta_{sv} = 0$, or $q^*_{su} = 0$, and thus $\delta_v = +1$, $\delta_{su} = \delta_{sv} = 0$, and $\delta_{uv} = +1$. In either case, the tightness of the TRI constraint is preserved under the perturbation. Finally, if sv is weak, then the tight TRI constraint implies that either su is weak, and thus $\delta_v = \delta_{uv} = 0$ and $\delta_{su} = \delta_{sv} = -1$, or $q^*_{su} = 0$, in which case $\delta_v = \delta_{su} = 0$ and $\delta_{sv} = -1$, $\delta_{uv} = +1$. In either case, the tightness of the TRI constraint is preserved under the perturbation. □

Thus, we have established that the perturbation δ defined in Section 3.5.5 is such that $q^* \pm \varepsilon\delta \in \mathbb{L}^s_3(G)|_{q_s=t}$. Thus, if δ is non-zero, then q^* cannot be extremal in the polytope $\mathbb{L}^s_3(G)|_{q_s=t}$. But δ is non-zero whenever there is a locking component distinct from the s-locking component and the 0-1-locking component. Therefore, only variables with singleton marginals in $\{0,1\}$ and variables locked to s can exist at an extremal point.

We thus obtain Theorem 3.6, from which in turn the linearity of $F^s_{\mathbb{L}^s_3(G)}$ follows from the discussion in Section 3.5.3, and hence Theorem 3.2 follows.

3.6 Discussion

We have analysed the tightness of LP relaxations for MAP inference in binary pairwise graphical models in model classes that generalise the classically understood balanced case. Theorem 3.2 on hybrid conditions (combining restrictions on topology and potentials) for tightness of $\mathbb{L}^s_3(G)$ is interesting for several reasons. It improves our understanding of why and when the relaxation will perform well. It supports the interesting characterisation of almost balanced models, which, to our knowledge, was not much considered prior to Weller (2015). Finally, it provides an important step into hybrid characterisations, combining conditions on potential type and model topology, as mentioned in Section 3.1. The core of the analysis relies of a combination of structural arguments, and the construction of a primal perturbation. An important question for future work is to what extent these approaches can be extended to Sherali–Adams polytopes with consistency conditions imposed over higher-order clusters, to make further progress in understanding the effectiveness of LP relaxations in practice. $\mathbb{L}_2(G)$ is tight for any balanced model and we now know that $\mathbb{L}^s_3(G)$ is tight for any almost balanced model; a natural next question is whether $\mathbb{L}_4(G)$, or some generalised Sherali–Adams polytope utilising 4-cluster constraints, is tight for any model that can be rendered balanced by deleting two variables.

It may be tempting to conjecture that if $\mathbb{L}_r(G)$ is tight over a model class for some r, then if an extra variable is added with arbitrary interactions, $\mathbb{L}_{r+1}(G)$ will be tight on the larger model. However, this is false. Consider the class of planar binary pairwise model with no singleton potentials and only pure 2-potentials. The triplet-consistent relaxation $\mathbb{L}_3(G)$ is tight for such models (Barahona, 1983); yet if one adds a new variable connected to all of the original ones, one obtains a model which is an uprooting (see Chapter 2) of a general planar binary pairwise model. Since MAP inference on an uprooted model is known to be equivalent to that on the original model (see Lemma 2.7), and MAP inference on planar binary pairwise models (with singleton and pairwise potentials) is known to be NP-hard (Barahona, 1982), this newly obtained class of models must be NP-hard, and therefore not solvable under $\mathbb{L}_4(G)$.

Chapter 4

Conditions Beyond Treewidth for Tightness of LP Relaxations

This chapter is based on results from the following publication:

- Rowland, M., Pacchiano, A., and Weller, A. (2017). Conditions beyond treewidth for tightness of higher-order LP relaxations. In *Artifical Intelligence and Statistics (AISTATS)*.

Our primary contribution in this chapter is to examine the sufficient condition given in Theorem 1.11 for tightness of Sherali–Adams relaxations, in terms of treewidth of the underlying graph, and to understand whether or not this condition is necessary. Along the way, we reprove some classical results with a new approach. Work was carried out jointly with Adrian Weller and Aldo Pacchiano. Many sections of the original paper have been rewritten for the purposes of this thesis.

4.1 Introduction

As with Chapter 3, the focus of this chapter is on theoretical understanding of tightness of Sherali–Adams relaxations for *pairwise* graphical models. Wainwright and Jordan (2004) showed that the Sherali–Adams relaxation $\mathbb{L}_r(G)$ (for $r \geq 2$) is exact for any model over a graph G with treewidth $\leq r-1$ (see Theorem 1.11). A natural question to ask is whether the converse also holds; given a graph G with treewidth $> r-1$, is there a model over G such that the relaxation $\mathbb{L}_r(G)$ is *not* tight for this model? In other words, is the graph condition "G has treewidth $\leq r-1$" necessary and sufficient for the property $\mathbb{L}_r(G) = \mathbb{M}(G)$? Weller (2016a) studied this question specifically for $\mathbb{L}_3(G)$ in the case of binary pairwise models;

using connections to graph minor theory and characterisations of weakly bipartite graphs, it was shown that this characterisation *does* indeed hold for $\mathbb{L}_3(G)$ and this class of models. Here, we build on the approach of Weller (2016a), and study the corresponding question for the quadruplet-consistent polytope, $\mathbb{L}_4(G)$, showing that the converse result in fact breaks down. Whereas Weller (2016a) made extensive use of powerful earlier results in combinatorics, including two results which won the Fulkerson prize (Guenin, 2001; Lehman, 1990), our analysis takes a different, geometric approach (developed in Sections 4.3 and 4.4), which may be of independent interest.

We conclude this section by emphasising the following contributions:

- In Section 4.2: Background graph-theoretic material, drawing on Weller (2016a), concluding with the statement of our main original result, Theorem 4.1, which strengthens tightness guarantees for $\mathbb{L}_4(G)$ due to Wainwright and Jordan (2004), via graph minor theory.
- In Section 4.3: Characterisations of tight models for Sherali–Adams relaxations based on polyhedral geometry, and new proofs of the earlier stated Theorems 1.8 and 1.14, characterising tightness of $\mathbb{L}_2(G)$ for models on trees and balanced models.
- In Sections 4.4 and Appendix Section 4.A: Proof of Theorem 4.1.

4.2 Graph minors and conditions for tightness

In this section, we review existing results on tightness of Sherali–Adams relaxations in terms of graph minor theory, and state our main theorem. Intuitively speaking, the broad approach of analysing Sherali–Adams relaxations in this way is to make statements such as "if a graph G does not have any bad substructures, then the relaxation $\mathbb{L}_r(G)$ is guaranteed to be tight for all models on G". Of course, without precise definitions of *substructures* or their *badness*, this does not yet mean anything. We now review the relevant concepts needed from graph theory, so as to be able to make precise versions of statements such as the one above. This background is provided in Sections 4.2.1 to 4.2.3, and we then state our main results in Section 4.2.4.

4.2.1 Graph minor theory

We begin with a brief review of the graph minor theory required to make the statement of our main result precise; for further background, see Diestel, 2010, Chapter 12. Given a graph $G = (V, E)$, a graph H is a *minor* of G (written $H \leq G$) if it can be obtained from G via a series of edge deletions, vertex deletions, and edge *contractions*. The result of contracting an edge $uv \in E$ is the graph $G' = (V', E')$ where u and v are 'merged' to form a new vertex w which is adjacent to any vertex that was previously adjacent to either u or v. That is, $V' = V \setminus \{u, v\} \cup \{w\}$, and $E' = \{e \in E \mid u, v \notin e\} \cup \{wx \mid ux \in E \text{ or } vx \in E\}$. This is illustrated in Figure 4.1.

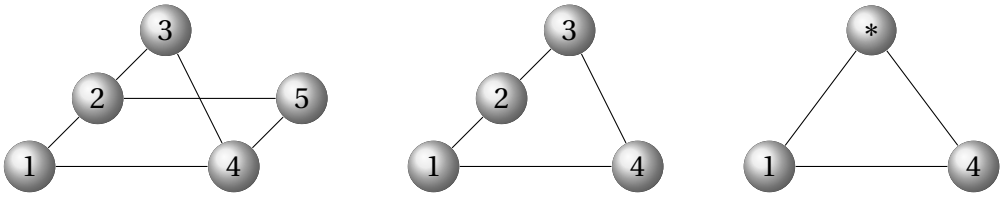


Figure 4.1. The left graph has a K_3 minor, which may be exhibited by first deleting the vertex 5 (resulting in the central graph), and then contracting the edge between vertices 2 and 3 to create a new vertex $*$ (resulting in the right graph).

A property of graphs is *closed under taking minors* or *minor-closed* if whenever G has the property and $H \leq G$, then H also has the property. The *graph minor theorem* of Robertson and Seymour (2004) proves that any minor-closed graph property may be characterised by a unique finite set $\{H_1, \ldots, H_m\}$ of *minimal forbidden minors*. The term *minimal* here is with respect to the partial order $\leq$ on graphs described above. That is, a graph G has the property iff it does not contain any H_i as a minor. Checking to see if a graph contains some fixed H as a minor may be performed in polynomial time (Robertson and Seymour, 1995). A classical example of a minor-closed property, and the corresponding set of minimal forbidden minors, is given by Wagner's theorem, which states that a graph is planar if and only if it does not contain K_5 (the complete graph on five vertices) or $K_{3,3}$ (the complete bipartite graph on two sets of three vertices) as a minor.

4.2.2 Treewidth and tightness as minor-closed properties

The graph property of *having treewidth* $\leq r-1$ is closed under taking minors Robertson and Seymour (1986), and hence by graph minor theorem, this property may be characterised by forbidding a unique finite set of minimal *forbidden minors*. In other words, of all the

graphs with treewidth $\geq r$, there is a unique finite set T_r of graphs which are minimal with respect to minor operations.

Hence, the sufficient condition of Wainwright and Jordan (2004) for tightness of $\mathbb{L}_r(G)$ may be reframed as the following statement: if a graph G does not contain any member of the set T_r as a minor, then $\mathbb{L}_r(G)$ is guaranteed to be tight for all models over G.

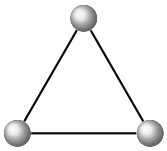


Figure 4.2. K_3, the only element of T_2, the set of minimal forbidden minors for treewidth ≤ 1.

The relevant sets of forbidden minors for treewidth ≤ 1 and ≤ 2 are particularly simple with just one member each: $T_2 = \{K_3\}$ and $T_3 = \{K_4\}$ (Bodlaender, 1998), where K_n is the complete graph on n vertices. For higher values of r, T_r always contains K_r but there are also other forbidden minors, and their number grows rapidly; T_4 has 4 members (Arnborg et al., 1990) while T_5 has over 70 (Sanders, 1993). The sets of forbidden minors for treewidth 1, ≤ 2 and ≤ 3 are illustrated in Figures 4.2, 4.3, and 4.4, respectively.

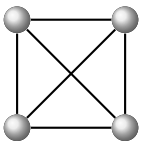
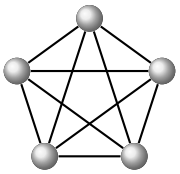
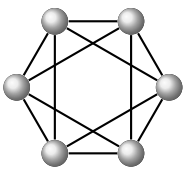


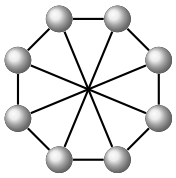
Figure 4.3. K_4, the only element of T_3, the set of minimal forbidden minors for treewidth ≤ 2.



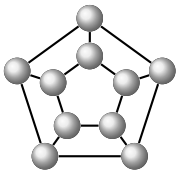
(a) K_5



(b) Octahedral graph



(c) Wagner graph



(d) Pentagonal prism graph

Figure 4.4. The four members of T_4, the set of minimal forbidden minors (unsigned) for treewidth ≤ 3.

Weller (2016a) showed that, for any r, the graph property of $\mathbb{L}_r(G)$ *being tight for all models on the graph* G is also closed under taking minors. Hence, by the graph minor theorem, this property may be characterised by forbidding a unique set of minimal forbidden minors, which we denote by U_r. It was shown that, in fact, $U_2 = T_2 = \{K_3\}$ and $U_3 = T_3 = \{K_4\}$. However, until this work, all that was known about U_4 is that it contains the complete graph K_5; it has been an open question as to whether $U_4 = T_4$.

One main contribution here is to show that $U_4 \neq T_4$. Indeed, we argue that $U_4 \cap T_4 = \{K_5\}$ and that U_4 must contain at least one other forbidden minor, which we cannot yet identify. This progress on understanding U_4 is a significant theoretical development, demonstrating

that in general, treewidth is not precisely the right way to characterise tightness of LP relaxations at a granular level.

4.2.3 Signed graphs and minors

Whilst the results described in Section 4.2.2 describe worst-case tightness results in terms of graph topology, low levels of the Sherali–Adams hierarchy are tight for many high-treewidth models arising in practice. This motivates an analysis that takes into account not only model topology, but also the types of potentials that may be present in a model. As an example, the results described in the previous section tell us that for any graph G with a K_3 minor (equivalent to having treewidth greater than 1), there is a model on the graph for which $\mathbb{L}_2(G)$ is not tight. However, Theorem 1.14 tells us that $\mathbb{L}_2(G)$ is tight for any binary pairwise graphical model with balanced potentials, regardless of the treewidth of G. Thus, in some sense it is not the presence of K_3 as a minor *per se* that prevents $\mathbb{L}_2(G)$ from being tight on graphs which are not trees, but the presence of the K_3 minor in combination with the types of potentials present in the model. Theorem 1.14 motivates us to take the attractiveness/repulsiveness of pairwise potentials into account in our analysis, and as Weller (2016a) showed, signed graphs are a natural way in which to do this.

A *signed graph* is a graph $G = (V, E)$ together with a function $\Sigma : E \to \{\text{even}, \text{odd}\}$ that assigns a *sign* to each edge of the graph. We may associate a signed graph (G, Σ) with a class of models $\mathcal{M}(G, \Sigma)$ on G by restricting pairwise potentials to agree with the signs dictated by Σ. More precisely, a model $M = (G, (\theta_{\mathcal{E}} | \mathcal{E} \in E))$ is in the class $\mathcal{M}(G, \Sigma)$ if for all edges $\mathcal{E} \in E$, we have

$$\begin{aligned} \Sigma(\mathcal{E}) = \text{ even } &\implies \theta_{\mathcal{E}} \text{ is attractive,} \\ \Sigma(\mathcal{E}) = \text{ odd } &\implies \theta_{\mathcal{E}} \text{ is repulsive.} \end{aligned} \tag{4.1}$$

We say that $\mathbb{L}_r(G)$ is tight for a signing Σ of G if $\mathbb{L}_r(G)$ is tight for every model in $\mathcal{M}(G, \Sigma)$. Weller (2016a) also showed that the property of a signed graph (G, Σ) that $\mathbb{L}_r(G)$ is tight for all models in the class $\mathcal{M}(G, \Sigma)$ can also be characterised by forbidding particular substructures, in a sense that we now review in more detail.

A *signed minor* of a signed graph is obtained by edge deletion (of any edge), vertex deletion, and edge contraction specifically along *even* edges. Further, any *resigning* operation is also allowed, in which a subset of vertices $S \subseteq V$ is selected and then all edges with exactly one end in S are flipped even $\leftrightarrow$ odd. Hence, to contract an odd edge, one may first flip

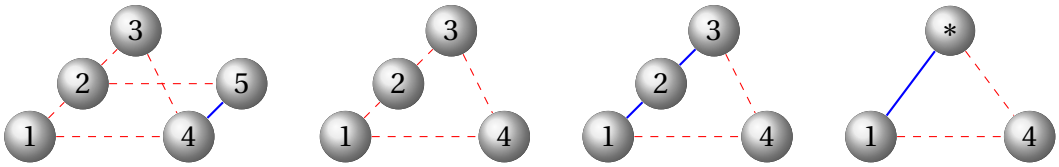


Figure 4.5. The far-right graph is a signed minor of the far-left graph, which is exhibited by deleting vertex 5 (mid-left), flipping vertex 2 (mid-right), and then contracting along the even edge between 2 and 3.

either end of the edge to make it even, then contract. See Figure 4.5 for an illustration of the process of taking a signed minor. In this illustration and all that follow, an odd edge is illustrated by a dashed red line, and an even edge by a solid blue line. Incidentally, this notion of flipping is closely related to the notion of flipping of potentials introduced in Section 1.4.1. Indeed, if a model belonging to a class $\mathcal{M}(G,\Sigma)$ has a subset of vertices $S \subseteq V$ flipped, the new flipped model belongs to the class $\mathcal{M}(G,\Sigma^{(S)})$, where $\Sigma^{(S)}$ is the signed graph obtained from (G,Σ) by flipping the subset S of variables.

The graph minor theorem of Robertson and Seymour generalises to signed graphs (Geelen et al., 2014; Huynh, 2009): any property of signed graphs which is closed under taking signed minors may be characterised by a unique finite set of minimal forbidden signed minors. We note that Watanabe (2011) used the theory of signed graph minors to understand when belief propagation algorithms are guaranteed to have unique fixed points.

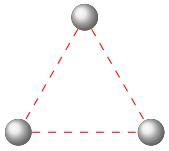


Figure 4.6. Odd-K_3, the unique element of U'_2, the set of minimal forbidden signed minors for tightness of $\mathbb{L}_2(G)$.

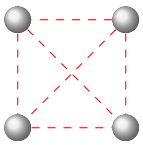


Figure 4.7. Odd-K_4, the unique element of U'_3, the set of minimal forbidden signed minors for tightness of $\mathbb{L}_3(G)$.

The property of a signed graph (G,Σ) that $\mathbb{L}_r(G)$ is tight for Σ is closed under taking signed minors. There is therefore a finite set of minimal forbidden signed minors, U'_r, such that if a signed graph (G,Σ) does not have any member of U'_r as a signed minor, then $\mathbb{L}_r(G)$ is tight for any model M on G respecting Σ. Weller (2016a) showed that U'_2 consists of a single signed minor (the odd-K_3 — a K_3 in which all edges are declared odd), and that U'_3 consists of a single signed minor (the odd-K_4); see Figures 4.6 and 4.7 for illustrations of these signed minors.

4.2.4 Statement of main results

With the necessary background established in the previous sections, we now state our main results.

Theorem 4.1. The only non-tight signing for $\mathbb{L}_4(G)$ of any graph G in the set T_4 (the minimal forbidden minors for treewidth ≤ 3) is the odd-K_5.

The minimal forbidden signed minor concerned is illustrated in Figure 4.8.

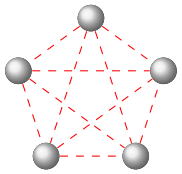


Figure 4.8. Odd-K_5, the unique signing of an element of T_4, the set of minimal forbidden unsigned minors for treewidth ≤ 3, that appears in U_4', the set of minimal forbidden signed minors for $\mathbb{L}_4(G)$, as shown by Theorem 4.1.

This result is perhaps surprising, as it shows that the exact correspondence between the graph properties "treewidth $\leq r$" and exactness of the Sherali–Adams relaxation $\mathbb{L}_{r+1}(G)$, which holds for $r = 1, 2$, breaks down for $r = 3$. This first result, stated in terms of forbidden signed minors, immediately implies the following weaker result regarding unsigned graph minors.

Corollary 4.2. Of the four minimal forbidden signed minors characterising the graph property of *having treewidth* ≤ 3 (see Figure 4.4), the only one which is a minimal forbidden minor for $\mathbb{L}_4(G) = \mathbb{M}(G)$ is K_5.

A natural conjecture for $\mathbb{L}_4(G)$ was that one must forbid just some signings of the four graphs in T_4; see Figure 4.4. Now given Theorem 4.1, it would seem sensible to conjecture that $\mathbb{L}_4(G)$ is tight for all valid models over a signed graph (G, Σ) iff the signed graph does not contain an odd-K_5 as a signed minor. However, this must be false (unless P=NP), since if it were true, we would have that $\mathbb{L}_4(G)$ is tight for any model over a graph G not containing K_5 (as an unsigned minor). As mentioned in Section 4.2, Wagner's theorem characterises planar graphs as those without K_5 or $K_{3,3}$ as a minor. Planar graphs are therefore a subclass of graphs which are K_5-free. Hence, we would have a polynomial time method to solve MAP inference for any planar binary pairwise model. Yet it is not hard to see that we may encode minimal vertex cover in such a model, and it is known that planar minimum vertex cover is NP-hard (Lichtenstein, 1982).

Having stated our main results, we now develop several geometric tools that will be useful in obtaining their proofs; these tools are described in Section 4.3, and the proofs are then given in Section 4.4 and in the Appendix Section 4.A.

4.3 The geometry of Sherali–Adams relaxations

Here we introduce several geometric notions for the Sherali–Adams polytopes which we shall apply in Section 4.4 in the proof of Theorem 4.1. In what follows, we will use the minimal representation of the Sherali–Adams polytopes described in Section 3.3, so that a configuration of pseudomarginals $((\mu_v | v \in V), (\mu_{\mathcal{E}} | \mathcal{E} \in E)) \in \mathbb{L}_r(G)$ is represented as $((q_v | v \in V), (q_{ij} | ij \in E))$, where q_v represents the marginal probability of the event $\{X_v = 1\}$ for each vertex $v \in V$, and q_{ij} represents the marginal probability of the event $\{X_i = 1, X_j = 1\}$ for each edge $ij \in E$. We shall also use this minimal representation for the marginal polytope. As in earlier sections, to lighten notation, we will identify $\mathbb{L}_r(G)$ (and $\mathbb{M}(G)$) with its image under this parametrisation, and hence will write statements such as $((q_v | v \in V), (q_{ij} | ij \in E)) \in \mathbb{L}_r(G)$ without further comment.

4.3.1 The geometry of the Sherali–Adams polytopes

The study of tightness of LP relaxations is naturally expressed in the language of polyhedral combinatorics. We introduce key notions from the literature, which will be used throughout the chapter, then provide new proofs of characterisations of tightness for the local polytope $\mathbb{L}_2(G)$ with these geometric ideas.

Given a polytope $\Pi \subset \mathbb{R}^m$ and an extremal point (vertex) $z \in \mathrm{Ext}(\Pi)$, the *normal cone* to Π at z, denoted $N_\Pi(z)$, is the polyhedral cone defined by

$$N_\Pi(z) = \left\{ c \in \mathbb{R}^m \,\middle|\, z \in \arg\max_{y \in \Pi} \langle c, y \rangle \right\}. \tag{4.2}$$

We define the *conical hull* of a finite set $X \subset \mathbb{R}^m$ by $\mathrm{Cone}(X) = \{\sum_{x \in X} \lambda_x x \mid \lambda_x \geq 0 \ \forall x \in X\}$. The following characterisation of normal cones will be particularly useful.

Lemma 4.3 (Theorem 2.4.9, Schneider, 1993)**.** Let $\Pi = \{y \in \mathbb{R}^m | Ay \leq b\}$ be a polytope for some $A = [a_1, \ldots, a_k]^\top \in \mathbb{R}^{k \times m}$, $b \in \mathbb{R}^k$ (for some $k \in \mathbb{N}$). Then for $z \in \text{Ext}(\Pi)$, we have

$$N_\Pi(z) = \text{Cone}(\{a_i | \langle a_i, z \rangle = b_i\}). \tag{4.3}$$

Further, if the representation $\{x \in \mathbb{R}^m | Ax \leq b\}$ has no redundant constraints, then $\{a_i | \langle a_i, z \rangle = b_i\}$ is a complete set of extremal rays of $N_\Pi(z)$ (up to scalar multiplication).

With these geometric notions in hand, we have a succinct, albeit indirect, characterisation of the set of potentials for which a given Sherali–Adams relaxation is tight.

Lemma 4.4. The set of potentials which are tight with respect to $\mathbb{L}_r(G)$ is exactly given by the following union of cones:

$$\bigcup_{z \in \text{Ext}(\mathbb{M}(G))} N_{\mathbb{L}_r(G)}(z)\,. \tag{4.4}$$

The proof of this characterisation is immediate from the definition of normal cones in Equation (4.2). This concise characterisation, together with the explicit parametrisation of normal cones given by Lemma 4.3 and the form of the linear constraints defining the local polytope $\mathbb{L}_2(G)$ given in Section 3.3, yields an efficient algorithm for generating arbitrary potentials which are tight with respect to $\mathbb{L}_2(G)$. This formulation may also be used to identify classes of potentials for which $\mathbb{L}_2(G)$ is guaranteed to be tight. We demonstrate the power of our geometric approach by providing new proofs in Section 4.A.1 of the following well-known results, stated as Theorems 1.8 and 1.14 earlier in this thesis. We restate these results here for clarity.

Lemma 4.5. If G is a tree, then $\mathbb{L}_2(G)$ is tight for all potentials, i.e. $\mathbb{L}_2(G) = \mathbb{M}(G)$.

Lemma 4.6. For an arbitrary graph G, $\mathbb{L}_2(G)$ is tight for the set of balanced models on G.

The proofs proceed by explicitly demonstrating that a given potential lies in a cone $N_{\mathbb{L}_2(G)}(z)$ for some vertex z of the marginal polytope, by expressing the potential as a conical combination of extremal rays of the cone.

4.3.2 The symmetry of the Sherali–Adams polytopes

The Sherali–Adams polytopes have rich symmetries which can be exploited when classifying tightness of LP relaxations using the tools discussed in Section 4.3.1. Intuitively, these symmetries arise either by considering relabellings of the vertices of the graph G

(permutations), or relabellings of the state space of individual variables (flippings). The key result is that a Sherali–Adams polytope is tight for a collection of potentials ϕ iff it is tight for any permutation or flipping of ϕ; we use this in Section 4.4 to reduce the number of linear programs that must be checked to verify Theorem 4.1.

The permutation group. Let $\sigma \in \mathrm{Aut}(G) \leq \mathcal{S}_V$ be a relabelling of the vertices of the graph G, where $\mathcal{S}_V$ is the symmetric group on the set V, and $\mathrm{Aut}(G)$ is the automorphism group of G, defined by

$$\mathrm{Aut}(G) = \left\{\sigma \in \mathcal{S}_V \,|\, ij \in E \iff \sigma(i)\sigma(j) \in E \quad \text{for all } i, j \in V\right\}. \tag{4.5}$$

The permutation σ then induces an affine bijective map $Y_\sigma : \mathbb{L}_r(G) \to \mathbb{L}_r(G)$ (which naturally lifts to an affine bijective map on $\mathbb{R}^{V \cup E}$), given by applying the corresponding relabelling to the components of the pseudomarginal vectors:

$$Y_\sigma((q_i)_{i\in V}, (q_{ij})_{ij\in E}) = ((q_{\sigma(i)})_{i\in V}, (q_{\sigma(i)\sigma(j)})_{ij\in E}), \qquad \forall ((q_i)_{i\in V}, (q_{ij})_{ij\in E}) \in \mathbb{L}_r(G). \tag{4.6}$$

The element $\sigma \in \mathrm{Aut}(G)$ also naturally induces a linear map on the space of potentials, which is formally the dual space $(\mathbb{R}^{V\cup E})^\star$, although we will frequently simply write $\mathbb{R}^{V\cup E}$. We denote the map on the space of potentials by $Y_\sigma^\dagger : \mathbb{R}^{V\cup E} \to \mathbb{R}^{V\cup E}$, given by

$$Y_\sigma^\dagger((\phi_i)_{i\in V}, (\phi_{ij})_{ij\in E}) = ((\phi_{\sigma(i)})_{i\in V}, (\phi_{\sigma(i)\sigma(j)})_{ij\in E}), \quad \forall ((\phi_i)_{i\in V}, (\phi_{ij})_{ij\in E}) \in \mathbb{R}^{V\cup E}. \tag{4.7}$$

The sets $\{Y_\sigma | \sigma \in \mathrm{Aut}(G)\}$ and $\{Y_\sigma^\dagger | \sigma \in \mathrm{Aut}(G)\}$ obey the group axioms (under the operation of composition), and hence form groups of symmetries on $\mathbb{L}_r(G)$ and $\mathbb{R}^{V\cup E}$ respectively; they are both naturally isomorphic to $\mathrm{Aut}(G)$.

These symmetry groups provide a useful formalism for thinking about tightness of Sherali–Adams relaxations. We provide one such result in this language, with proof in Section 4.A.2.

Lemma 4.7. $\mathbb{L}_r(G)$ is tight for a given potential $\phi \in \mathbb{R}^{V\cup E}$ iff it is tight for all potentials $Y_\sigma^\dagger(\phi)$, $\sigma \in \mathrm{Aut}(G)$.

The flipping group. Whilst the permutation group described above corresponds to permuting the labels of vertices in the graph, it is also useful to consider the effect of permuting the labels of the states of individual variables. In the case of binary models, this is precisely the notion of flipping described in Section 1.4.1. In this chapter, we denote

the transformation of pseudomarginals corresponding to flipping the variable associated with vertex $v \in V$ by $F_{(v)} : \mathbb{L}_r(G) \to \mathbb{L}_r(G)$. These flipping maps commute and have order 2, hence the group generated by them, $\langle F_{(v)} | v \in V \rangle$, is isomorphic to $\mathbb{Z}_2^V$. A general element of this group can be thought of as simultaneously flipping a subset $I \subseteq V$ of variables, written as $F_{(I)} : \mathbb{L}_r(G) \to \mathbb{L}_r(G)$. Flipping a subset of variables $I \subseteq V$ also naturally induces a map $F_{(I)}^\dagger : \mathbb{R}^{V \cup E} \to \mathbb{R}^{V \cup E}$ on the space of potentials. We have already seen in Proposition 1.19 that an analogous result to Lemma 4.7 also holds for the group of flipping symmetries.

The joint symmetry group of the Sherali–Adams polytopes. Tying together the remarks on the permutation group and flipping group above, we observe that in general the symmetries of the flipping and permutation groups on $\mathbb{L}_r(G)$ do not commute. In fact, observe that

$$Y_\sigma^{-1} \circ F_{(I)} \circ Y_\sigma = F_{(\sigma^{-1}(I))}, \qquad \text{for all } \sigma \in \mathrm{Aut}(G), I \subseteq V. \tag{4.8}$$

Thus the group of symmetries of $\mathbb{L}_r(G)$ generated by permutations and flippings is isomorphic to a semidirect product $\mathrm{Aut}(G) \rtimes \mathbb{Z}_2^V$.

4.4 Identifying forbidden signed minors

In this section, we set out the high-level approach to establishing Theorem 4.1. First, note that $\mathbb{L}_4(G)$ is tight for a potential $\phi \in \mathbb{R}^{V \cup E}$ if and only if

$$\max_{q \in \mathbb{L}_4(G)} \langle \phi, q \rangle = \max_{z \in \mathbb{M}(G)} \langle \phi, z \rangle, \tag{4.9}$$

or equivalently if and only if

$$\max_{q \in \mathbb{L}_4(G)} \min_{z \in \mathbb{M}(G)} \left[\langle \phi, q \rangle - \langle \phi, z \rangle \right] = 0. \tag{4.10}$$

Since $\mathbb{M}(G) \subseteq \mathbb{L}_4(G)$, we have $\max_{q \in \mathbb{L}_4(G)} \langle \phi, q \rangle \geq \max_{z \in \mathbb{M}(G)} \langle \phi, z \rangle \ \forall \phi \in \mathbb{R}^{V \cup E}$. Hence, it follows that $\mathbb{L}_4(G)$ is not tight for some potential $\phi \in \mathcal{M}(G, \Sigma)$ (the set of potentials respecting a signing Σ of G, see Section 4.2.3) iff the following optimisation problem has a non-zero optimum:

$$\max_{\phi \in \mathcal{M}(G,\Sigma)} \max_{q \in \mathbb{L}_4(G)} \min_{z \in \mathbb{M}(G)} \left[\langle \phi, q \rangle - \langle \phi, z \rangle \right]. \tag{4.11}$$

For the graphs in T_4, this is a high-dimensional indefinite quadratic program which is intractable to solve. However, using the geometric ideas of Section 4.3, we may decompose this problem into a collection of tractable linear programs.

4.4.1 Linearising the problem

We first convert this indefinite quadratic program over a conic region to an optimisation problem over a polytope, by intersecting the feasible region with some convex polytope containing the origin in its interior:

Lemma 4.8. Problem (4.11) has a non-zero optimum iff

$$\max_{\phi \in \mathcal{M}(G,\Sigma) \cap \Pi} \max_{q \in \mathbb{L}_4(G)} \min_{z \in \mathbb{M}(G)} \left[\langle \phi, q \rangle - \langle \phi, z \rangle \right] \tag{4.12}$$

has a non-zero optimum, where Π is any polytope containing the origin in its interior.

Problem (4.12) is a non-definite quadratic program, and so is difficult to solve directly. Instead, we convert it into a tractable number of linear programs, using the geometric constructions introduced in Section 4.3, as follows.

Lemma 4.9. Problem (4.12) has a non-zero optimum iff at least one of the following set of problems (indexed by $z \in \mathrm{Ext}(\mathbb{M}(G))$) has a non-zero optimum:

$$P'_z : \quad \max_{\phi \in \mathcal{M}(G,\Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z)} \max_{q \in \mathbb{L}_4(G)} \left[\langle \phi, q \rangle - \langle \phi, z \rangle \right]. \tag{4.13}$$

The problem P'_z in Expression (4.13) still has a quadratic objective, but has some particularly desirable properties that will allow us to make progress. Firstly, it is now purely a maximisation problem. Secondly, the region of optimisation is the direct product of two polytopes, and the only quadratic terms in the objective are "cross-terms" between variables in these two polytopes.

Lemma 4.10. An optimum $(\phi^\star, q^\star)$ of P'_z (see Expression (4.13)) occurs with $\phi^\star$ (resp. $q^\star$) extremal in $\mathcal{M}(G,\Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z)$ (resp. $\mathbb{L}_4(G)$).

The above lemma tells us that to establish whether P'_z has a non-zero optimum, it suffices to check whether any of the following linear programs $P''_{z,\phi}$ have a non-zero optimum:

$$P''_{z,\phi} : \quad \max_{q \in \mathbb{L}_4(G)} \left[\langle \phi, q \rangle - \langle \phi, z \rangle \right], \tag{4.14}$$

for each $z \in \text{Ext}(\mathbb{M}(G))$, and for each $\phi \in \text{Ext}(\mathcal{M}(G,\Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z))$. Each problem $P''_{z,\phi}$ is a linear program over $\mathbb{L}_4(G)$, and is therefore efficiently solvable.

Note that the number of problems $P''_{z,\phi}$ we have to solve scales with the size of the set $\text{Ext}(\mathcal{M}(G,\Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z))$, and so for computational reasons it is preferable to select the polytope Π containing the origin in its interior so that the polytope $\mathcal{M}(G,\Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z)$ has as few vertices as possible. One way in which to do this is to consider a H-representation for $N_{\mathbb{M}(G)}(z)$ as $\{\phi \in \mathbb{R}^{V \cup E} | A\phi \geq \mathbf{0}\}$ for some matrix A. Note that as $N_{\mathbb{M}(G)}(z)$ is a cone based at the origin, all halfspaces intersect the origin, so we may conclude that the linear inequalities defining its H-representation are all of the form $A_i x \geq 0$, where A_i is the i^{th} row of A. We may then enforce the additional linear constraint on ϕ that $\mathbf{1}^\top A\phi \leq 1$, where $\mathbf{1}$ is a vector of ones. This ensures that we obtain a polytope with number of vertices equal to the number of unique (up to scalar multiplication) extremal rays of the cone (plus an additional vertex at the origin).

Having arrived at the collection of linear programs of the form in Problem (4.14) to be solved in order to verify Theorem 4.1, we now utilise the notions of symmetry of Sherali–Adams relaxations discussed in Section 4.3.2 to reduce the number of linear programs that must be checked.

4.4.2 Exploiting symmetries of the marginal polytope and Sherali–Adams relaxations

Whilst we have reduced Problem (4.12) to finding the greatest optimal value in a set of linear programs described in Problem (4.14), we note that a priori there are a vast number of such LPs that must be solved. Given a graph $G = (V,E)$, there are $2^{|E|}$ possible signings of the graph, whilst there are $2^{|V|}$ vertices of the marginal polytope, and therefore in general many vertices of the polytopes $\mathcal{M}(G,\Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z)$ to consider. Therefore, if possible, we would like to make arguments to reduce the number of linear programs that must be solved to establish whether or not Problem (4.12) has a non-zero optimum. Fortunately, the rich symmetry groups associated with the marginal polytope and Sherali–Adams relaxations, as discussed in Section 4.3, allow large computational savings to be made.

From the remarks in Section 4.3, $\mathbb{L}_4(G)$ is not tight for some potential $\phi \in \mathbb{R}^{V \cup E}$ iff it is not tight for $F(\phi)$ for any $F \in \text{Aut}(G) \rtimes \mathbb{Z}_2^{|V|}$, the symmetry group of the Sherali–Adams polytope. Note that as well as acting naturally on the space of potentials, the symmetry group $\text{Aut}(G) \rtimes \mathbb{Z}_2^{|V|}$ of the polytope $\mathbb{L}_4(G)$ also acts naturally on the set of signings of G.

From this, we deduce that a given signing Σ of G is non-tight iff $F(\Sigma)$ is non-tight for any (all) $F \in \mathrm{Aut}(G) \rtimes \mathbb{Z}_2^{|V|}$. Therefore it suffices to check the optimal value of Problem (4.14) only for one representative signing of G in each orbit under the action of $\mathrm{Aut}(G) \rtimes \mathbb{Z}_2^{|V|}$.

Finally, since we have an explicit H-representation of the polytope $\mathcal{M}(G,\Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z)$, we can find a V-representation (i.e. an exhaustive collection of extremal points), and hence recover a list of all LPs of the form of Problem (4.14) that we need to solve in order to find the optimum of Problem (4.12). The results of these calculations for each of the four graphs in T_4, the minimal forbidden minors for treewidth ≤ 3, are detailed in Sections 4.A.4 to 4.A.7. The calculations show that the only signing of a graph in T_4 which is not tight for $\mathbb{L}_4$ is the odd-K_5 in Figure 4.8, from which Theorem 4.1 follows.

4.5 Discussion

In previous work, Weller (2016a) showed that tightness of $\mathbb{L}_2(G)$ for all valid potentials on a signed graph (G,Σ) may be characterised exactly by forbidding just an odd-K_3 as signed minor, and that a similar result for $\mathbb{L}_3(G)$ holds by forbidding just an odd-K_4. These are precisely the odd versions of the forbidden minors for the respective treewidth conditions given by Wainwright and Jordan (2004). In this work, we have shown that this relationship between forbidden minors for bounded treewidth and exactness of Sherali–Adams relaxations breaks down beyond this point, through the use of a geometric characterisation of the models for which Sherali–Adams relaxations are tight.

A natural subject for further work is the identification of other minimal forbidden minors for tightness of $\mathbb{L}_4(G)$; we note that one natural candidate for further investigation is some signing of a $k \times k$ grid of sufficient size, since this is planar with treewidth k. Unification with results such as those for almost-balanced models appearing in Chapter 3 is also of interest. Our treatment here is for Sherali–Adams relaxations that apply higher-order consistency constraints uniformly across the graph; it will be interesting to see to what extent a graph minor-based theory can be built up around localised relaxations. More broadly, the use of signed minors for analysing approximate inference algorithms on graphical models is a young area, and there is much scope for further work in understanding a variety of approximate inference algorithms in this way.

Appendix 4.A Proofs

4.A.1 Geometric proofs of results for the pairwise Sherali–Adams relaxation

In this section, we give proofs of Lemmas 4.5 and 4.6 using the geometric insights of Section 4.3.

Lemma 4.5. If G is a tree, then $\mathbb{L}_2(G)$ is tight for all potentials, i.e. $\mathbb{L}_2(G) = \mathbb{M}(G)$.

Proof. We will show that given arbitrary potentials $\phi \in \mathbb{R}^{V \cup E}$ with a MAP optimum configuration $z \in \text{Ext}(\mathbb{M}(G))$, we have $\phi \in N_{\mathbb{L}_2(G)}(z)$, from which it immediately follows from the definition of normal cones that z is optimal for ϕ in $\mathbb{L}_2(G)$, and hence the result follows. It will be helpful to denote the standard basis of $\mathbb{R}^{V \cup E}$ by $\{\mathbf{e}_i | i \in V\} \cup \{\mathbf{e}_{ij} | ij \in E\}$, so that a potential $\phi \in \mathbb{R}^{V \cup E}$ may be expressed as $\phi = \sum_{i \in V} \phi_i \mathbf{e}_i + \sum_{ij \in E} \phi_{ij} \mathbf{e}_{ij}$.

We first argue that it is sufficient to consider potentials $\phi \in \mathbb{R}^{V \cup E}$ such that $\phi \in N_{\mathbb{M}(G)}(z^{(0)})$, where $z^{(0)} \in \{0,1\}^{V \cup E}$ is the configuration where every random variable X_v in the graphical model is set to 0 — this will neaten the argument. To see this claim, suppose that $\phi \in \mathbb{R}^{V \cup E} \setminus N_{\mathbb{M}(G)}(z^{(0)})$. Since $\mathbb{R}^{V \cup E} = \cup_{z \in \text{Ext}(\mathbb{M}(G))} N_{\mathbb{M}(G)}(z)$, there exists $z' \in \text{Ext}(\mathbb{M}(G))$ such that $\phi \in N_{\mathbb{M}(G)}(z')$. Now consider the flipping map F (see Section 4.3.1) that maps $z' \in \mathbb{M}(G)$ to $z^{(0)} \in \mathbb{M}(G)$, and recall by Proposition 1.19 that $\mathbb{L}_2(G)$ is tight for ϕ iff it is tight for $F^\dagger(\phi)$, and that since $z^{(0)}$ is optimal for $F^\dagger(\phi)$, we have $F^\dagger(\phi) \in N_{\mathbb{M}(G)}(z^{(0)})$.

Thus, we assume $\phi \in N_{\mathbb{M}(G)}(z^{(0)})$. We now proceed by induction on the number of vertices n of G. Our method will be to show that $\phi \in N_{\mathbb{L}_2(G)}(z^{(0)})$ by directly demonstrating that ϕ can be written as a non-negative linear combination of extremal vectors of $N_{\mathbb{L}_2(G)}(z^{(0)})$. The cases $n = 1, 2$ are trivial, as $\mathbb{L}_2(G) = \mathbb{M}(G)$ for such graphs immediately from the definition of Sherali–Adams polytopes. For the inductive step, let $u \in V$ be a leaf, with neighbour $u' \in V$. As the variables associated with u, u' are both equal to 0 at the marginal optimum, there are three tight local constraints (see Section 3.3 for a recap of these constraints) corresponding to the pair u, u' which are tight at the marginal polytope optimum $z^{(0)}$, namely $q_{uu'} \leq q_u$, $q_{uu'} \leq q_{u'}$, and $q_{uu'} \geq 0$. By Lemma 4.3, the outward-pointing normal to each hyperplane defined by these constraints is an extremal vector of the cone $N_{\mathbb{L}_2(G)}(z^{(0)})$; using the notation for basis vectors introduced above, we can write the outward pointing normals for these constrains as $\mathbf{e}_{uu'} - \mathbf{e}_u$, $\mathbf{e}_{uu'} - \mathbf{e}_{u'}$, and $-\mathbf{e}_{uu'}$. Note that as $z^{(0)}$ is an optimal configuration, setting either (or both) of the variables corresponding to u and u'

to 1 cannot result in a higher scoring configuration, from which we obtain the inequalities $\phi_u, \phi_{u'} \leq 0$, and $\phi_u + \phi_{u'} + \phi_{uu'} \leq 0$.

We now consider two cases: (i) $\phi_{uu'} \geq -\phi_u$; and (ii) $\phi_{uu'} < -\phi_u$. First, suppose (i) holds, so that $\phi_{uu'} \geq -\phi_u$. We consider the following conical combination of outward-pointing normals:

$$-\phi_u(\mathbf{e}_{uu'} - \mathbf{e}_u) + (\phi_{uu'} + \phi_u)(\mathbf{e}_{uu'} - \mathbf{e}_{u'}). \tag{4.15}$$

Note that the coefficients of $\mathbf{e}_u$ and $\mathbf{e}_{uu'}$ exactly match the singleton potential on u and the edge potential on uu' for the model. We now consider the residual potentials (that is, the model whose potentials are the difference between the potentials of the original model, and the potentials given by the conical combination of extremal vectors in Expression (4.15)). It is sufficient for the residual potentials to have the same optimal marginal vertex for the inductive step to be complete, since the residual model has one fewer variable; the coefficient of the singleton potential on u and pairwise potential on uu' in Expression (4.15) matches that of the original model, so the random variable corresponding to u may be removed. Note that the residual potentials are the same for all singletons except u and u'; the new potential on u is 0 and the new potential on u' is $\phi_u + \phi_{u'} + \phi_{uu'}$. If some other configuration yields a higher score than $z^{(0)}$ on this residual model, it must be the case that $X_{u'}$ is set to 1 in this configuration — else the same configuration would yield a greater score than that of $z^{(0)}$ in the original model. But if $X_{u'}$ is set to 1, the residual model yields the same score as setting both z_u and $z_{u'}$ equal to 1 in the original model. Therefore we deduce that $z^{(0)}$ is still optimal for the residual model, and the inductive step is complete.

We now address case (ii), where $\phi_{uu'} < -\phi_u$. Consider the following conical combination

$$-\phi_u(\mathbf{e}_{uu'} - \mathbf{e}_u) + (-\phi_{uu'} - \phi_u)(-\mathbf{e}_{uu'}). \tag{4.16}$$

Note again that the coefficients of $\mathbf{e}_u$ and $\mathbf{e}_{uu'}$ exactly match the singleton potential on u and the edge potential on uu' for the model. The residual model therefore has 0 potential for the singleton u and the edge uu', and the original singleton potential for u'. It is immediate that an optimal marginal vertex for the residual potentials is again $z^{(0)}$; if any other vertex yielded a higher score, then by ensuring that the u variable is set to 0 in this new configuration, this vertex would yield a higher score than $z^{(0)}$ in the original model, a contradiction. Therefore we deduce that $z^{(0)}$ is optimal for the residual model, and the inductive step is complete. □

Lemma 4.6. For an arbitrary graph G, $\mathbb{L}_2(G)$ is tight for the set of balanced models on G.

Proof. Let $\phi = ((\phi_i)_{i\in V}, (\phi_{ij})_{ij\in E}) \in \mathbb{R}^{V\cup E}$ be a balanced set of potentials on G. We may assume that no edge potentials are repulsive. To see this, partition V into two disjoint sets V_a and V_b such that all edges between V_a and V_b are repulsive, whilst all within either V_a or V_b are attractive (this is possible as ϕ is assumed balanced). Then flip the set V_a (i.e. apply the map $F^{\dagger}_{(V_a)}$ described in Section 4.A.2 to ϕ). By construction, $F^{\dagger}_{(V_a)}(\phi)$ has all edge potentials attractive, and by Proposition 1.19, $\mathbb{L}_2(G)$ is tight for ϕ iff it is tight for $F^{\dagger}_{(V_a)}(\phi)$.

Let $z^\star \in \mathbb{M}(G)$ be the optimal marginal vertex for ϕ, and partition V into two disjoint sets $V_0 = \{i \in V | z^\star_i = 0\}$ and $V_1 = \{i \in V | z^\star_i = 1\}$. Given $i \in V_0$ and $j \in V_1$, note that the three tight local constraints corresponding to the edge ij are $q_{ij} \leq q_i$, $q_{ij} \geq 0$, and $q_{ij} \geq q_i + q_j - 1$. In the notation introduced in the proof of Lemma 4.5, the three extremal vectors contributed by this edge to the cone $N_{\mathbb{L}_2(G)}(z^\star)$ are $\mathbf{e}_{ij} - \mathbf{e}_i$, $-\mathbf{e}_{ij}$ and $\mathbf{e}_i + \mathbf{e}_j - \mathbf{e}_{ij}$.

We construct the vector

$$\sum_{i\in V_0, j\in V_1} \phi_{ij}(\mathbf{e}_{ij} - \mathbf{e}_i)\,, \tag{4.17}$$

which lies in $N_{\mathbb{L}_2(G)}(z^\star)$, as each vector in the sum is extremal for $N_{\mathbb{L}_2(G)}(z^\star)$, and all coefficients are non-negative by assumption of attractiveness of ϕ. We then consider the model given by the residual potentials ϕ', which is the vector of potentials given by taking the difference between the original potentials and the potentials given by the vector in Expression (4.17) above:

$$\phi'_i = \phi_i + \sum_{j\in V_1} \phi_{ij} \text{ for } i \in V_0\,, \quad \phi'_{ij} = 0 \text{ for } i \in V_0, j \in V_1\,, \tag{4.18}$$

with all other components of ϕ' equal to the corresponding components of ϕ . It is sufficient to show that (i) the residual model has the same optimal vertex $z^\star$ as the model given by ϕ, and (ii) that the residual potentials can be written as the conical combination of extremal vectors in $N_{\mathbb{L}_2(G)}(z^\star)$.

We begin with (i). Let $z' \in \mathbb{M}(G)$ be a new optimal marginal vertex for the residual model with potentials ϕ'. We further refine our partition $V = V_0 \cup V_1$ by defining

$$V_{0\to 0} = \{i \in V | z^\star_i = 0, z'_i = 0\}\,, \quad V_{0\to 1} = \{i \in V | z^\star_i = 0, z'_i = 1\}\,, \tag{4.19}$$

$$V_{1\to 0} = \{i \in V | z^\star_i = 1, z'_i = 0\}\,, \quad V_{1\to 1} = \{i \in V | z^\star_i = 1, z'_i = 1\}\,. \tag{4.20}$$

If $V_{0\to 1} = V_{1\to 0} = \emptyset$, we are done as $z^\star = z'$. We note that in fact, it is sufficient to show $V_{0\to 1} = \emptyset$, since then the scores of both z' and $z^\star$ for ϕ' are the same as their respective scores for ϕ; this follows since the only components of ϕ and ϕ' that differ are singleton potentials for variables in V_0, and pairwise potentials between variables in V_0 and variables in V_1. Thus, from $V_{0\to 1} = \emptyset$ it immediately follows that $z^\star$ is optimal for ϕ' too.

Thus, take the configuration z', and consider the effect on the score with respect to the residual potentials ϕ' by flipping all variables in the set $V_{0\to 1}$ from taking value 1 to taking value 0. The effect is to decrease the score by the following quantity:

$$\sum_{i\in V_{0\to 1}} \phi'_i + \sum_{i,j\in V_{0\to 1}} \phi'_{ij} + \sum_{\substack{i\in V_{0\to 1}\\ j\in V_{1\to 1}}} \phi'_{ij}\,. \tag{4.21}$$

By the relationship in Expression (4.18), we may rewrite this in terms of the potentials ϕ as follows:

$$\sum_{i\in V_{0\to 1}} \phi_i + \sum_{\substack{i\in V_{0\to 1}\\ j\in V_1}} \phi_{ij} + \sum_{i,j\in V_{0\to 1}} \phi_{ij}\,. \tag{4.22}$$

But now note that this is precisely the amount by which the score of $z^\star$ for ϕ would increase if the variables in $V_{0\to 1}$ (which are set to 0 in $z^\star$) were set to 1 in this configuration. Thus, by optimality of $z^\star$ for ϕ, Expression (4.22) must be non-negative. It then follows that Expression (4.21) is non-negative, and so the score of the configuration given by taking z' and flipping the variables in $V_{0\to 1}$ with respect to ϕ' is at least as great as that of z' itself. Thus, we may take $V_{0\to 1} = \emptyset$, and hence (i) is proven.

It now remains to deal with (ii): demonstrating that the residual potentials ϕ' can be written as a conical combination of vectors in $N_{\mathbb{L}_2(G)}(z^\star)$. Connected components can be treated separately, so by passing to the residual model as in (i) above, we need only consider cases where all edge potentials are attractive, and we may assume that the optimal marginal vertex has all variables set to the same value, as is the case for the connected components of the residual model in (i) above.

Again, by flipping all variables in the model if necessary, we may assume that the optimal value of all variables is 0. We then demonstrate how to subtract a sequence of vectors in $N_{\mathbb{L}_2(G)}(z^\star)$ from the potential, so as to sequentially remove edges, until the residual is a model on a spanning tree. At that point, we can apply Lemma 4.5 to show the residual is in $N_{\mathbb{L}_2(G)}(z^\star)$.

We therefore take an edge $ij \in E$, noting that the tight LOC constraints corresponding to this edge (since the optimal configuration has $z_i = z_j = 0$) are $q_{ij} \leq q_i, q_{ij} \leq q_j$ and $q_{ij} \geq 0$, corresponding to outward-pointing normals in $N_{\mathbb{L}_2(G)}(z)$ of $\mathbf{e}_{ij} - \mathbf{e}_i$, $\mathbf{e}_{ij} - \mathbf{e}_j$ and $-\mathbf{e}_{ij}$. We then consider the vector $\alpha(\mathbf{e}_{ij} - \mathbf{e}_i) + \beta(\mathbf{e}_{ij} - \mathbf{e}_j) \in N_{\mathbb{L}_2(G)}(z^\star)$, with the aim of identifying $\alpha, \beta \geq 0$, so that the residual model has 0 edge potential for the edge $ij \in E$, and so that the optimal marginal vertex for the model is unchanged. We now demonstrate that such α, β exist.

To ensure that the residual model has 0 edge marginal for the edge ij, we must take $\beta = \phi_{ij} - \alpha$. Therefore, it is sufficient to find $0 \leq \alpha \leq \phi_{ij}$ so that the residual model also has optimal marginal vertex given by taking all random variables to have value 0. Denoting the potentials for the residual model by $\phi' = ((\phi'_i)_{i \in V}, (\phi'_{ij})_{ij \in E})$, we note that

$$\phi'_i = \alpha + \phi_i\,, \quad \phi'_j = (\phi_{ij} - \alpha) + \phi_j\,, \quad \phi'_{ij} = 0\,, \tag{4.23}$$

with all other components of ϕ' equal to the respective component of ϕ. Now note that any configuration of the random variables in which $z_i = z_j$ has the same score with respect to ϕ' and with respect to ϕ, meaning that such a configuration cannot score more highly with respect to ϕ' than the configuration with all variables set to 0. Now consider a configuration optimal for ϕ' subject to $z_i = 1$ and $z_j = 0$ (denote by T_i the set of indices corresponding to variables set to 1 in this configuration, so that the optimal configuration has $z_{T_i} = 1$ and $z_{V \setminus T_i} = 0$), and similarly consider a configuration optimal for ϕ' subject to $z_i = 0$ and $z_j = 1$ (denote by T_j the set of indices corresponding to variables set to 1 in this configuration). It now suffices to show that both of these configurations have lower score with respect to ϕ' than the configuration with all variables set to 0 (which has score 0). For general subsets $A, B \subseteq V$, we write

$$\Theta_A = \sum_{i \in A} \phi_i + \sum_{\substack{ij \in E \\ i,j \in A}} \phi_{ij}\,, \quad \text{and} \quad \Theta_{A,B} = \sum_{\substack{ij \in E \\ i \in A, j \in B}} \phi_{ij}\,, \tag{4.24}$$

so that the score of the configuration $(z_{T_i} = 1, z_{V \setminus T_i} = 0)$ with respect to ϕ' is $\alpha + \Theta_{T_i}$, and the score of the configuration $(z_{T_j} = 1, z_{V \setminus T_j} = 0)$ with respect to ϕ' is $(\phi_{ij} - \alpha) + \Theta_{T_j}$. We note that it therefore suffices to find $0 \leq \alpha \leq \phi_{ij}$ such that

$$\alpha + \Theta_{T_i} \leq 0\,, \quad (\phi_{ij} - \alpha) + \Theta_{T_j} \leq 0\,. \tag{4.25}$$

It is therefore sufficient to show that $\phi_{ij} + \Theta_{T_i} + \Theta_{T_j} \leq 0$. Note that (by optimality of the score 0 for ϕ) we have

$$0 \geq \Theta_{T_i \cup T_j} = \Theta_{T_i} + \Theta_{T_j} - \Theta_{T_i \cap T_j} + \Theta_{\{j\}, T_i \setminus T_j \setminus \{i\}} + \Theta_{\{i\}, T_j \setminus T_i \setminus \{j\}} + \Theta_{T_i \setminus T_j \setminus \{i\}, T_j \setminus T_i \setminus \{j\}} + \phi_{ij}\,, \tag{4.26}$$

and so

$$\Theta_{T_i} + \Theta_{T_j} + \phi_{ij} \leq \Theta_{T_i \cap T_j} - \Theta_{\{j\}, T_i \setminus T_j \setminus \{i\}} - \Theta_{\{i\}, T_j \setminus T_i \setminus \{j\}} - \Theta_{T_i \setminus T_j \setminus \{i\}, T_j \setminus T_i \setminus \{j\}} \leq 0\,, \tag{4.27}$$

where the second inequality above follows as each individual term is non-positive. The result then follows as described above. □

4.A.2 Group-theoretic result from Section 4.3.2

Lemma 4.7. $\mathbb{L}_r(G)$ is tight for a given potential $\phi \in \mathbb{R}^{V \cup E}$ iff it is tight for all potentials $Y_\sigma^\dagger(\phi)$, $\sigma \in \mathrm{Aut}(G)$.

Proof. It is sufficient to show that if $\mathbb{L}_k(G)$ is tight for ϕ, then given $\sigma \in \mathrm{Aut}(G)$, $\mathbb{L}_k(G)$ is tight for $Y_\sigma^\dagger(\phi)$. Recalling that for $\sigma \in \mathrm{Aut}(G)$, Y_σ maps any Sherali–Adams polytope $\mathbb{L}_r(G)$ to itself bijectively, we obtain

$$\max_{q \in \mathbb{L}_k(G)} \langle \phi, q \rangle = \max_{q \in \mathbb{L}_k(G)} \langle \phi, Y_{\sigma^{-1}} q \rangle = \max_{q \in \mathbb{L}_k(G)} \left\langle Y_\sigma^\dagger(\phi), q \right\rangle, \tag{4.28}$$

and

$$\max_{q \in \mathbb{M}(G)} \langle \phi, q \rangle = \max_{q \in \mathbb{M}(G)} \langle \phi, Y_{\sigma^{-1}} q \rangle = \max_{q \in \mathbb{M}(G)} \left\langle Y_\sigma^\dagger(\phi), q \right\rangle. \tag{4.29}$$

It follows that $\mathbb{L}_k(G)$ is not tight for ϕ (i.e. $\max_{q \in \mathbb{L}_k(G)} \langle \phi, q \rangle > \max_{q \in \mathbb{M}(G)} \langle \phi, q \rangle$) if and only if it is not tight for $Y_\sigma^\dagger(\phi)$ (i.e. $\max_{q \in \mathbb{L}_k(G)} \langle Y_\sigma^\dagger(\phi), q \rangle > \max_{q \in \mathbb{M}(G)} \langle Y_\sigma^\dagger(\phi), q \rangle$). □

4.A.3 Proof of results from Section 4.4

We provide a proof of the following result from Section 4.2.4.

Theorem 4.1. The only non-tight signing for $\mathbb{L}_4(G)$ of any graph G in the set T_4 (the minimal forbidden minors for treewidth ≤ 3) is the odd-K_5.

Recall from Section 4.4 that for a given signed graph (G,Σ), $\mathbb{L}_4(G)$ is tight for all models on G respecting the signing Σ (the set of which is denoted by $\mathcal{M}(G,\Sigma)$) iff the following problem has optimal value 0:

$$\max_{\phi\in\mathcal{M}(G,\Sigma)}\max_{q\in\mathbb{L}_4(G)}\min_{z\in\mathbb{M}(G)}\left[\langle\phi,q\rangle-\langle\phi,z\rangle\right] \tag{4.30}$$

We restate and prove the chain of results in Section 4.4 that allow for this problem to be reduced to a manageable collection of linear programs.

Lemma 4.8. Problem (4.11) has a non-zero optimum iff

$$\max_{\phi\in\mathcal{M}(G,\Sigma)\cap\Pi}\max_{q\in\mathbb{L}_4(G)}\min_{z\in\mathbb{M}(G)}\left[\langle\phi,q\rangle-\langle\phi,z\rangle\right] \tag{4.12}$$

has a non-zero optimum, where Π is any polytope containing the origin in its interior.

Proof. Since $\mathcal{M}(G,\Sigma)\cap\Pi\subseteq\mathcal{M}(G,\Sigma)$, if Problem (4.12) has a non-zero optimum then so does Problem (4.30). For the opposite direction, suppose Problem (4.30) has a non-zero optimum, so in particular there exists $\phi'\in\mathcal{M}(G,\Sigma)$ achieving the non-zero optimal value in Expression (4.12), with corresponding configurations $q'\in\mathbb{L}_4(G)$ and $z'\in\mathbb{M}(G)$, such that

$$\langle\phi',q'\rangle-\langle\phi',z'\rangle>0\,. \tag{4.31}$$

As $\mathbf{0}\in\mathrm{int}(\Pi)$, Π contains an open ball around the origin, and so there exists $\lambda>0$ such that $\lambda\phi'\in\Pi$. The optimal elements of $\mathbb{L}_4(G)$ and $\mathbb{M}(G)$ are unchanged by this scaling of ϕ', and so the value of the objective attained by the potentials $\lambda\phi'$ is

$$\langle\lambda\phi',q'\rangle-\langle\lambda\phi',z'\rangle=\lambda\left(\langle\phi',q'\rangle-\langle\phi',z'\rangle\right)>0\,, \tag{4.32}$$

and so it follows that (4.12) has a non-zero optimum. □

Lemma 4.9. Problem (4.12) has a non-zero optimum iff at least one of the following set of problems (indexed by $z\in\mathrm{Ext}(\mathbb{M}(G))$) has a non-zero optimum:

$$P'_z:\quad\max_{\phi\in\mathcal{M}(G,\Sigma)\cap\Pi\cap N_{\mathbb{M}(G)}(z)}\max_{q\in\mathbb{L}_4(G)}\left[\langle\phi,q\rangle-\langle\phi,z\rangle\right]. \tag{4.13}$$

Proof. We have $\mathbb{R}^{V\cup E}=\cup_{z\in\mathrm{Ext}(\mathbb{M}(G))}N_{\mathbb{M}(G)}(z)$, as $\mathbb{M}(G)$ is a polytope. Therefore if (ϕ',q',z') is optimal for Problem (4.12), then it is optimal for $P_{z'}$, since $\phi'\in N_{\mathbb{M}(G)}(z')$, as z' is optimal

for ϕ' in $\mathbb{M}(G)$. Conversely, the optimal value for Problem (4.12) is the maximum of the optimal values for each problem P_z for $z \in \text{Ext}(\mathbb{M}(G))$, since for any $\phi \in N_{\mathbb{M}(G)}(z)$, we have $\langle \phi, z \rangle = \max_{y \in \mathbb{M}(G)} \langle \phi, y \rangle$. □

Lemma 4.10. An optimum $(\phi^\star, q^\star)$ of P'_z (see Expression (4.13)) occurs with $\phi^\star$ (resp. $q^\star$) extremal in $\mathcal{M}(G, \Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z)$ (resp. $\mathbb{L}_4(G)$).

Proof. Let $(\phi^\star, q^\star) \in (\mathcal{M}(G, \Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z)) \times \mathbb{L}_4(G)$ be optimal for P'_z. Consider the objective of P'_z with $q^\star$ fixed as a function of the potential parameter ϕ. The resulting objective is linear in ϕ, and so we can move $\phi^\star$ to some extremal point of $\mathcal{M}(G, \Sigma) \cap \Pi \cap N_{\mathbb{M}(G)}(z)$ without reducing the score attained with $q^\star$ fixed. Therefore we may take $\phi^\star$ to be extremal. By an analogous argument holding $\phi^\star$ fixed, the objective is affine in $q^\star$, so we may take $q^\star$ to be extremal. □

Having established these intermediate results, we now give the details of the linear programs of the form given in Problem (4.14) required to ascertain Theorem 4.1. Code for evaluating these linear programs is available upon request.

4.A.4 Forbidden signed versions of K_5

We initially calculate the orbits of the set of signings of K_5 under the action of the joint permutation and flipping group $\mathcal{S}_V \rtimes \mathbb{Z}_2^{|V|}$, as discussed in Section 4.4.2. There are 7 different orbits of signings of K_5 — one representative of each is shown in Figure 4.9. Here, as elsewhere in this thesis, solid blue edges are even (representing attractive edge potentials) and dashed red edges are odd (representing repulsive edge potentials).

The tightness of some classes can be determined by known results:

- Figure 4.9g corresponds to balanced signings; it is known that $\mathbb{L}_2(K_5)$ is tight for such potentials (by Lemma 4.6, for example), and so it follows that $\mathbb{L}_4(K_5)$ is too, as $\mathbb{L}_4(G) \subseteq \mathbb{L}_2(G)$.
- Figures 4.9d and 4.9e are almost balanced, so are tight for $\mathbb{L}_3(K_5)$ (by Theorem 3.2 of Chapter 3) and so are tight for $\mathbb{L}_4(K_5)$ too, as $\mathbb{L}_4(G) \subseteq \mathbb{L}_3(G)$.

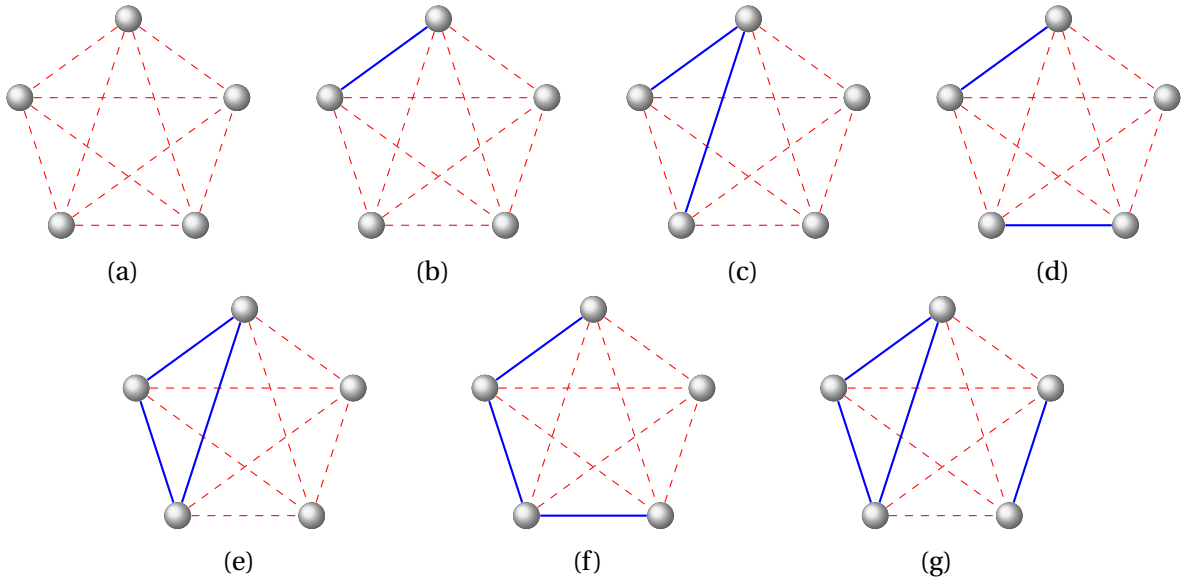


Figure 4.9. One representative of each of the 7 orbits of signings of K_5 under the action of the joint flipping and permutation group for this graph.

Checking the relevant LPs of the form given by Problem (4.14) for the remaining signings reveals that the signings shown in Figures 4.9b, 4.9c and 4.9f are tight with respect to $\mathbb{L}_4(K_5)$, whilst Figure 4.9a is not. We therefore deduce that the only forbidden signing of K_5 for $\mathbb{L}_4(K_5)$ is that in Figure 4.9a: the odd-K_5.

4.A.5 Forbidden signed versions of the octahedral graph

We initially calculate the orbits of the set of signings of the octahedral graph O_6 under the action of the joint flipping and permutation group for this graph. Note that the permutation group for this graph is not $\mathcal{S}_V$, since not all permutations in $\mathcal{S}_V$ respect the topology of O_6; the automorphism group of O_6 is in fact of order 48, and isomorphic to the group of rigid symmetries of the cube (this can be seen by identifying vertices of the graph with faces of the cube, and edges in the graph encoding adjacency of faces). There are 14 different orbits of signings of O_6 under this group action — one representative of each is shown in Figure 4.10.

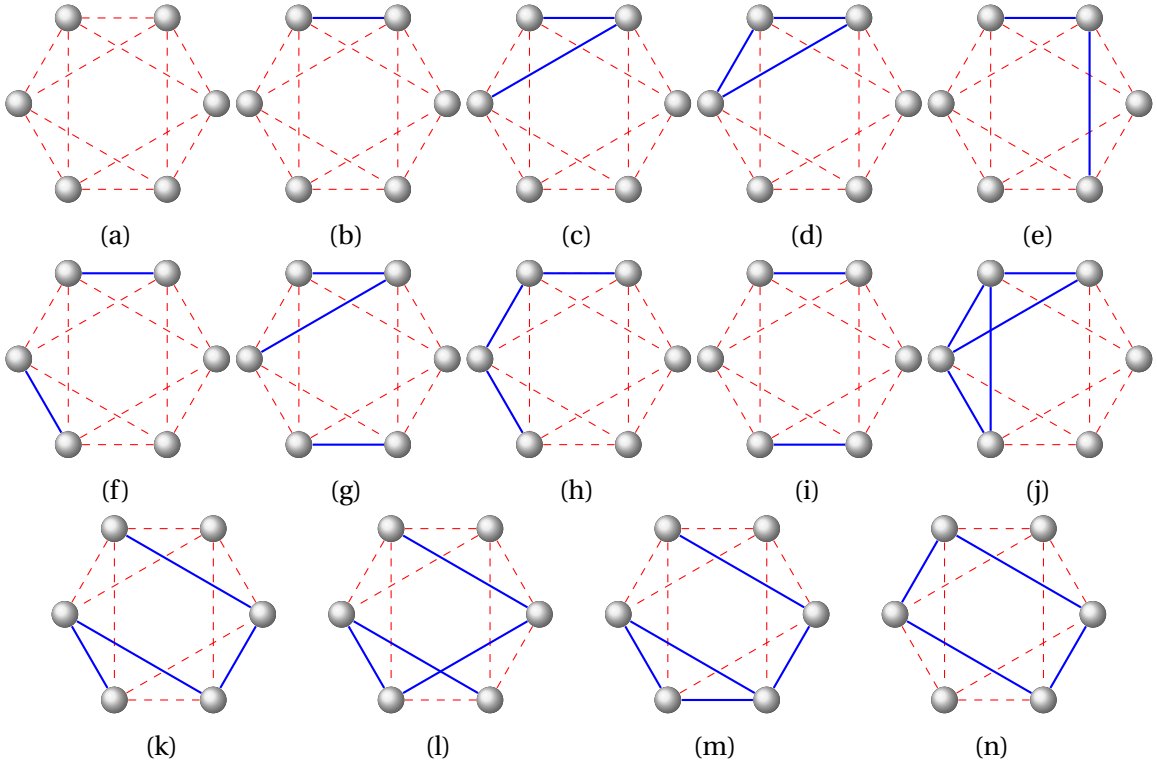


Figure 4.10. One representative of each of the 14 orbits of signings of O_6 under the action of the joint flipping and permutation group for this graph.

Checking the relevant LPs of the form given by Problem (4.14) for all 14 signings reveals that all signings are tight with respect to $\mathbb{L}_4(O_6)$, and it immediately follows that $\mathbb{L}_4(O_6)$ is tight for any potentials on O_6, and so O_6 need not be forbidden as a minor of a graph G to ensure tightness of $\mathbb{L}_4(G)$.

4.A.6 Forbidden signed versions of the Wagner graph

We initially calculate the orbits of the set of signings of the Wagner graph M_8 under the action of the joint flipping and permutation group for this graph, as discussed in Section 4.4.2. There are 8 different orbits of signings of M_8 — one representative of each is shown in Figure 4.11.

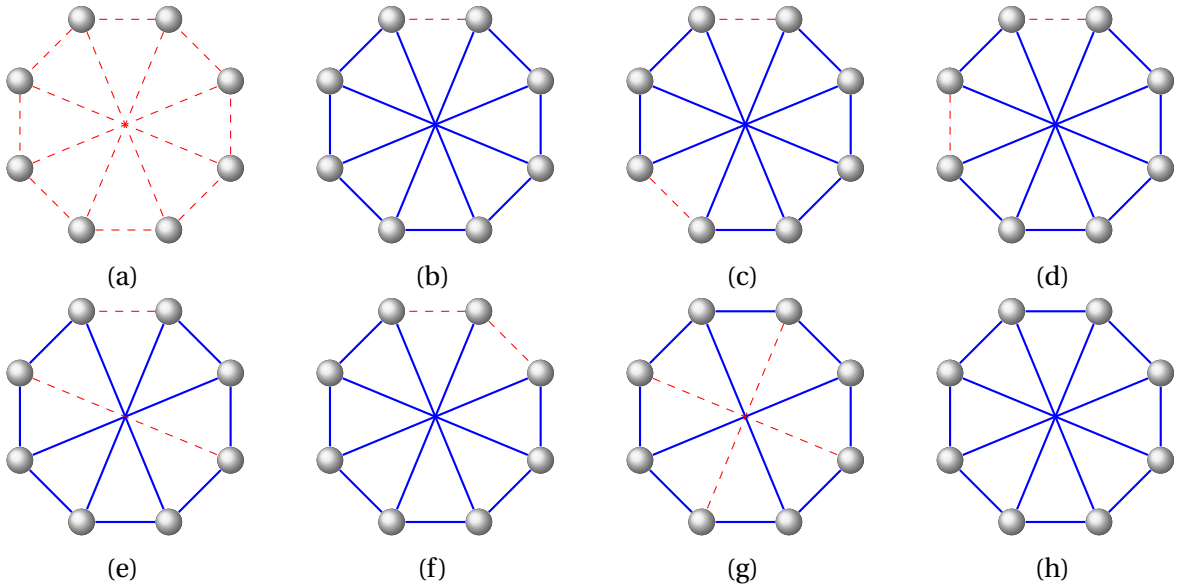


Figure 4.11. One representative of each of the 8 orbits of signings of M_8 under the action of the joint flipping and permutation group of this graph.

Checking the relevant LPs of the form given by Problem (4.14) for all 8 signings reveals that all signings are tight with respect to $\mathbb{L}_4(M_8)$, and it immediately follows that $\mathbb{L}_4(M_8)$ is tight for any potentials on M_8, and so M_8 need not be forbidden as a minor of a graph G to ensure tightness of $\mathbb{L}_4(G)$.

4.A.7 Forbidden signed versions of the pentagonal prism graph

We initially calculate the orbits of the set of signings of the pentagonal prism graph Y_5 under the action of the joint flipping and permutation group of the graph, as discussed in Section 4.4.2. There are 12 different orbits of signings of Y_5 — one representative of each is shown in Figure 4.12.

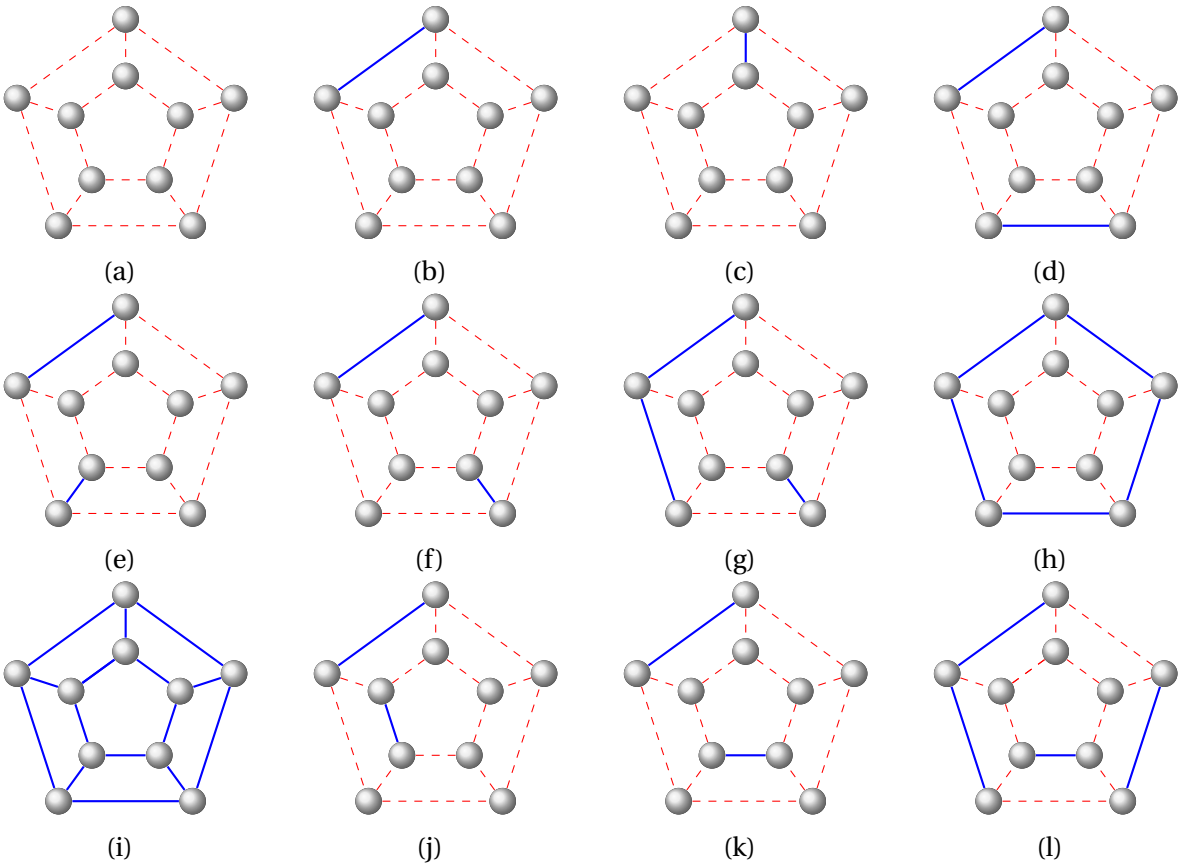


Figure 4.12. One representative of each of the 12 orbits of signings of Y_5 under the action of the joint flipping and permutation group of this graph.

Checking the relevant LPs of the form given by Problem (4.14) for all 14 signings reveals that all signings are tight with respect to $\mathbb{L}_4(Y_5)$, and it immediately follows that $\mathbb{L}_4(Y_5)$ is tight for any potentials on Y_5, and so Y_5 need not be forbidden as a minor of a graph G to ensure tightness of $\mathbb{L}_4(G)$.

Part II

Structure in Monte Carlo Methods

Chapter 5

Monte Carlo Methods for Kernel Approximation and Dimensionality Reduction

The focus of the second half of this thesis is Monte Carlo methods for both linear dimensionality reduction and approximation of kernel methods. We are concerned with sampling methods that go beyond i.i.d. sampling, and instead introduce shared structure between samples, resulting in an improvement in statistical efficiency, computational efficiency, space (memory) efficiency, or a combination of these factors, relative to standard approaches. This first chapter briefly establishes the necessary background material and context for our original work, which appears in Chapters 6 and 7, focusing on random projections, and random feature approximations to kernel methods.

5.1 Kernel methods

Kernel methods are widely used in statistics and machine learning, and include important model classes such as Gaussian processes (Rasmussen and Williams, 2005) and support vector machines (Cortes and Vapnik, 1995; Schölkopf and Smola, 2001), as well as other applications such as reinforcement learning (Ormoneit and Sen, 2002). We begin by reviewing kernel methods as non-parametric extensions of linear methods, then discuss issues with their scalability, and finally describe random feature approximations as solutions to these scalability issues.

5.1.1 Linear methods

Linear methods form the backbone of many disciplines in machine learning and statistics, including regression, classification and unsupervised learning, using methods such as (Bayesian) linear regression, linear support vector machines (SVMs), and principal component analysis, respectively. Such methods are generally well understood theoretically, but can perform poorly on datasets which do not exhibit clear linear structure. A classical example of this phenomenon is the inability of a linear SVM to correctly classify data classes that are not linearly separable, such as the "XOR" example in Figure 5.1, in which an element of $\mathbb{R}^2$ is classified as "red" if exactly one of its coordinates is positive, and is classified as "blue" otherwise.

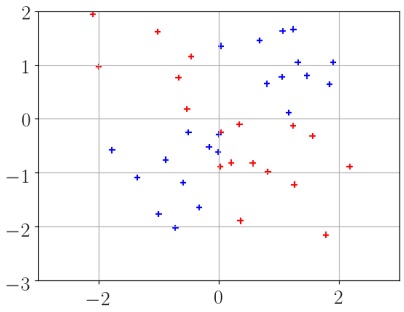


Figure 5.1. The red and blue classes of data are not linearly separable.

One way around this is to engineer non-linear features of the data which reveal some linear structure, and to use these non-linear features as inputs to a linear algorithm. One resolution to the XOR example above in this vein is to augment the representation of each data point (x_1, x_2) by adding a third coordinate given by the product of the first two, so the inputs to the SVM have the form (x_1, x_2, x_1x_2). The data are now linearly separable by checking the sign of the third coordinate — see Figure 5.2.

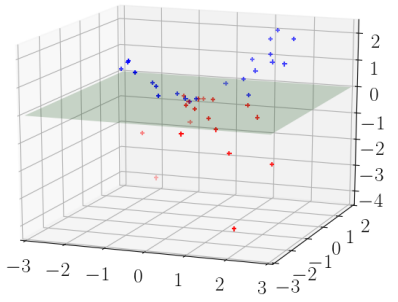


Figure 5.2. With the addition of a non-linear feature, the red and blue classes are now linearly separable.

This combination of hand-crafted non-linear features and linear learning algorithms greatly expands the scope of linear methods in statistics and machine learning, and can allow experts to incorporate domain-specific knowledge into their models. However, there are several significant drawbacks to this approach. Firstly, it may be extremely difficult to engineer useful features by hand, and even when possible, feature engineering is often a very time-consuming process. Secondly, the larger the designed feature representation for the data in question, the longer the running time for any subsequently used linear learning algorithm. Much of modern machine learning can be thought of as addressing these problems, by either learning useful features directly from data, or by implicitly working with high- (or infinite-) dimensional representations of data in a computationally

tractable way. Kernel methods take the latter approach, making use of the so-called kernel trick.

5.1.2 The kernel trick

Kernel methods make use of the insight that learning algorithms for these linear methods often do not require access to the raw input features themselves, but just inner products between the feature vectors. Thus, if one has data $(\mathbf{x}_i)_{i=1}^N$ lying in some base space $\mathbb{X}$, and wishes to run a linear learning algorithm on non-linear features $(\phi(\mathbf{x}_i))_{i=1}^N$ obtained through a mapping $\phi : \mathbb{X} \to \mathcal{H}$ for some Hilbert space $\mathcal{H}$, it is sufficient to have access to an oracle function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ that returns $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{H}}$. With this oracle in hand, we can replace all instances of inner products between data points $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ in a linear method with evaluations of the oracle $K(\mathbf{x}_i, \mathbf{x}_j)$, and implicitly run the linear method on the feature vectors $(\phi(\mathbf{x}_i))_{i=1}^N$, without ever computing the features themselves.

The Moore–Aronszajn theorem (Aronszajn, 1950) classifies precisely the set of functions $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ that arise as oracle functions above; the theorem states that a function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ satisfies

$$K(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{H}} \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{X} \tag{5.1}$$

for some Hilbert space $\mathcal{H}$ and function $\phi : \mathbb{X} \to \mathcal{H}$ if and only if K is a kernel function, as defined below.

Definition 5.1 (**Kernel functions**)**.** A function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a *kernel* if it is symmetric (i.e. $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{X}$) and is positive semidefinite, in the sense that for any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{X}$ and any $a_1, \ldots, a_n \in \mathbb{R}$, we have $\sum_{i,j=1}^n a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.

Thus, we need not even explicitly specify the non-linear feature mapping $\phi : \mathbb{X} \to \mathcal{H}$ we wish to use, and may instead directly specify a kernel function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$. Indeed, the kernel formalism allows us to apply linear learning algorithms to spaces $\mathbb{X}$ which do not even possess any intrinsic linear structure, such as discrete graphs (see e.g. Vishwanathan et al., 2010) and rankings (see e.g. Kondor and Barbosa, 2010). Intuitively, the kernel function can be thought of as specifying similarities between pairs of data points, with large positive values encoding similarity, and with a value of 0 corresponding to unrelatedness. We list some commonly used kernel functions over Euclidean space in Table 5.1.

Name	Gaussian kernel	Matérn kernel	White noise kernel
$K(\mathbf{x},\mathbf{y})$	$\sigma^2 \exp\left(-\frac{1}{2\lambda^2}\|\mathbf{x}-\mathbf{y}\|_2^2\right)$	$\sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\|\mathbf{x}-\mathbf{y}\|_2\right)^{\nu} K_\nu\left(\sqrt{2\nu}\|\mathbf{x}-\mathbf{y}\|_2\right)$	$\sigma^2 \mathbb{1}_{\{\mathbf{x}=\mathbf{y}\}}$

Table 5.1. Expressions for some commonly used kernels over Euclidean space, evaluated for an input pair $\mathbf{x},\mathbf{y}$. Each kernel has several hyperparameters: σ^2 and λ^2 in the case of the Gaussian kernel, and an additional hyperparameter ν in the case of the Matérn kernel. K_ν is the modified Bessel function of the second kind.

5.1.3 Gaussian process regression

As a running example in this section to aid intuition, we consider the problem of Bayesian linear regression, and its kernelised variant, Gaussian process (GP) regression. We posit a model relating an input $\mathbf{x} \in \mathbb{R}^d$ to a corresponding output $y \in \mathbb{R}$ according to

$$y \sim \mathcal{N}(\langle \beta, \mathbf{x} \rangle, \sigma^2), \tag{5.2}$$

with $\beta \in \mathbb{R}^d$ some unknown parameter vector, $\sigma^2 > 0$ a known variance parameter, and the Gaussian noise independent for each observation. We place a prior distribution $N(\mathbf{0}, I)$ over β, and consider a training data set $(\mathbf{x}_i, y_i)_{i=1}^N$. This induces a posterior distribution $p(\beta|(\mathbf{x}_i, y_i)_{i=1}^N)$ on the model parameters, which is available analytically (see Rasmussen and Williams, 2005, Chapter 2 for further details). In particular, writing $\mathbf{X} \in \mathbb{R}^{N\times d}$ for the matrix with i^{th} row $\mathbf{x}_i$ and $\mathbf{y} \in \mathbb{R}^N$ for the vector with i^{th} element y_i, the mean of the posterior predictive distribution for a novel input point $\mathbf{x}^* \in \mathbb{R}^d$ is

$$(\mathbf{x}^*)^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top + \sigma^2 I\right)^{-1} \mathbf{y}. \tag{5.3}$$

Note that $\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{N\times N}$ (often referred to as the Gram matrix) has $(i,j)^{\text{th}}$ element given by $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$, and $(\mathbf{x}^*)^\top \mathbf{X}^\top$ is a row vector with i^{th} element given by $\langle \mathbf{x}^*, \mathbf{x}_i \rangle$. Thus, the posterior predictive mean really does depend only on inner products between the input data points, illustrating the general principle discussed at the beginning of this section. Replacing these inner products with evaluations of a kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ yields the formula for the posterior predictive mean associated with a *Gaussian process*, the non-parametric kernel extension of Bayesian linear regression:

$$(\mathbf{k}^*)^\top \left(\mathbf{K} + \sigma^2 I\right)^{-1} \mathbf{y}, \tag{5.4}$$

where $\mathbf{K} \in \mathbb{R}^{N \times N}$ has $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ for all i, j, and $\mathbf{k}^* \in \mathbb{R}^N$, with $\mathbf{k}^*_i = K(\mathbf{x}^*, \mathbf{x}_i)$ for all i. This serves as an illustration for the *conceptual* ease by which linear methods can be extended to non-parametric kernel methods.

However, this example also serves to illustrate the computational burden associated with kernel methods as the number of data points N grows. Evaluating Expression (5.4) requires inverting the $N \times N$ matrix $\mathbf{K} + \sigma^2 I$, or at the very least solving a linear system involving this matrix. Either of these two linear-algebraic operations has complexity $\mathcal{O}(N^3)$ in full generality, which means that practically, exact computation with kernel methods quickly becomes expensive. The scalability of kernel methods is therefore limited, and in practical terms it is currently difficult to carry out *exact* inference for datasets of more than approximately $10^4 - 10^5$ examples[1] without further structure (Lee et al., 2018). Note that in contrast, the linear case is often more computationally tractable than the general non-linear case; in our particular example of the posterior predictive mean, we observe that the linear case in Expression (5.3) can be rewritten as

$$(\mathbf{x}^*)^\top (\mathbf{X}^\top \mathbf{X} + \sigma^2 I)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{5.5}$$

Assuming $d < N$, evaluating this expression now requires solving a d-dimensional linear system, with $\mathcal{O}(d^3)$ complexity, compared to the $\mathcal{O}(N^3)$ complexity associated with evaluating Expression (5.3). The overall complexity of evaluating Expression (5.5) is therefore dominated by construction of the matrix $\mathbf{X}^\top \mathbf{X}$, which has $\mathcal{O}(d^2 N)$ complexity. We emphasise that whilst this example deals specifically with Gaussian processes, linear structure often leads to fast algorithms for a wide range of problems, such as SVMs (Joachims, 2006).

There is therefore great interest in methods for *approximate* computation in kernel methods that scale more efficiently with dataset size than the exact methods discussed above. Several recent approaches to scaling kernel methods feature prominently in the literature, such as the Nyström method for low-rank approximation of the Gram matrix (Williams and Seeger, 2001). Here, we focus on another popular method, random feature methods, with perhaps the best known variety, random Fourier features, proposed by Rahimi and Recht (2007). Random feature methods exploit the fact that computation in linear learning algorithms is often much more tractable than in their non-linear counterparts, as illustrated in the Gaussian process example above.

[1] Under certain additional assumptions, such as particular structure in the set of input points (Izmailov et al., 2018) or in one-dimensional settings for specific kernels (Solin, 2016), it is possible to scale more efficiently than this.

5.1.4 Random features

Generally, random feature approximations for kernel methods rely on expressing a kernel $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ in the following manner:

$$K(\mathbf{x}, \mathbf{y}) = C\, \mathbb{E}_{\mathbf{w}\sim\mu}\left[\langle f(\mathbf{w}, \mathbf{x}), f(\mathbf{w}, \mathbf{y})\rangle\right], \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{X}, \tag{5.6}$$

for some distribution $\mu \in \mathscr{P}(\mathbb{R}^d)$, some fixed measurable function $f : \mathbb{R}^d \times \mathbb{X} \to \mathbb{R}^l$ (for some $l \in \mathbb{N}$), and some constant C. We take $C = 1$ in the following discussion to simplify notation. We shall see below that, perhaps surprisingly, such expressions are known for a wide variety of kernels of interest. The idea is then to take samples $\mathbf{w}_1, \ldots, \mathbf{w}_m \sim \mu$ (typically independently), and approximate Expression (5.6) using the standard Monte Carlo estimator:

$$K(\mathbf{x}, \mathbf{y}) \approx \widehat{K}(\mathbf{x}, \mathbf{y}) := \frac{1}{m}\sum_{i=1}^{m} \langle f(\mathbf{w}_i, \mathbf{x}), f(\mathbf{w}_i, \mathbf{y})\rangle, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{X}. \tag{5.7}$$

We then observe that defining the *random feature map* $\Phi_m : \mathbb{R}^d \to \mathbb{R}^{l\times m}$ as follows:

$$\Phi_m(\mathbf{x}) = \left(\frac{1}{\sqrt{m}} f(\mathbf{w}_i, \mathbf{x})\right)_{i=1}^{m}, \qquad \forall \mathbf{x} \in \mathbb{X}, \tag{5.8}$$

reveals that $\widehat{K}(\mathbf{x}, \mathbf{y}) = \langle \Phi_m(\mathbf{x}), \Phi_m(\mathbf{y})\rangle$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{X}$. Thus, a non-linear kernel method on a collection of input data $(\mathbf{x}_i)_{i=1}^N$ may be approximated by applying the corresponding linear method to the random features $(\Phi_m(\mathbf{x}_i))_{i=1}^N$. If the number of features m required for this approximation to be satisfactorily accurate is less than the number of data points N, this typically leads to computational gains, through the use of algorithms that exploit the linear structure of the problem, as in the case of GP regression and Bayesian linear regression described above. We refer the reader to Rahimi and Recht, 2007; Sutherland and Schneider, 2015 for example analyses of random feature approximation error.

In some sense, random feature approximations bring kernel methods round full circle; kernel methods are flexible, non-parametric, but computationally burdensome versions of linear methods, and a random feature approximation of a kernel method is a linear method which, in ideal circumstances, approaches the flexibility of the kernel method with the computational benefits associated with linear methods.

Having explained random feature approximations in the abstract, we now give several concrete examples.

Example 5.2 (**Random Fourier features**)**.** Rahimi and Recht (2007) introduced *random Fourier features* for stationary kernels $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (that is, kernels K for which $K(\mathbf{x}, \mathbf{y})$ depends only on $\mathbf{x} - \mathbf{y}$) via Bochner's theorem. The theorem states that for such a kernel K, there exists a positive finite measure $\eta \in \mathscr{M}(\mathbb{R}^d)$ such that the following Fourier identity holds:

$$K(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} \exp(\mathrm{i}\langle \mathbf{x} - \mathbf{y}, \mathbf{w} \rangle) \eta(\mathrm{d}\mathbf{w}) \,. \tag{5.9}$$

By standard trigonometric identities, this fits into the general random feature framework in Expression (5.6) by taking $\mu = \eta / \eta(\mathbb{R}^d)$, $C = \eta(\mathbb{R}^d)$, and $f(\mathbf{w}, \mathbf{x}) = (\cos(\langle \mathbf{w}, \mathbf{x} \rangle), \sin(\langle \mathbf{w}, \mathbf{x} \rangle))$ for all $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$.

Example 5.3 (**The angular kernel**)**.** The *angular kernel* $K : (\mathbb{R}^d \setminus \{\mathbf{0}\}) \times (\mathbb{R}^d \setminus \{\mathbf{0}\}) \to \mathbb{R}$ is defined by

$$K(\mathbf{x}, \mathbf{y}) = 1 - \frac{2}{\pi} \arccos\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \, \|\mathbf{y}\|} \right) . \tag{5.10}$$

The angular kernel, and the closely related notion of angular distance, have many important applications, including for example natural language processing (Sundaram et al., 2013), image processing (Honeine and Richard, 2010), and speaker representations (Schmidt et al., 2014). This kernel is also amenable to random feature approximation, via the identity

$$K(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\mathbf{w} \sim \eta} \left[\operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \operatorname{sign}(\langle \mathbf{w}, \mathbf{y} \rangle) \right] , \tag{5.11}$$

for any isotropic distribution $\eta \in \mathscr{P}(\mathbb{R}^d)$ satisfying $\eta(\{\mathbf{0}\}) = 0$.

5.2 Dimensionality reduction via random projections

Whilst kernel methods seek to work with (implicitly) high- (or infinite-) dimensional representations of data, it is often of interest generally to reduce the dimensionality of data before statistical analysis, a machine learning algorithm, or some other computational procedure is applied. Computational concerns are often a key motivation; data with reduced dimensionality typically takes less memory to store, and computation with such data typically takes less time.

Dimensionality reduction is thus of great concern in many algorithmic fields, and as such has been widely studied, with many different variants existing. Some methods work entirely with linear transformations of data (such as principal component analysis) whilst others make use of highly non-linear transformations (such as neural-network-based autoencoders). Additionally, some methods tailor the transforms used to reduce dimensionality to the particular data in question (such as in manifold learning), whilst other methods work in an entirely data-independent way.

Here, we give a brief account of random projections, a collection of data-independent methods for linear dimensionality reduction. Whilst non-linear methods and methods depending on the data in question may be able to recover additional low-dimensional structure, it turns out that if certain metric properties of the dataset are of concern, then it is possible to achieve very good performance with linear, data-independent mechanisms. Random projections also turn out to be instructive in our study of random features for kernel methods.

5.2.1 Random projections and the Johnson–Lindenstrauss transform

The key fundamental result behind randomised linear dimensionality reduction is the following, due to Johnson and Lindenstrauss (1984). It concerns distortions of pairwise distances of a given dataset; these are of particular interest, since many algorithms depend only on pairwise distances between points or inner products between pairs of points, which are necessarily approximately preserved when pairwise distances are approximately preserved.

Theorem 5.4 (**The Johnson–Lindenstrauss Lemma**)**.** Let $(\mathbf{x}_i)_{i=1}^N \subset \mathbb{R}^d$, and let $\varepsilon \in (0,1)$. Then if $m > 8\varepsilon^{-2}\log N =: k(\varepsilon, N)$, there exists a linear map $A : \mathbb{R}^d \to \mathbb{R}^m$ such that

$$(1-\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|A(\mathbf{x}_i) - A(\mathbf{x}_j)\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2, \quad \text{for all } i, j \in \{1, \ldots, N\}. \tag{5.12}$$

Intuitively, the result states that it is possible to linearly project N points in $\mathbb{R}^d$ into a space with dimension proportional to $\log N$ (with no dependence on d), and approximately preserve all pairwise distances — a result that at first appears somewhat surprising. Perhaps even more surprisingly, it turns out that such maps are not "hard" to compute given a dataset, but can often be found with high probability by sampling from a distribution over projections in an entirely data-independent way. Indeed, the original proof of Theorem 5.4 considers a randomised projection onto a $k(\varepsilon, N)$-dimensional subspace

of $\mathbb{R}^d$ chosen according to Haar measure on the corresponding Grassmannian manifold (informally, chosen "uniformly at random" across all such subspaces), which achieves the low-distortion embedding of Theorem 5.4 with high probability. This observation — that high-quality dimensionality reduction can be achieved in an entirely data-independent manner — motivates the study of random projections.

Later work shows that even simpler random projections can be used to achieve the same asymptotic dependence on N and ε. Dasgupta and Gupta (2003) show, for example, that taking A to be multiplication by a matrix with i.i.d. Gaussian entries can also achieve the low-distortion embedding of Theorem 5.4 with high probability.

Finally, as well as being an important problem in its own right, we remark that studying random projections for dimensionality reduction is also instructive in terms of understanding random feature maps for kernel methods. Indeed, random projections can be interpreted as an approximate random feature method for the linear kernel $K(\mathbf{x},\mathbf{y}) = \langle \mathbf{x},\mathbf{y}\rangle$, with the representation $K(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \left[\langle \mathbf{w},\mathbf{x}\rangle \langle \mathbf{w},\mathbf{y}\rangle\right] \eta(\mathrm{d}\mathbf{w})$, for suitable measures $\eta \in \mathscr{M}(\mathbb{R}^d)$. The lack of a non-linearity in this representation of the kernel makes analysis particularly tractable, and insights that can be gained from the analysis of this simple case inform the more general case of random feature approximations.

5.3 Improvements to the Johnson–Lindenstrauss transform and random feature methods

There are several key computational and statistical concerns associated with random projections and random feature methods. In this thesis, we consider the following:

(i) **Sampling costs**: the computational (pre-processing) cost of sampling the random projection (in the case of dimensionality reduction), or the random feature map in the case of kernel approximation. In some circumstances, we may also be interested in the *randomness budget* of an algorithm; that is, how many (pseudo)random bits an implementation requires, if for example generation or (secure) communication of these bits is expensive. For further discussion of these points in the context of random projections and random feature maps, see e.g. Choromanski and Sindhwani (2016); Yu et al. (2016).

(ii) **Storage of projection map**: the memory (space) cost of storing the random projection or random feature map. Along with embedding costs, this is one of the principal

motivating factors for improvements to random projection and random feature methods, and is discussed extensively in the literature; see Ailon and Chazelle (2009); Dasgupta et al. (2010); Le et al. (2013) for representative discussions.

(iii) **Embedding computation**: the computational cost of computing the projection of random features associated with each datapoint.

(iv) **Accuracy**: the statistical accuracy of the reconstructions of the quantities of interest, such as pairwise distances or kernel evaluations, from the random projections and/or random features. Whilst many improvements to random projections and random feature maps focus on computational issues, there has been recent progress in improving accuracy too, such as utilising orthogonal couplings (Yu et al., 2016) — this point is discussed in much more detail in Chapters 6 and 7.

Improvements in these computational, space, and statistical considerations leads to more accurate methods that can scale to larger quantities of data, and as such there has been much recent interest in improving methods for random projection and random feature maps in these directions. We present a brief overview of previous work in these two domains below.

5.3.1 Scalable random projections

To fix ideas appearing in Section 5.2.1, we begin with a precise definition of the Johnson–Lindenstrauss transform. Suppose we want to project a d-dimensional dataset $(\mathbf{x}_i)_{i=1}^N \subset \mathbb{R}^d$ into m dimensions. The *Johnson–Lindenstrauss transform* (JLT) is then defined by taking $\mathbf{A} \in \mathbb{R}^{m\times d}$ to be a draw from a probability distribution over $m \times d$ real matrices, where each entry has independent distribution $\mathcal{N}(0,1)$. Each datapoint $(\mathbf{x}_i)_{i=1}^N$ is then embedded into $\mathbb{R}^m$ according to the map

$$\mathbf{x} \mapsto \frac{1}{\sqrt{m}}\mathbf{A}\mathbf{x}. \tag{JLT}$$

It can be shown that with $m = \mathcal{O}(\varepsilon^{-2}\log N)$, as in Theorem 5.4, this linear map attains the condition in Expression (5.12) with high probability. We comment briefly on the four considerations introduced above in the context of the JLT. (i) Sampling costs: all elements of $\mathbf{A}$ are sampled independently, resulting in a sampling cost of $\mathcal{O}(md)$, and a randomness budget of md times the number of pseudorandom bits used to construct a full-precision float. (ii) Map storage: since the transformation matrix $\mathbf{A}$ is unstructured, all elements of the matrix must be stored, requiring $\mathcal{O}(md)$ space. (iii) Embedding computation: since

the matrix $\mathbf{A}$ is unstructured, computing the embedding for a single datapoint involves a single matrix-vector multiplication, requiring $\mathcal{O}(md)$ floating-point multiplications. Using $m = \mathcal{O}(\varepsilon^{-2}\log N)$ as in Theorem 5.4, this gives a dependence on N as $\mathcal{O}(\varepsilon^{-2}d\log N)$. (iv) Accuracy: the embedding follows a concentration inequality-type guarantee in accordance with Theorem 5.4.

Substantial progress regarding improved computational properties of random projections was achieved by Ailon and Chazelle (2006), who introduced the *fast Johnson–Lindenstrauss transform* (FJLT), which exploits fast matrix-vector computations, and expresses the transform in terms of three matrices, $\mathbf{P}, \mathbf{H}, \mathbf{D}$, as follows:

$$\mathbf{x} \mapsto \mathbf{PHDx}. \tag{FJLT}$$

Here $\mathbf{D}$ is a random diagonal matrix with i.i.d. Unif($\{\pm 1\}$) entries (Rademacher random variables) on the diagonal, $\mathbf{H}$ is the normalised $d \times d$ Hadamard matrix (a deterministic orthogonal matrix supporting $\mathcal{O}(d\log d)$ matrix-vector products, for which we give a full account in Chapter 6), and $\mathbf{P}$ is a sparse $m \times d$ projection matrix, with independent entries having distribution given by a mixture $q\delta_0 + (1-q)\mathcal{N}(0, q^{-1})$, where $q = \min(\Theta(d^{-1}\log(N)^2), 1)$, and $\delta_0 \in \mathscr{P}(\mathbb{R})$ is the Dirac delta distribution at 0. The transform is given by sampling the matrices $\mathbf{P}$ and $\mathbf{D}$ from their corresponding distributions, and computing the embedding.

The concentration-inequality-based accuracy results that Ailon and Chazelle (2006) give are comparable to those achieved by the standard JLT; the principal contribution of the FJLT is to lower the cost of computing the embedding of a dataset. Indeed, matrix-vector products with $\mathbf{D}$ take $\mathcal{O}(d)$ floating-point multiplications, matrix-vector products with $\mathbf{H}$ take $\mathcal{O}(d\log d)$ floating-point multiplications, and matrix-vector products with $\mathbf{P}$ take $\mathcal{O}(\|\mathbf{P}\|_0)$ floating-point operations, where $\|\mathbf{P}\|_0$ is the number of non-zero entries of $\mathbf{P}$. It can be shown (Ailon and Chazelle, 2009) that with high probability, this latter quantity is close to its mean, $m\log^2 N$. The FJLT additionally allows some savings to be made in terms of sampling costs (each diagonal element of $\mathbf{D}$ requires a single random bit, and all zero entries of $\mathbf{P}$ require only a single random bit), and map storage (the projection matrix $\mathbf{P}$ may be more efficiently stored than the dense matrix $\mathbf{A}$ of the standard JLT by using coding techniques that exploit its sparsity), although the asymptotic dependence is still $\mathcal{O}(md)$.

Following Ailon and Chazelle (2006), a variety of other computationally efficient random projection methods were proposed, notably the *sparse Johnson–Lindenstrauss transform* (Dasgupta et al., 2010), which defines a distribution over sparse projection matrices, and

exploits fast sparse matrix-vector multiplication algorithms to achieve speed-ups in computing embeddings. Other approaches include utilising fast matrix-vector multiplication algorithms for circulant matrices (Hinrichs and Vybíral, 2011; Vybíral, 2011; Zhang and Cheng, 2013).

5.3.2 Scalable random features

Work on scaling up random feature approximations appeared in the literature several years after the key works scaling up random projections described in Section 5.3.1. Following on from the explanation of random Fourier features in Example 5.2, we note that the random features computed take the particular form

$$\Phi_m(\mathbf{x}) = (\cos(\langle \mathbf{w}_i, \mathbf{x} \rangle), \sin(\langle \mathbf{w}_i, \mathbf{x} \rangle))_{i=1}^m , \tag{5.13}$$

with the vectors $(\mathbf{w}_i)_{i=1}^m$ drawn i.i.d. from the Fourier distribution $\eta \in \mathscr{P}(\mathbb{R}^d)$ of interest. We may re-express this as first computing a linear embedding $\mathbf{x} \mapsto \mathbf{W}\mathbf{x}$, followed by applying sine and cosine non-linearities coordinate-wise to the resulting vector, where $\mathbf{W} \in \mathbb{R}^{m \times d}$ has i^{th} row given by $\mathbf{w}_i$, for $i = 1, \ldots, m$.

Le et al. (2013) proposed replacing the dense unstructured matrix-vector multiplication $\mathbf{x} \mapsto \mathbf{W}\mathbf{x}$ involved in computing the random features above with the following product:

$$\mathbf{x} \mapsto \mathbf{SHG\Pi HD}\mathbf{x}. \tag{5.14}$$

Here, $\mathbf{H}$ is the scaled Hadamard matrix, and $\mathbf{D}$ is a random diagonal matrix, both of which appeared in the Fast Johnson–Lindenstrauss transform described in Section 5.3.1. $\mathbf{G}$ is a diagonal matrix with random Gaussian entries, $\mathbf{\Pi}$ is a random permutation matrix, and $\mathbf{S}$ is a random diagonal scaling matrix (with the exact distribution of $\mathbf{S}$ depending on the kernel to be approximated). The authors observed that this transform leads to improved theoretical embedding computational costs and map storage relative to unstructured random features, and achieved both comparable concentration results and empirical performance; in their experiments, it was possible to achieve orders of magnitude improvements in runtime and RAM usage on SVM classification experiments. Following this paper, several other approaches to computationally cheaper random feature methods have been proposed in the literature, such as that of Choromanski and Sindhwani (2016), which utilises circulant matrices and associated fast matrix-vector product algorithms rely-

ing on the discrete Fourier transform. In Chapter 7, we investigate alternative approaches to the structured embedding in Expression (5.14).

5.4 Outline of original work

Having established the relevant background material, we now give a brief outline of our original contributions in this second part of the thesis. A more detailed list of contributions may be found in each of Chapters 6 and 7.

- In Chapter 6, we study linear dimensionality reduction methods using random orthogonal matrices, and in particular study fast discrete approximation sampling algorithms based on Hadamard matrices. We establish theoretical results that quantify the performance of random projections using these approximate sampling methods, and evaluate their performance empirically on a variety of datasets.
- In Chapter 7, we study orthogonality as a means of improving statistical efficiency of random feature approximations. We contribute new theoretical results for stationary kernels and for the angular kernel, and demonstrate the empirical benefits of orthogonality for random feature maps in Gram matrix estimation and approximate Gaussian process regression.

Chapter 6

Fast Dimensionality Reduction with Hadamard–Rademacher Random Matrices

This chapter is based on material from the following publication:

- Choromanski, K.*, Rowland, M.*, and Weller, A. (2017). The unreasonable effectiveness of structured random orthogonal embeddings. In *Neural Information Processing Systems (NIPS)*. [*=equal contribution].

More precisely, the chapter gives an account of the work on linear dimensionality reduction appearing in this paper (the other work in the paper, on random features for the angular kernel, appears in Chapter 7). Work was carried out jointly with Krzysztof Choromanski and Adrian Weller; the theoretical results in this chapter are principally due to the author of the thesis and Krzysztof Choromanski, and the experiments in Section 6.5 were designed, implemented, and run by the author of the thesis. The original paper was written jointly with Krzysztof Choromanski and Adrian Weller, although many parts have been rewritten for the purposes of this thesis.

6.1 Introduction

In this chapter, we consider the problem of linear dimensionality reduction via random projections, as described in Section 5.2. We consider several families of probability distributions over matrices, and then study random projection algorithms based on these

random matrices which offer improvements over the standard JLT in terms of the four criteria discussed in Section 5.3, namely (i) sampling costs, (ii) map storage, (iii) embedding cost, and (iv) statistical efficiency.

Our approach is centred around random projections based on random matrix distributions which are concentrated on matrices with orthogonal rows. We refer to random matrices following these distributions as random ortho-matrices. Matrices with orthogonal rows are natural to consider for dimensionality reduction, and were amongst the first constructions given for random projections by Johnson and Lindenstrauss (1984). In addition, orthogonality has recently been observed to improve statistical performance in related tasks, such as random feature approximations for kernel methods (Yu et al., 2016). We will consider several families of probability distributions over such matrices.

First, we consider a distribution over such matrices where the direction of each matrix row is marginally distributed uniformly on the unit hypersphere, and has length distribution given by that corresponding to a Gaussian random vector. We term these *Gaussian orthogonal matrices*. We will see that working with such matrices can lead to benefits in statistical performance in comparison to random matrices using independent rows, as in the standard JLT. However, this approach comes at the cost of increased sampling costs, as non-trivial linear algebra computations are required to sample Gaussian orthogonal matrices in the first place.

To avoid these increased sampling costs, and to make improvements in embedding costs, several authors have recently explored alternative random matrix distributions, based on composing combinations of deterministic Hadamard matrices with random diagonal matrices (precise definitions are given in Section 6.2). We refer to the composition of a Hadamard matrix $\mathbf{H}$ with a random diagonal matrix $\mathbf{D}$ as a $\mathbf{HD}$ block. For dimensionality reduction, the fast JLT of Ailon and Chazelle (2006) (see Section 5.3) uses a single $\mathbf{HD}$ block as a way of spreading out (with high probability) the mass of a vector over all dimensions before applying a sparse Gaussian projection matrix. Choromanski and Sindhwani (2016) also used just one $\mathbf{HD}$ block as part of a larger structure in approximate random feature maps. Bojarski et al. (2017) and Andoni et al. (2015) discussed using 3 $\mathbf{HD}$ blocks for locality-sensitive hashing methods but gave no concrete results for their application to dimensionality reduction or kernel approximation, and left open the theoretical question as to why $k = 3$ blocks work well whilst 1 or 2 blocks can have noticeably worse performance. Yu et al. (2016) also explore using a number k of $\mathbf{HD}$ blocks for Gaussian kernel approximation empirically, observing good computational and statistical performance for $k = 3$, but without any theoretical accuracy guarantees. It was left as an open question as

to why matrices formed by a small number of **HD** blocks can outperform non-discrete transforms. To address some of the open questions raised, we unify these collections of distributions, considering a general family of random orthogonal matrices, which we term **SD**-product matrices. These are formed by multiplying some number k of **SD** blocks, each of which is highly structured. Here **S** is a structured matrix satisfying certain conditions (the Hadamard matrix **H** being a special case), and **D** is a random diagonal matrix. These random matrices are defined precisely in Section 6.2.

Here we provide non-asymptotic results which show that using random ortho-matrices, both Gaussian orthogonal and **SD**-product, yield improved MSE for dimensionality reduction. We provide analysis to understand why $k = 3$ can outperform the JLT, why odd k yields different behaviour to even k, and why higher values of k deliver decreasing additional benefits (see Section 6.3 and Section 6.4). Indeed, a large source of motivation for this work is to develop theoretical understanding of the structured random matrices described above, and the linear dimensionality setting turns out to be particularly amenable to such analysis.

These structured random matrices turn out to offer improved sampling costs, randomness budget, map storage costs, and embedding computation costs. This comes at the cost of losing the marginal uniformity of row direction distributions, but it turns out that these random projections still lead to improved MSE in inner product reconstruction relative to the standard JLT.

We conclude this section by highlighting the following contributions:

- In Section 6.3: The *Orthogonal Johnson–Lindenstrauss Transform* (OJLT) for dimensionality reduction. We prove that the OJLT offers superior sampling costs, map storage, embedding computation, and accuracy (as measured by MSE) relative to the standard JLT.
- In Section 6.4: A Markov chain perspective on Hadamard–Rademacher matrices, a paticular variant of **SD**-product matrices, that builds intuitive understanding of the OJLT, and casts light on the effectiveness, observed in other domains, of Hadamard–Rademacher random matrices. In particular, it will also be useful in understanding the use of Hadamard–Rademacher random matrices in Chapter 7.
- In Section 6.5: Empirical validation of our theoretical results and further experiments on a range of data sets.

6.2 Random ortho-matrices

We begin by precisely defining the random matrices we work with for the remainder of the chapter. We consider two general families of probability distributions over matrices with orthogonal rows, introduced below.

Definition 6.1 (**Gaussian orthogonal matrices**)**.** Let $m \leq d$. A Gaussian orthogonal matrix is a random matrix $\mathbf{G}_{\text{ort}}$ taking values in $\mathbb{R}^{m\times d}$ whose distribution is given by the regular conditional probability distribution of the unstructured Gaussian matrix $\mathbf{G}$, conditional on the event that the rows of $\mathbf{G}$ are orthogonal.

There are several ways in which a Gaussian orthogonal matrix may be sampled. One method is to (i) sample a uniformly random orthonormal set of m vectors in $\mathbb{R}^d$ (more precisely, sample an element of the Steifel manifold from the unique probability distribution invariant to the action of the orthogonal group on the Steifel manifold), then (ii) independently scale each of the resulting vectors so that their norms match those of Gaussian vectors in distribution, and finally (iii) stack the vectors as rows of a matrix.

The sampling of a uniformly random orthogonal set can be performed by sampling a set of m linearly independent vectors in $\mathbb{R}^d$ from any isotropic distribution on $\mathbb{R}^d$, and then performing Gram-Schmidt orthogonalisation on this set.

Our second class of distributions is motivated by the desire to obtain similar statistical benefits of orthogonality to $\mathbf{G}_{\text{ort}}$, whilst gaining computational efficiency by employing more structured matrices.

Definition 6.2 (**SD-product matrices**)**.** An **SD**-product matrix is a random matrix of the form

$$\prod_{i=1}^{k} \mathbf{SD}_i = (\mathbf{SD}_k)\cdots(\mathbf{SD}_1)\,. \tag{6.1}$$

Here $\mathbf{S} \in \mathbb{R}^{d\times d}$ is a fixed matrix with orthogonal rows, with $|\mathbf{S}_{ij}| = \frac{1}{\sqrt{d}}\ \forall\, i, j \in \{1,\ldots,d\}$, and the $(\mathbf{D}_i)_{i=1}^{k}$ are independent random diagonal matrices. Some authors refer to $\mathbf{S}$ as a (scaled) Hadamard matrix, but we reserve this term for a more specific family of matrices satisfying these conditions — see Expression (6.2).

This class includes as particular cases several recently introduced random matrices (see e.g. Andoni et al. (2015); Yu et al. (2016)), where good *empirical* performance was observed. We go further to establish strong theoretical guarantees in Section 6.3.

A prominent example of **S** matrices is the family of normalised *Hadamard* matrices, defined recursively by $\mathbf{H}_1 = (1) \in \mathbb{R}^{1\times 1}$, and then for $i > 1$:

$$\mathbf{H}_l = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{H}_{l-1} & \mathbf{H}_{l-1} \\ \mathbf{H}_{l-1} & -\mathbf{H}_{l-1} \end{pmatrix} \in \mathbb{R}^{2^{l-1}\times 2^{l-1}}. \tag{6.2}$$

Importantly, matrix-vector products with a Hadamard matrix $\mathbf{H} \in \mathbb{R}^{d\times d}$ are computable in $O(d \log d)$ time via the fast Walsh-Hadamard transform, yielding large computational savings relative to matrix-vector products with unstructured dense matrices. In addition, Hadamard matrices enable a significant space advantage for map storage: since the fast Walsh-Hadamard transform can be computed without explicitly storing **H**, only kd bits are required, to store the diagonal elements of $(\mathbf{D}_i)_{i=1}^k$. Note that these $\mathbf{H}_l$ matrices are defined only for d a power of 2, but if needed, one can always adjust the input data by padding with 0s to enable the use of "the next smallest **H**", doubling the number of dimensions in the worst case.

The Hadamard matrices **H** are representatives of a much larger family in **S** which also attains computational savings. These are L_2-normalised versions of Kronecker-product matrices of the form $\mathbf{A}_1 \otimes \ldots \otimes \mathbf{A}_l \in \mathbb{R}^{d\times d}$ for $l \in \mathbb{N}$, where $\otimes$ stands for the Kronecker product and the blocks $\mathbf{A}_i$ have entries of the same magnitude and orthogonal rows. For these matrices, matrix-vector products are also computable in $\mathcal{O}(d \log d)$ time (Zhang et al., 2015).

For the random diagonal matrices $(\mathbf{D}_i)_{i=1}^k$, we principally consider the distribution given by taking diagonal elements to be i.i.d. Rademacher random variables (i.e. Unif($\{\pm 1\}$), leading to the following random matrices.

Definition 6.3 (**S-Rademacher random matrices**)**.** The **S**-*Rademacher* random matrix with $k \in \mathbb{N}$ blocks is an **SD**-product random matrix, given specifically by the following formula, where $(\mathbf{D}_i^{(\mathscr{R})})_{i=1}^k$ are independent random diagonal matrices, with i.i.d. Rademacher random variables (i.e. Unif($\{\pm 1\}$)) on the diagonals:

$$\prod_{i=1}^{k} \mathbf{S}\mathbf{D}_i^{(\mathscr{R})}. \tag{6.3}$$

We will refer to **S**-Rademacher random matrices that specifically use a Hadamard structured matrix **H** as *Hadamard–Rademacher random matrices*. We will also be interested

more generally in distributions over the unit circle S^1 in the complex plane $\mathbb{C}$ — see Section 6.3.5.

Finally, since **SD**-product matrices are square, we will consider *subsampled* versions of them to achieve dimensionality reduction.

Definition 6.4 (**Subsampled SD-product matrix**)**.** A subsampled **SD**-product matrix is given by prepending a coordinate projection matrix $\mathbf{P} \in \mathbb{R}^{m \times d}$ to an **SD**-product matrix, resulting in a random matrix of the form

$$\mathbf{PSD}_k \cdots \mathbf{SD}_1 \,. \tag{6.4}$$

We will consider three types of coordinate projection matrices:

1. *Sampling with replacement*, in which each row of **P** is drawn independently from the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ of canonical basis vectors;
2. *Sampling without replacement*, in which **P** follows the distribution of a sampling with replacement coordinate projection matrix as above, conditioned on the event that all rows are distinct;
3. *First m rows*, in which the **P** is deterministic and simply selects the first m elements of the vector for projection.

Our motivation for considering these three varieties is as follows. Sampling coordinates without replacement eliminates the possibility of duplication, but otherwise treats all coordinates symmetrically. Sampling with replacement allows for the possibility of duplication, which we might expect to worsen statistical performance — indeed this is the case (see Theorem 6.11). Because of this, we will primarily be interested in subsampling without replacement, but see Section 6.3.6 for further discussion of subsampling with replacement. Finally, sampling the first m coordinates breaks symmetry, which we might again expect to have knock-on effects on statistical performance. However, in the case where the structured matrix **S** is a Hadamard matrix **H**, there are additional adjustments that can be made to the fast Walsh-Hadamard transform to exploit the fact that only a small collection of contiguous coordinates are required — we explain this point in more detail in Section 6.3.6. Unless specifically mentioned, in what follows we assume that the sampling without replacement strategy is used.

Having described our two main families of random matrices, the Gaussian orthogonal matrix $\mathbf{G}_{\text{ort}}$ and subsampled **SD**-product matrices, we next analyse them in the context of dimensionality reduction.

6.3 The Orthogonal Johnson–Lindenstrauss Transform

Let $\mathcal{X} = (\mathbf{x}_i)_{i=1}^N \subset \mathbb{R}^d$ be a dataset of d-dimensional real vectors. As described in Section 5.3, the standard JLT random projection uses the randomised linear map $\mathbf{x} \mapsto \frac{1}{\sqrt{m}}\mathbf{G}\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^d$, where $\mathbf{G} \in \mathbb{R}^{m\times d}$ is a draw from the random matrix distribution where each entry is drawn independently from $\mathcal{N}(0,1)$.

Our Orthogonal Johnson–Lindenstrauss Transform (OJLT) is obtained by replacing the unstructured random matrix $\mathbf{G} \in \mathbb{R}^{m\times d}$ with a random ortho-matrix from Section 6.2: either $\mathbf{G}_{\text{ort}}$, yielding the random projection $\mathbf{x} \mapsto \frac{1}{\sqrt{m}}\mathbf{G}_{\text{ort}}\mathbf{x}$, or a subsampled (without replacement) $\mathbf{S}$-Rademacher matrix $\mathbf{PSD}_k^{(\mathscr{R})}\cdots\mathbf{SD}_1^{(\mathscr{R})}$, yielding the random projection $\mathbf{x} \mapsto \frac{\sqrt{d}}{\sqrt{m}}\mathbf{PSD}_k^{(\mathscr{R})}\cdots\mathbf{SD}_1^{(\mathscr{R})}\mathbf{x}$.

We now compare these three random projection methods according to the four criteria introduced in Section 5.3, namely: (i) sampling costs; (ii) map storage; (iii) embedding computation; (iv) statistical accuracy. In general, we will be concerned with the variant of coordinate subsampling that operates without replacement of indices (as described in Definition 6.4), but discuss alternatives in Section 6.3.6.

6.3.1 Sampling costs

To sample an unstructured Gaussian matrix $\mathbf{G} \in \mathbb{R}^{m\times d}$ for the JLT, we require md independent Gaussian random variables to be sampled, representing an overall computational cost of $\mathcal{O}(md)$. Further, a randomness budget of md times the number of random bits required to represent a draw from a scalar Gaussian random variable (the exact quantity will depend on the number of bits used to represent a floating-point number in a given implementation) is required to sample from the matrix distribution. In contrast, as alluded to in Section 6.1, sampling $\mathbf{G}_{\text{ort}}$ requires additional linear algebra to orthogonalise the directions of the vectors concerned. Thus, the randomness budget is the same as for sampling the unstructured Gaussian matrix, but the computational cost of computing the transformation matrix is $\mathcal{O}(m^2 d)$, due to the linear-algebraic overhead involved in computing the orthonormal basis (via e.g. Gram-Schmidt orthogonalisation). However, for **SD**-product matrices, the only sampling required is of the elements of the diagonal matrices $(\mathbf{D}_i)_{i=1}^k$ (requiring kd random bits), and the selection of projection coordinates for the matrix $\mathbf{P}$ (which can be implemented with $m\lceil\log(d+m)\rceil$ random bits).

6.3.2 Map storage

For $\mathbf{G}$ and $\mathbf{G}_{\text{ort}}$, the dense sampled matrices themselves must be stored, requiring memory for md floating-point numbers. In contrast, for **SD**-product matrices, if matrix-vector products with $\mathbf{S}$ are computable via a fast transform algorithm, then it is not necessary to store the matrix $\mathbf{S}$ itself in memory; this is the case, for example, with the normalised Hadamard matrix $\mathbf{H}$, for which the fast Walsh-Hadamard transform may be used to evaluate matrix-vector products. Only the kd diagonal elements of the $(\mathbf{D}_i)_{i=1}^k$ matrices and the projection indices of $\mathbf{P}$ need be stored, requiring kd bits for the diagonal elements, and m integers for the projection indices.

6.3.3 Embedding computation

Both $\mathbf{G}$ and $\mathbf{G}_{\text{ort}}$ are dense matrices without any particular structure that aids computation of matrix-vector products, and hence computing embeddings with these matrices costs the usual matrix-vector product cost of $\mathcal{O}(md)$. In contrast, embeddings with **SD**-product matrices for which $\mathbf{S}$ admits fast matrix-vector products may be computed faster. For example, embeddings with Hadamard–Rademacher random matrices may be computed as follows: matrix-vector products with a diagonal matrix $\mathbf{D}_i^{(\mathscr{R})}$ may be computed in $\mathcal{O}(d)$ time, matrix-vector products with $\mathbf{H}$ may be computed in $\mathcal{O}(d\log d)$ time via the fast Walsh-Hadamard transform, and finally coordinate projection via $\mathbf{P}$ may be computed in $\mathcal{O}(d)$ time. Considering that there are k **HD** blocks, the overall embedding cost is $\mathcal{O}(kd\log d)$. Thus, roughly speaking, when $m > k\log d$, we can expect computational advantages from using **SD**-product matrices; since we generally take $k \leq 3$, and m is typically taken to be of the order $\mathcal{O}(\log N)$, in line with Theorem 5.4, this is generally the case.

6.3.4 Statistical accuracy

Given two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, the corresponding estimators of their inner product after the random projection are denoted by:

$$\widehat{K}_m^{\text{base}}(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \langle \mathbf{G}\mathbf{x}, \mathbf{G}\mathbf{y} \rangle \,, \tag{6.5}$$

$$\widehat{K}_m^{\text{ort}}(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \langle \mathbf{G}_{\text{ort}}\mathbf{x}, \mathbf{G}_{\text{ort}}\mathbf{y} \rangle \,, \tag{6.6}$$

$$\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y}) = \frac{d}{m} \left\langle \mathbf{PSD}_k^{(\mathscr{R})} \cdots \mathbf{SD}_1^{(\mathscr{R})} \mathbf{x}, \mathbf{PSD}_k^{(\mathscr{R})} \cdots \mathbf{SD}_1^{(\mathscr{R})} \mathbf{y} \right\rangle . \tag{6.7}$$

Previous work on fast versions of random projection methods has focused on providing concentration guarantees at an asymptotically optimal rate; see (Ailon and Chazelle, 2009; Dasgupta et al., 2010) for example analyses. In contrast, we focus here on understanding mean squared error (MSE) of inner product estimators based on random projections. Precisely, the MSE of an estimator $\widehat{K}(\mathbf{x}, \mathbf{y})$ of the inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ is defined to be $\text{MSE}(\widehat{K}(\mathbf{x}, \mathbf{y})) = \mathbb{E}\left[(\widehat{K}(\mathbf{x}, \mathbf{y}) - \langle \mathbf{x}, \mathbf{y} \rangle)^2 \right]$. It is possible to perform exact analysis, and thus we are able to directly compare the performance of the three methods concerned, under the measure of MSE.

We contribute the following closed-form expressions, which exactly quantify the MSE for these three estimators.

Lemma 6.5. The unstructured JLT inner product estimator $\widehat{K}_m^{\text{base}}(\mathbf{x}, \mathbf{y})$ of the inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ using m-dimensional random projections is unbiased, with

$$\text{MSE}(\widehat{K}_m^{\text{base}}(\mathbf{x}, \mathbf{y})) = \frac{1}{m} (\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2) \,. \tag{6.8}$$

Theorem 6.6. The estimator $\widehat{K}_m^{\text{ort}}(\mathbf{x}, \mathbf{y})$ of the inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ is unbiased and satisfies[1], for $d \geq 4$:

$$\text{MSE}(\widehat{K}_m^{\text{ort}}(\mathbf{x}, \mathbf{y})) = \text{MSE}(\widehat{K}_m^{\text{base}}(\mathbf{x}, \mathbf{y})) - \frac{m-1}{m(d-1)} \left(\langle \mathbf{x}, \mathbf{y} \rangle^2 + \frac{d}{d+2} \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \right) . \tag{6.9}$$

Theorem 6.7. The OJLT estimator $\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})$ of the inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, with k blocks, using m-dimensional random projections and uniform subsampling policy

[1] The author thanks Sergio Bacallado for providing the simplified expression for the MSE given here.

without replacement, is unbiased with

$$\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})) = \frac{1}{m}\left(\frac{d-m}{d-1}\right)\Big((\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \sum_{r=1}^{k-1}\frac{(-1)^r 2^r}{d^r}(2\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \frac{(-1)^k 2^k}{d^{k-1}}\sum_{i=1}^{d} x_i^2 y_i^2\Big). \tag{6.10}$$

Proof (Sketch). For $k = 1$, the random projection matrix is given by subsampling rows from $\mathbf{SD}_1$, and the computation can be carried out directly. For $k > 1$, the proof proceeds by induction. The random projection matrix in the general case is given by subsampling rows of the matrix $\mathbf{SD}_k \cdots \mathbf{SD}_1$. By writing the MSE as an expectation and using the law of conditional expectations (conditioning on the value of the first $k-1$ random matrices $\mathbf{D}_{k-1},\ldots,\mathbf{D}_1$), the statement of the theorem for 1 **SD** block and for $k-1$ **SD** blocks can be neatly combined to yield the result. □

To our knowledge, it has not previously been possible to provide precise quantitative guarantees that **SD**-product matrices outperform i.i.d. matrices. Combining Lemma 6.5 with Theorem 6.7 yields the following important result.

Corollary 6.8 (**Theoretical guarantee of improved performance**)**.** The OJLT estimator $\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})$ (using subsampling without replacement) yields guaranteed lower MSE than $\widehat{K}_m^{\mathrm{base}}(\mathbf{x},\mathbf{y})$, for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$.

It is not yet clear when $\widehat{K}_m^{\mathrm{ort}}$ is better or worse than $\widehat{K}_m^{(k)}$; we explore this empirically in Section 6.5. Theorem 6.7 reveals several interesting facts about how random projection performance of the OJLT varies with the number of blocks k used, which sheds light on the properties of **SD**-product matrices more generally. From Expression (6.10), it is clear that there are diminishing MSE benefits to using a large number k of **SD** blocks. Interestingly, there is also an alternating effect as k changes between odd and even. We explore this notion in greater depth in Section 6.4, through connections to periodicity in Markov chains. We observe that the results above offer quantitative confirmation of empirical observations recently noted in the works of Yu et al. (2016) and Andoni et al. (2015) on the behaviour of Hadamard–Rademacher random matrices.

Theorem 6.7 is a key result in this chapter, demonstrating that **SD**-product matrices yield both statistical and computational improvements compared to the standard JLT procedure, which is widely used in practice. In the remainder of this section, we discuss several further variations on the OJLT.

6.3.5 Complex variants of the OJLT

We show that the MSE benefits of Theorem 6.7 may be markedly improved by using **SD**-product matrices with complex entries. Specifically, we consider the variant **S**-*Hybrid* random matrix $\mathbf{SD}_k^{(\mathscr{U})} \prod_{i=1}^{k-1} \mathbf{SD}_i^{(\mathscr{R})}$, where $\mathbf{D}_k^{(\mathscr{U})}$ is a diagonal matrix with i.i.d. Unif(S^1) random variables on the diagonal, independent of $(\mathbf{D}_i^{(\mathscr{R})})_{i=1}^{k-1}$, and S^1 is the unit circle of $\mathbb{C}$. Observe that an **S**-Hybrid matrix is unitary, but generally not orthogonal. We thus use the real part of the Hermitian product between projections as an inner product estimator; recalling the definitions of Section 6.2, we obtain the random projection $\mathbf{x} \mapsto \frac{\sqrt{d}}{\sqrt{m}} \mathbf{PSD}_k^{(\mathscr{U})} \left[\prod_{i=1}^{k-1} \mathbf{SD}_i^{(\mathscr{R})} \right] \mathbf{x}$, and denote the corresponding inner product estimator for two points $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ by

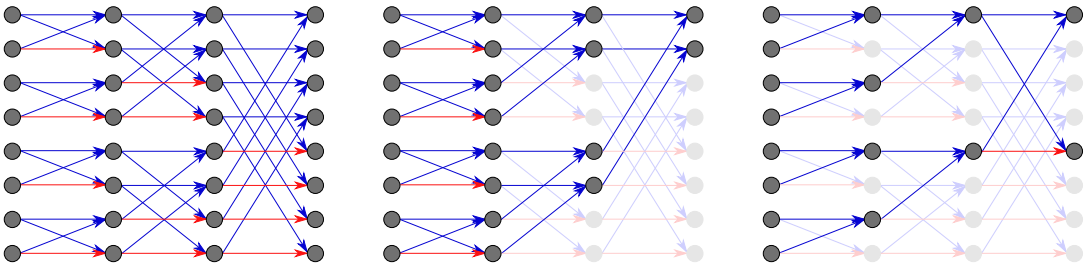
$$\widehat{K}_m^{\mathscr{H},(k)}(\mathbf{x},\mathbf{y}) = \frac{d}{m} \operatorname{Re} \left[\left\langle \overline{\mathbf{PSD}_k^{(\mathscr{U})} \left[\prod_{i=1}^{k-1} \mathbf{SD}_i^{(\mathscr{R})} \right] \mathbf{x}}, \mathbf{PSD}_k^{(\mathscr{U})} \left[\prod_{i=1}^{k-1} \mathbf{SD}_i^{(\mathscr{R})} \right] \mathbf{y} \right\rangle \right]. \tag{6.11}$$

Here, and in what follows, the bar notation denotes element-wise complex conjugation. Perhaps remarkably, this complex variant yields exactly half the MSE of the OJLT estimator.

Theorem 6.9. The estimator $\widehat{K}_m^{\mathscr{H},(k)}(\mathbf{x},\mathbf{y})$, applying uniform subsampling without replacement, is unbiased and satisfies: $\mathrm{MSE}(\widehat{K}_m^{\mathscr{H},(k)}(\mathbf{x},\mathbf{y})) = \frac{1}{2}\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y}))$.

This large factor of 2 improvement could instead be obtained by doubling m for $\widehat{K}_m^{(k)}$. However, this would require doubling the number of parameters for the transform, whereas the **S**-Hybrid estimator requires additional storage only for the complex parameters in the final diagonal matrix $\mathbf{D}_k^{(\mathscr{U})}$. Strikingly, it is straightforward to extend the proof of Theorem 6.9 (see Section 6.7) to show that rather than taking the complex random variables in $\mathbf{D}_k^{(\mathscr{U})}$ to be Unif(S^1), it is possible to take them to be Unif($\{1,-1,\mathrm{i},-\mathrm{i}\}$) and still obtain exactly the same benefit in MSE.

Theorem 6.10. For the estimator $\widehat{K}_m^{\mathscr{H},(k)}$ defined in Equation (6.11), replacing the random matrix $\mathbf{D}_k^{(\mathscr{U})}$ (which has i.i.d. Unif(S^1) elements on the diagonal) with a random diagonal matrix having i.i.d. Unif($\{1,-1,\mathrm{i},-\mathrm{i}\}$) elements on the diagonal does not affect the MSE of the estimator.



(a) Full FWHT transform.

(b) First 2 coordinates subsampled.

(c) Example of optimal selection of 2 coordinates with respect to computation.

Figure 6.1. Computational graph representing the Fast Walsh-Hadamard transform in 8 dimensions, and partial computation of the transform when only certain output coordinates are required. The value of each node is given by the sum of its inputs; a blue arrow represents the identity operation, whilst a red arrow represents multiplication by -1. Illustrations of this kind are common in the literature — see for example Aung et al. (2009).

6.3.6 Alternative subsampling strategies

Our results above focus on **SD**-product matrices where rows have been subsampled without replacement. As discussed in Definition 6.4, occasionally (e.g. for parallelisation) it may be of interest instead to subsample *with* replacement. As might be expected, this leads to worse MSE, which can be quantified precisely.

Theorem 6.11. For each of the estimators $\widehat{K}_m^{(k)}$ and $\widehat{K}_m^{\mathscr{H},(k)}$, if uniform subsampling *with* (rather than without) replacement is used, then the MSE is worsened by a multiplicative constant of $\frac{d-1}{d-m}$.

As mentioned in Definition 6.4, an alternative to sampling coordinates uniformly with and without replacement is to simply take, for example, the first m coordinates. Note whilst this approach is straightforward to implement, it results in a random projection that does not treat the coordinates of the input vectors symmetrically, which may lead to detrimental statistical performance.

In any case, with prior knowledge of which coordinates are to be sampled, it is possible to improve on the $\mathcal{O}(d \log d)$ cost of computing the matrix-vector product with the final copy of $\mathbf{H}$ in the embedding matrix. This is illustrated for the case of an 8-dimensional Fast Walsh-Hadamard transform in Figure 6.1. The first subfigure shows the operations involved in computing the full transform. The left-most column represents the

8-dimensional input, and the next three columns correspond to the three factors

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \tag{6.12}$$

that the 8-dimensional Hadamard matrix $\mathbf{H}$ may be expressed as a Kronecker product of. The second sub-figure shows the required computations if only the first two coordinates of the transformed vector are required. The third sub-figure shows that even fewer computations are required if only the first and fifth components of the output are of interest.

6.4 Understanding the effectiveness of orthogonality

Here we build intuitive understanding for the effectiveness of Hadamard–Rademacher matrices in particular, discussing a connection to random walks over orthogonal matrices.

We focus on Hadamard–Rademacher random matrices $\mathbf{HD}_k \ldots \mathbf{HD}_1$, a special case of the **SD**-product matrices described in Section 6.2. Our aim is to provide intuition for how the choice of k affects the quality of the random matrix, following our earlier observations just after Corollary 6.8, which indicated that for **SD**-product matrices, odd values of k yield greater benefits than even values, and that there are diminishing benefits from higher values of k. We proceed by casting the random matrices into the framework of Markov chains.

Definition 6.12 (**The Hadamard–Rademacher process**)**.** The Hadamard–Rademacher process in n dimensions is the Markov chain $(\mathbf{X}_k)_{k=0}^{\infty}$ taking values in the orthogonal group $O(n)$, with $\mathbf{X}_0 = \mathbf{I}$ almost surely, and $\mathbf{X}_k = \mathbf{HD}_k\mathbf{X}_{k-1}$ almost surely, where $\mathbf{H}$ is the normalised Hadamard matrix in n dimensions, and $(\mathbf{D}_k)_{k=1}^{\infty}$ are i.i.d. diagonal matrices with independent Rademacher random variables on their diagonals.

Constructing an estimator based on Hadamard–Rademacher matrices is equivalent to simulating several time steps from the Hadamard–Rademacher process. The quality of estimators based on Hadamard–

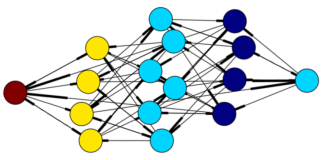


Figure 6.2. Visualisation of the Cayley graph explored by the Hadamard–Rademacher process in two dimensions. Nodes are coloured red, yellow, light blue, and dark blue for Cayley distances of 0, 1, 2, and 3 from the identity matrix, respectively.

Rademacher random matrices comes from a quick mixing property of the corresponding Markov chain. The following demonstrates attractive properties of the chain in low dimensions. The proof is entirely computational, involving enumerating all possible paths of length 4 that may be taken by the Markov chain through the state space, and is hence omitted.

Proposition 6.13. The Hadamard–Rademacher process, in the specific case of $d = 2$ dimensions, explores a state space of 16 orthogonal matrices, is ergodic with respect to the uniform distribution on this set, has period 2, the diameter of the Cayley graph associated with the random walk is 3, and the chain is fully mixed after 3 time steps.

This proposition, and the Cayley graph corresponding to the Markov chain's state space (Figure 6.2), illustrate the fast mixing properties of the Hadamard–Rademacher process in low dimensions; this agrees with the observations in Section 6.3 that there are diminishing returns associated with using a large number k of **HD** blocks in an estimator. The observation in Proposition 6.13 that the Markov chain has period 2 indicates that we should expect different behaviour for estimators based on odd and even numbers of blocks of **HD** matrices, which is reflected in the analytic expressions for MSE derived in Theorem 6.7 for the dimensionality reduction setup. We observe that in the expressions for MSE in Theorem 6.7, when $d > 2$, the expressions continue to change as k increases; this tells us that although the chain is completely mixed after 3 time steps when $d = 2$, this is not generally the case, although the exponential convergence (in k) of the MSE suggests that in general, the chain still enjoys fast mixing properties.

6.5 Experiments

We present comparisons of estimators introduced in Section 6.3, illustrating our theoretical results, and further demonstrating the empirical success of ROM-based estimators at the level of Gram matrix approximation. Throughout, we use the normalised Hadamard matrix **H** for the structured matrix **S**, and if not otherwise mentioned, we use uniform row sampling without replacement to select rows of Hadamard–Rademacher random matrices. We make four separate comparisons:

1. We compare the standard JLT based on the unstructured Gaussian random matrix with the OJLT based on the orthogonal Gaussian matrix, and based on the Hadamard–Rademacher random matrix with $k = 3$ blocks.

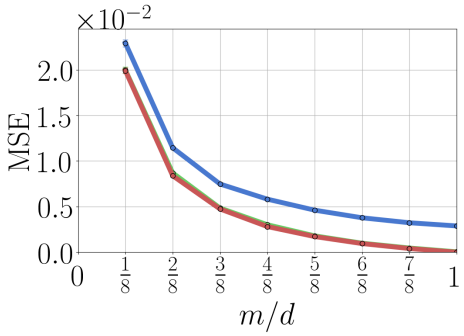
2. We compare the OJLT using Hadamard–Rademacher random matrices with varying numbers k of blocks.
3. We compare the OJLT using Hadamard–Rademacher random matrices with subsampling without replacement, subsampling with replacement, and subsampling the first m rows of the matrix.
4. We compare the OJLT using Hadamard–Rademacher random matrices with the hybrid complex variants introduced in Section 6.3.5.

We compare these methods on a variety of UCI datasets (Lichman, 2013); we present results on the insurance company dataset here, and several other datasets in the Appendix Section 6.B.

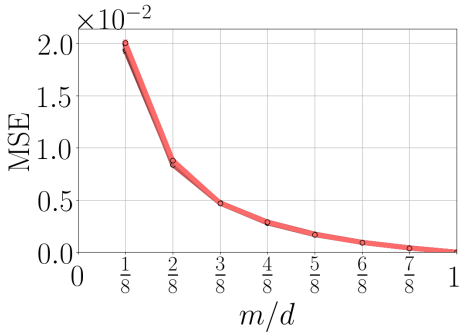
6.5.1 Pointwise inner product approximation

Complementing the theoretical results of Section 6.3, we provide several salient comparisons of the various methods described above, measured by MSE of pointwise inner product esimation — see Figure 6.3.

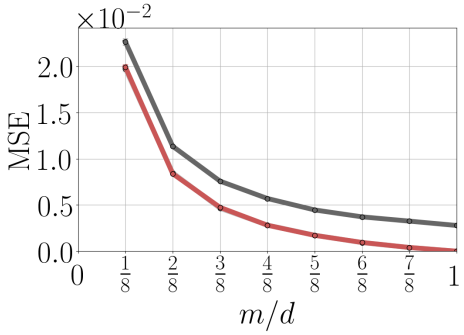
All empirical results support our earlier theoretical findings, and also serve to illustrate the behaviour of these methods that may be observed in practice. In particular, we note that Figure 6.3a demonstrates the superiority of transforms based on Gaussian orthogonal and Hadamard–Rademacher random matrices over those associated with unstructured Gaussian matrices, and also that Hadamard–Rademacher random matrices perform competitively with Gaussian orthogonal matrices. Figure 6.3b illustrates the MSE of transforms using Hadamard–Rademacher random matrices with varying numbers of blocks — performance is similar in all cases in this instance. The performance of various subsampling strategies are presented in Figure 6.3c, illustrating Theorem 6.11, and also demonstrating empirically (for this dataset) that a deterministic row sampling procedure (such as taking the first m rows) is competitive with uniform subsampling without replacement. Finally, Figure 6.3d illustrates the behaviour predicted by Theorems 6.9 and 6.10. We note that in this instance, sampling the first m rows leads to similar MSE performance as sampling uniformly without replacement.



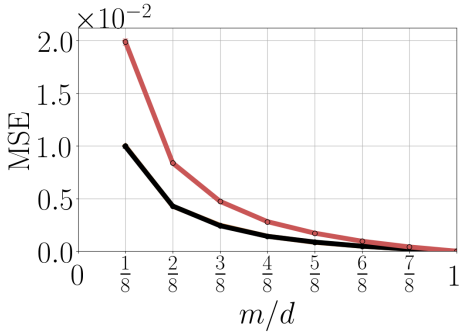
(a) Pointwise MSE using unstructured Gaussian (blue), orthogonal Gaussian (green), and Hadamard–Rademacher (red) random matrices.



(b) Pointwise MSE using Hadamard–Rademacher random matrices with $k = 1$ (dark) to $k = 4$ (light) blocks.



(c) Pointwise MSE using Hadamard–Rademacher random matrices with subsampling with replacement (dark grey), without replacement (red), and taking the first m rows (light grey).

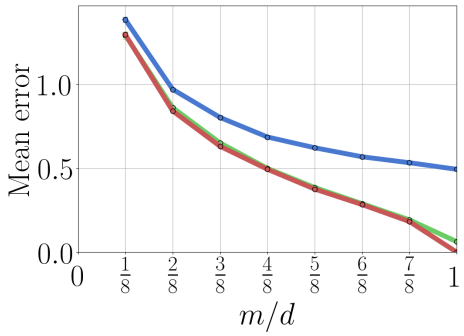


(d) Pointwise MSE using Hadamard–Rademacher random matrices (red), Hadamard-Hybrid with Unif(S^1) diagonals (gold), and Hadamard-Hybrid with Unif($\{1,-1,\mathrm{i},-\mathrm{i}\}$) diagonals (black).

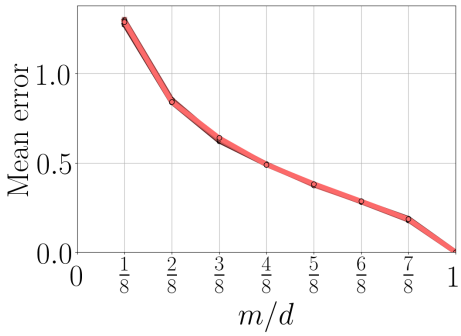
Figure 6.3. Comparisons of various random projection methods for pointwise inner product estimation on the insurance company UCI dataset.

6.5.2 Gram matrix approximation

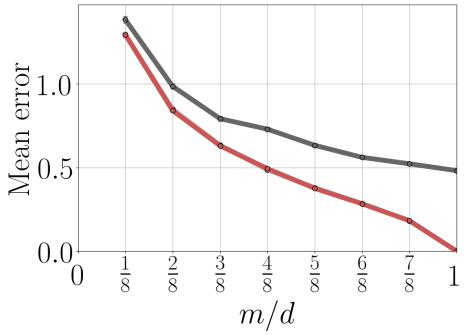
Moving beyond the theoretical guarantees established in Section 6.3, we show empirically that the statistical superiority of estimators based on ROMs is maintained at the level of Gram matrix approximation. We use the normalised Frobenius norm error ($\|\mathbf{K}-\widehat{\mathbf{K}}\|_2/\|\mathbf{K}\|_2$) as our metric for comparing an estimator $\widehat{\mathbf{K}}$ with the true matrix $\mathbf{K}$ (as used by Choromanski and Sindhwani, 2016), and plot the mean error for a range of random projection methods on full UCI datasets — see Figure 6.4 for results on the insurance company dataset, and Appendix Section 6.B for additional datasets.



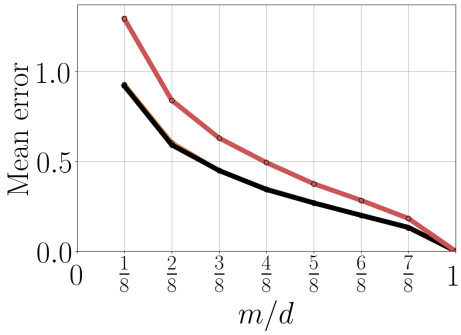
(a) Gram matrix estimation error using unstructured Gaussian (blue), orthogonal Gaussian (green), and Hadamard–Rademacher (red) random matrices.



(b) Gram matrix estimation error using Hadamard–Rademacher random matrices with $k = 1$ (dark) to $k = 4$ (light) blocks.



(c) Gram matrix estimation error using Hadamard–Rademacher random matrices with subsampling with replacement (dark grey), without replacement (red), and taking the first m rows (light grey).



(d) Gram matrix estimation error using Hadamard–Rademacher random matrices (red), Hadamard-Hybrid with Unif(S^1) diagonals (gold), and Hadamard-Hybrid with Unif($\{1,-1,\mathrm{i},-\mathrm{i}\}$) diagonals (black).

Figure 6.4. Comparisons of various random projection methods for Gram matrix estimation on the insurance company UCI dataset.

Broadly, the empirical results observed here for Gram matrix estimation echo the theoretical results obtained for pointwise estimation earlier. Figure 6.4a demonstrates that projections based on Gaussian orthogonal and Hadamard–Rademacher random matrices yield similar behaviour, with both offering superior performance relative to projections based on unstructured Gaussian matrices. We observe in Figure 6.4d that the Hadamard-Hybrid estimators using complex numbers consistently offer advantages relative to Hadamard–Rademacher random matrices, and switching from the uniform distribution Unif(S^1) to the discrete distribution Unif($\{1,-1,\mathrm{i},-\mathrm{i}\}$) has no effect on the mean Frobenius error. For comparison, we also list values of $\log N/d$ for each dataset concerned in Table 6.1,

which are proportional to the number of features required to accurately reconstruct inner products for each dataset, according the Johnson–Lindestrauss theory set out in Chapter 5. Thus these quantities give some context to the errors achieved in Figures 6.4 and 6.6. However, we point out several caveats in making these comparisons: (i) the number of features dictated by the Johnson–Lindenstrauss theory is linear in ε^{-2}, where ε is a prescribed accuracy level as described in Chapter 5; (ii) in our experiments, we evaluate normalised Frobenius norm error.

Dataset	Boston	Wine	Insurance Company	Parkinson	CPU
$\log N/d$	0.43	0.66	0.09	0.33	0.41

Table 6.1. Number of features dictated by Johnson–Lindenstrauss theory for the datasets used in Gram matrix reconstruction experiments, expressed as a proportion of the original dimensionality of the dataset, not including scaling by ε^{-2}, where ε is the prescribed accuracy level.

6.6 Discussion

The results in Section 6.3 establish quantitative results for the performance of Hadamard–Rademacher matrices, and more generally **SD**-product matrices, in linear dimensionality reduction. This represents a step towards understanding the strong empirical performance that has been observed in a variety of contexts when using Hadamard–Rademacher random matrices, as described in Section 6.1. The use of Hadamard–Rademacher random matrices for dimensionality reduction also yields an alternative interpretation of Hadamard–Rademacher blocks in existing random projection algorithms, such as the FJLT.

We highlight two natural questions for future work. Firstly, whether the Markov chain perspective described in Section 6.4 can cast further light on the empirical success of Hadamard–Rademacher matrices. Secondly, we emphasise that the statistical analysis performed in this chapter focuses on pointwise estimation of individual inner products between pairs of points. Ultimately, for many applications, we are often interested in estimating many such inner products in parallel. Our empirical results serve as preliminary indications of the kind of behaviour that may be observed, and developing a theoretical understanding how errors introduced by these structured random projections are correlated across given datasets is also an interesting question for further work.

Appendix 6.A Proofs

6.A.1 Proofs of results in Section 6.3.4

Lemma 6.5. The unstructured JLT inner product estimator $\widehat{K}_m^{\text{base}}(\mathbf{x},\mathbf{y})$ of the inner product $\langle \mathbf{x},\mathbf{y}\rangle$ for $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ using m-dimensional random projections is unbiased, with

$$\text{MSE}(\widehat{K}_m^{\text{base}}(\mathbf{x},\mathbf{y})) = \frac{1}{m}(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2)\,. \tag{6.8}$$

Proof. We begin by letting $R_i = \langle \mathbf{G}_i,\mathbf{x}\rangle\langle \mathbf{G}_i,\mathbf{y}\rangle$, where $\mathbf{g}^i$ stands for the i^{th} row of the unstructured Gaussian matrix $\mathbf{G}\in\mathbb{R}^{m\times d}$. With this notation, note that we have:

$$\widehat{K}_m^{\text{base}}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m} R_i\,. \tag{6.13}$$

Denote $\mathbf{G}_i = (g_{i1},...,g_{id})$. Notice that from the independence of the random variables $(g_{ij}|i=1,\ldots,m,\ \ j=1,\ldots,d)$ and the moments $\mathbb{E}[g_{ij}]=0$, $\mathbb{E}[g_{ij}^2]=1$ for all i,j, we obtain $\mathbb{E}[R_i]=\sum_{j=1}^d x_j y_j = \langle \mathbf{x},\mathbf{y}\rangle$. This establishes the unbiasedness of the estimator. Since the estimator is unbiased, we have $\text{MSE}(\widehat{K}_m^{\text{base}}(\mathbf{x},\mathbf{y})) = \text{Var}(\widehat{K}_m^{\text{base}}(\mathbf{x},\mathbf{y}))$. Since the random variables $(R_i)_{i=1}^m$ are independent and identically distributed, from Expression (6.13) it follows that we have the following expression for its MSE:

$$\text{MSE}(\widehat{K}_m^{\text{base}}(\mathbf{x},\mathbf{y})) = \frac{1}{m^2}\sum_{i=1}^{m}(\mathbb{E}[R_i^2]-(\mathbb{E}[R_i])^2) = \frac{1}{m}(\mathbb{E}[R_1^2]-\mathbb{E}[R_1]^2)\,. \tag{6.14}$$

From the unbiasedness of the estimator, we have $\mathbb{E}[R_1]=\langle \mathbf{x},\mathbf{y}\rangle$. Therefore we obtain:

$$\text{MSE}(\widehat{K}_m^{\text{base}}(\mathbf{x},\mathbf{y})) = \frac{1}{m}\left(\mathbb{E}[R_1^2]-\langle \mathbf{x},\mathbf{y}\rangle^2\right). \tag{6.15}$$

Now notice that

$$\mathbb{E}[R_1^2] = \mathbb{E}\left[\sum_{i,j,k,l=1}^{d} g_{1i}g_{1j}g_{1k}g_{1l}x_i y_j x_k y_l\right] = \sum_{i,j,k,l=1}^{d} x_i y_j x_k y_l \mathbb{E}\left[g_{1i}g_{1j}g_{1k}g_{1l}\right], \tag{6.16}$$

In the expression above the only non-zero terms correspond to quadruples (i,j,k,l), where no index appears an odd number of times. Therefore, from the inclusion-exclusion

principle and the fact that $\mathbb{E}[g_{1i}^2] = 1$ and $\mathbb{E}[g_{1i}^4] = 3$, we obtain

$$\begin{aligned}
\mathbb{E}[R_1^2] &= \sum_{\substack{i=j,k=l \\ i\neq k}} x_i y_j x_k y_l \mathbb{E}[g_{1i}g_{1j}g_{1k}g_{1l}] + \sum_{\substack{i=k,j=l \\ i\neq j}} x_i y_j x_k y_l \mathbb{E}[g_{1i}g_{1j}g_{1k}g_{1l}] \\
&\quad + \sum_{\substack{i=l,j=k \\ i\neq k}} x_i y_j x_k y_l \mathbb{E}[g_{1i}g_{1j}g_{1k}g_{1l}] + \sum_{i=j=k=l} x_i y_j x_k y_l \mathbb{E}[g_{1i}g_{1j}g_{1k}g_{1l}] \\
&= \sum_{\substack{i=j,k=l \\ i\neq k}} x_i y_j x_k y_l + \sum_{\substack{i=k,j=l \\ i\neq j}} x_i y_j x_k y_l \\
&\quad + \sum_{\substack{i=l,j=k \\ i\neq k}} x_i y_j x_k y_l + 3 \sum_{i=j=k=l} x_i y_j x_k y_l \\
&= \sum_{i,j=1}^{d} x_i y_i x_j y_j + \sum_{i,j=1}^{d} x_i^2 y_j^2 + \sum_{i,j=1}^{d} x_i y_i x_j y_j \\
&= (\|\mathbf{x}\|_2 \|\mathbf{y}\|_2)^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle^2 . \qquad (6.17)
\end{aligned}$$

Therefore, finally we have

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x},\mathbf{y})) = \frac{1}{m}\left((\|\mathbf{x}\|_2\|\mathbf{y}\|_2)^2 + 2\langle \mathbf{x},\mathbf{y}\rangle^2 - \langle \mathbf{x},\mathbf{y}\rangle^2\right) = \frac{1}{m}(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x},\mathbf{y}\rangle^2), \qquad (6.18)$$

which completes the proof. □

Theorem 6.6. The estimator $\widehat{K}_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y})$ of the inner product $\langle \mathbf{x},\mathbf{y}\rangle$ for $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ is unbiased and satisfies[2], for $d \geq 4$:

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y})) = \mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x},\mathbf{y})) - \frac{m-1}{m(d-1)}\left(\langle \mathbf{x},\mathbf{y}\rangle^2 + \frac{d}{d+2}\|\mathbf{x}\|^2\|\mathbf{y}\|^2\right). \qquad (6.9)$$

Proof. The unbiasedness of the Gaussian orthogonal estimator comes from the fact that every row of the Gaussian orthogonal matrix is marginally equal in distribution to a row of an unstructured Gaussian random matrix; unbiasedness therefore follows from Lemma 6.5. Now note that

$$\mathrm{Cov}(R_i, R_j) = \mathbb{E}[R_i R_j] - \mathbb{E}[R_i]\mathbb{E}[R_j], \qquad (6.19)$$

where $R_i = \langle \mathbf{r}_i,\mathbf{x}\rangle\langle \mathbf{r}_i,\mathbf{y}\rangle$, $R_j = \langle \mathbf{r}_j,\mathbf{x}\rangle\langle \mathbf{r}_j,\mathbf{y}\rangle$ and $\mathbf{r}_i,\mathbf{r}_j$ stand for the i^{th} and j^{th} row of the Gaussian orthogonal matrix respectively. From the fact that Gaussian orthogonal estimator

[2]The author thanks Sergio Bacallado for providing the simplified expression for the MSE given here.

is unbiased, we get

$$\mathbb{E}[R_i] = \langle \mathbf{x}, \mathbf{y} \rangle . \tag{6.20}$$

Let us now compute $\mathbb{E}[R_i R_j]$[3]. We begin by letting $\mathbf{S}$ be a random matrix drawn from Haar measure on the orthogonal group $O(d)$. Letting ξ_i, ξ_j be independent random variables, with $\xi_i^2, \xi_j^2 \sim \chi^2_d$, we have

$$(\mathbf{r}_i, \mathbf{r}_j) \overset{\mathcal{D}}{=} \begin{cases} (\xi_i \mathbf{S}_i, \xi_j \mathbf{S}_j) & \text{if } i \neq j \\ (\xi_i \mathbf{S}_i, \xi_i \mathbf{S}_j) & \text{if } i = j , \end{cases} \tag{6.21}$$

where $\mathbf{S}_i, \mathbf{S}_j$ denote the i$^{\text{th}}$ and j$^{\text{th}}$ rows of $\mathbf{S}$, and $\overset{\mathcal{D}}{=}$ denotes equality in distribution. Thus, computing moments of the relevant χ^2 random variables, we obtain

$$\mathbb{E}\left[R_i R_j\right] = \begin{cases} d^2 \mathbb{E}\left[\mathbf{x}^\top \mathbf{S}_i \mathbf{S}_i^\top \mathbf{y} \mathbf{x}^\top \mathbf{S}_j \mathbf{S}_j^\top \mathbf{y}\right] & \text{if } i \neq j \\ d(d+2)\mathbb{E}\left[\mathbf{x}^\top \mathbf{S}_i \mathbf{S}_i^\top \mathbf{y} \mathbf{x}^\top \mathbf{S}_j \mathbf{S}_j^\top \mathbf{y}\right] & \text{if } i = j . \end{cases} \tag{6.22}$$

Now, summing the expression $\mathbb{E}\left[\mathbf{x}^\top \mathbf{S}_i \mathbf{S}_i^\top \mathbf{y} \mathbf{x}^\top \mathbf{S}_j \mathbf{S}_j^\top \mathbf{y}\right]$ over the indices $i, j \in \{1, \ldots, d\}$, and using almost-sure orthonormality of the set $\{\mathbf{S}_1, \ldots, \mathbf{S}_d\}$, we obtain

$$\sum_{i,j=1}^{d} \mathbb{E}\left[\mathbf{x}^\top \mathbf{S}_i \mathbf{S}_i^\top \mathbf{y} \mathbf{x}^\top \mathbf{S}_j \mathbf{S}_j^\top \mathbf{y}\right] = \mathbb{E}\left[\mathbf{x}^\top \mathbf{S}^\top \mathbf{S} \mathbf{y} \mathbf{x}^\top \mathbf{S}^\top \mathbf{S} \mathbf{y}\right] = \langle \mathbf{x}, \mathbf{y} \rangle^2 . \tag{6.23}$$

As derived in Lemma 6.5, we have

$$\mathbb{E}\left[\mathbf{x}^\top \mathbf{S}_i \mathbf{S}_i^\top \mathbf{y} \mathbf{x}^\top \mathbf{S}_i \mathbf{S}_i^\top \mathbf{y}\right] = \frac{1}{d(d+2)} \mathbb{E}\left[R_i^2\right] = \frac{1}{d(d+2)} \left(\|\mathbf{x}\|^2 \|\mathbf{y}\|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle^2 \right) . \tag{6.24}$$

By symmetry, we can thus calculate $\mathbb{E}\left[R_i R_j\right]$ in the case $i \neq j$:

$$\begin{aligned} \mathbb{E}\left[R_i R_j\right] &= \frac{d^2}{d(d-1)} \left(\langle \mathbf{x}, \mathbf{y} \rangle^2 - \frac{d}{d(d+2)} \left(\|\mathbf{x}\|^2 \|\mathbf{y}\|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle^2 \right) \right) \\ &= \frac{d^2}{(d+2)(d-1)} \langle \mathbf{x}, \mathbf{y} \rangle^2 - \frac{d}{(d-1)(d+2)} \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 . \end{aligned} \tag{6.25}$$

[3] The author thanks Sergio Bacallado for providing the following straightforward argument for calculating this expectation; a previous version of this proof relied on calculations in hyperspherical coordinates, as in the proof of Theorem 7.5.

Plugging this result into the formula for the MSE of the estimator concerned, we obtain

$$\begin{aligned}
&\mathrm{MSE}(K_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y})) = \mathrm{MSE}(K_m^{\mathrm{base}}(\mathbf{x},\mathbf{y})) + \frac{m-1}{m}\left(\mathbb{E}\left[R_1R_2\right] - \mathbb{E}[R_1]^2\right)\\
=&\mathrm{MSE}(K_m^{\mathrm{base}}(\mathbf{x},\mathbf{y})) + \frac{(m-1)}{m}\left(\frac{d^2}{(d+2)(d-1)}\langle\mathbf{x},\mathbf{y}\rangle^2 - \frac{d}{(d-1)(d+2)}\|\mathbf{x}\|^2\|\mathbf{y}\|^2 - \langle\mathbf{x},\mathbf{y}\rangle^2\right)\\
=&\mathrm{MSE}(K_m^{\mathrm{base}}(\mathbf{x},\mathbf{y})) + \frac{(m-1)}{m(d-1)}\left(\langle\mathbf{x},\mathbf{y}\rangle^2 + \frac{d}{d+2}\|\mathbf{x}\|^2\|\mathbf{y}\|^2\right), \qquad (6.26)
\end{aligned}$$

as required for the statement of the theorem. □

Theorem 6.7. The OJLT estimator $\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})$ of the inner product $\langle\mathbf{x},\mathbf{y}\rangle$ for $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, with k blocks, using m-dimensional random projections and uniform subsampling policy without replacement, is unbiased with

$$\begin{aligned}
\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})) = \frac{1}{m}\left(\frac{d-m}{d-1}\right)\Big((\langle\mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \qquad (6.10)\\
\sum_{r=1}^{k-1}\frac{(-1)^r 2^r}{d^r}(2\langle\mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \frac{(-1)^k 2^k}{d^{k-1}}\sum_{i=1}^{d}x_i^2 y_i^2\Big).
\end{aligned}$$

We obtain Theorem 6.7 through a sequence of smaller propositions. Broadly, the strategy is first to show that the estimators of Theorem 6.7 are unbiased (Proposition 6.14). An expression for the mean squared error of the estimator $\widehat{K}_m^{(1)}$ with one matrix block is then derived (Proposition 6.15). Finally, a straightforward recursive formula for the mean squared error of the general estimator is derived (Proposition 6.16), and the result of the theorem then follows.

Proposition 6.14. The estimator $\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})$ is unbiased, for all $k,d\in\mathbb{N}$, $1\le m\le d$, and $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$.

Proof. Notice first that since rows of $\mathbf{S}$ are orthogonal and are L_2-normalised, the matrix $\mathbf{S}$ is an isometry. Thus each block $\mathbf{SD}_i^{(\mathscr{R})}$ is also an isometry. Therefore it suffices to prove the claim for $k=1$.

Then, denoting by $\mathbf{J}=(J_1,\ldots,J_m)$ the indices of the randomly selected rows of $\mathbf{SD}_1^{(\mathscr{R})}$, note that the estimator $\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})$ may be expressed in the form

$$\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m}\left(\sqrt{d}(\mathbf{SD}_1^{(\mathscr{R})})_{J_i}\mathbf{x}\times\sqrt{d}(\mathbf{SD}_1^{(\mathscr{R})})_{J_i}\mathbf{y}\right), \qquad (6.27)$$

where $(\mathbf{SD}_1^{(\mathscr{R})})_i$ is the i^{th} row of $\mathbf{SD}_1^{(\mathscr{R})}$. Since each of the rows of $\mathbf{SD}_1^{(\mathscr{R})}$ has the same marginal distribution, it suffices to demonstrate that $\mathbb{E}[\mathbf{y}^T\mathbf{D}_1^{(\mathscr{R})}\mathbf{S}_1^\top\mathbf{S}_1\mathbf{D}_1^{(\mathscr{R})}\mathbf{x}] = \frac{\mathbf{x}^\top\mathbf{y}}{d}$, where $\mathbf{S}_1$ is the first row of $\mathbf{S}$. Now simply note that

$$\mathbb{E}[\mathbf{y}^\top\mathbf{D}\mathbf{S}_1^\top\mathbf{S}_1\mathbf{D}\mathbf{x}] = \frac{1}{d}\mathbb{E}\left[\sum_{i=1}^d y_i\delta_i \times \sum_{i=1}^d x_i\delta_i\right] = \frac{1}{d}\mathbb{E}\left[\sum_{i=1}^d x_i y_i\delta_i^2\right] + \mathbb{E}\left[\sum_{i\neq j} x_i y_j\delta_i\delta_j\right] = \frac{\mathbf{x}^\top\mathbf{y}}{d}, \tag{6.28}$$

where $(\delta_i)_{i=1}^d$ are the i.i.d. Rademacher random variables on the diagonal of $\mathbf{D}_1^{(\mathscr{R})}$. □

With Proposition 6.14 in place, the mean square error for the estimator $\widehat{K}_m^{(1)}$ using one matrix block can be derived.

Proposition 6.15. The MSE of the single $\mathbf{SD}^{(\mathscr{R})}$-block m-feature estimator $\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})$ for $\langle\mathbf{x},\mathbf{y}\rangle$ using the *without replacement* row subsampling strategy is

$$\text{MSE}(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})) = \frac{1}{m}\left(\frac{d-m}{d-1}\right)\left(\|\mathbf{x}\|^2\|\mathbf{y}\|^2 + \langle\mathbf{x},\mathbf{y}\rangle^2 - 2\sum_{i=1}^d x_i^2 y_i^2\right). \tag{6.29}$$

Proof. First note that since $\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})$ is unbiased, the mean squared error is simply the variance of this estimator. Secondly, denoting the indices of the m randomly selected rows by $\mathbf{J} = (J_1,\ldots,J_m)$, by conditioning on $\mathbf{J}$ we obtain the following:

$$\begin{aligned}&\text{Var}\left(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})\right) = \\ &\frac{d^2}{m^2}\left(\mathbb{E}\left[\text{Var}\left(\sum_{p=1}^m (\mathbf{SDx})_{J_p}(\mathbf{SDy})_{J_p}\middle|\mathbf{J}\right)\right] + \text{Var}\left(\mathbb{E}\left[\sum_{p=1}^m (\mathbf{SDx})_{J_p}(\mathbf{SDy})_{J_p}\middle|\mathbf{J}\right]\right)\right),\end{aligned} \tag{6.30}$$

where we have dropped the indices on the random diagonal matrix $\mathbf{D}$ to keep notation uncluttered.

Now note that the conditional expectation in the second term is constant as a function of J, since conditional on whichever rows are sampled, the resulting estimator is unbiased. Taking the variance of this constant therefore causes the second term to vanish. Now

consider the conditional variance that appears in the first term:

$$\begin{aligned}\operatorname{Var}\left(\sum_{p=1}^{m}(\mathbf{SDx})_{J_p}(\mathbf{SDy})_{J_p}\middle|\mathbf{J}\right)&=\sum_{p=1}^{m}\sum_{p'=1}^{m}\operatorname{Cov}\left((\mathbf{SDx})_{J_p}(\mathbf{SDy})_{J_p},(\mathbf{SDx})_{J_{p'}}(\mathbf{SDy})_{J_{p'}}\middle|\mathbf{J}\right)\\&=\sum_{p,p'=1}^{m}\sum_{i,j,k,l=1}^{d}s_{J_pi}s_{J_pj}s_{J_{p'}k}s_{J_{p'}l}x_iy_jx_ky_l\operatorname{Cov}\left(\delta_i\delta_j,\delta_k\delta_l\right),\end{aligned}\tag{6.31}$$

where we write $\mathbf{D}=\operatorname{Diag}(\delta_1,\ldots,\delta_d)$. Now note that $\operatorname{Cov}\left(\delta_i\delta_j,\delta_k\delta_l\right)$ is non-zero iff i,j are distinct, and $\{i,j\}=\{k,l\}$, in which case the covariance is 1. We therefore obtain:

$$\begin{aligned}&\operatorname{Var}\left(\sum_{p=1}^{m}(\mathbf{SDx})_{J_p}(\mathbf{SDy})_{J_p}\middle|\mathbf{J}\right)=\\&\sum_{p,p'=1}^{m}\sum_{i\neq j}^{d}\left(s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}x_i^2y_j^2+s_{J_pi}s_{J_pj}s_{J_{p'}j}s_{J_{p'}i}x_iy_jx_jy_i\right).\end{aligned}\tag{6.32}$$

Substituting this expression for the conditional variance into the decomposition of the MSE of the estimator, we obtain the result of the theorem:

$$\begin{aligned}\operatorname{Var}\left(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})\right)&=\frac{d^2}{m^2}\mathbb{E}\left[\sum_{p,p'=1}^{m}\sum_{i\neq j}^{d}\left(s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}x_i^2y_j^2+s_{J_pi}s_{J_pj}s_{J_{p'}j}s_{J_{p'}i}x_iy_jx_jy_i\right)\right]\\&=\frac{d^2}{m^2}\sum_{p,p'=1}^{m}\sum_{i\neq j}^{d}\left(x_i^2y_j^2+x_ix_jy_iy_j\right)\mathbb{E}\left[s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}\right].\end{aligned}\tag{6.33}$$

We now consider the law on the index variables $\mathbf{J}=(J_1,\ldots,J_m)$ induced by the subsampling strategy without replacement to evaluate the expectation in this last term. If $p=p'$, the integrand of the expectation is deterministically $1/d^2$. If $p\neq p'$, then we obtain:

$$\begin{aligned}\mathbb{E}\left[s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}\right]&=\mathbb{E}\left[s_{J_pi}s_{J_pj}\mathbb{E}\left[s_{J_{p'}i}s_{J_{p'}j}\middle|J_p\right]\right]\\&=\mathbb{E}\Big[s_{J_pi}s_{J_pj}\Big[\left(\frac{1}{d}\left(\frac{d/2-1}{d-1}\right)-\frac{1}{d}\left(\frac{d/2}{d-1}\right)\right)\mathbb{1}_{\{s_{J_pi}s_{J_pj}=1/d\}}+\\&\qquad\left(\frac{1}{d}\left(\frac{d/2}{d-1}\right)-\frac{1}{d}\left(\frac{d/2-1}{d-1}\right)\right)\mathbb{1}_{\{s_{J_pi}s_{J_pj}=-1/d\}}\Big]\Big]\\&=\frac{1}{d(d-1)}\mathbb{E}\left[s_{J_pi}s_{J_pj}\left(\mathbb{1}_{\{s_{J_pi}s_{J_pj}=-1/d\}}-\mathbb{1}_{\{s_{J_pi}s_{J_pj}=1/d\}}\right)\right]\\&=\frac{-1}{d^2(d-1)},\end{aligned}\tag{6.34}$$

where we have used the fact that the products $s_{J_p i} s_{J_p j}$ and $s_{J_{p'} i} s_{J_{p'} j}$ take values in $\{\pm 1/d\}$, and because distinct rows of $\mathbf{S}$ are orthogonal, the marginal probability of each of the two values is $1/2$. A simple adjustment, using almost-sure distinctness of J_p and $J_{p'}$, yields the conditional probabilities needed to evaluate the conditional expectation that appears in the calculation above.

Substituting the values of these expectations back into the expression for the variance of $\widehat{K}_m^{(1)}(\mathbf{x}, \mathbf{y})$ then yields

$$\begin{aligned}
\mathrm{Var}(\widehat{K}_m^{(1)}(\mathbf{x}, \mathbf{y})) &= \frac{d^2}{m^2} \sum_{i \neq j}^{d} \left(x_i^2 y_j^2 + x_i x_j y_i y_j \right) \left(m \times \frac{1}{d^2} - m(m-1) \times \frac{1}{d^2(d-1)} \right) \\
&= \frac{1}{m} \left(1 - \frac{m-1}{d-1} \right) \sum_{i \neq j}^{d} \left(x_i^2 y_j^2 + x_i x_j y_i y_j \right) \\
&= \frac{1}{m} \left(1 - \frac{m-1}{d-1} \right) \left(\sum_{i,j=1}^{d} (x_i^2 y_j^2 + x_i x_j y_i y_j) - 2 \sum_{i=1}^{d} x_i^2 y_i^2 \right) \\
&= \frac{1}{m} \left(\frac{d-m}{d-1} \right) \left(\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 - 2 \sum_{i=1}^{d} x_i^2 y_i^2 \right), \qquad (6.35)
\end{aligned}$$

as required. □

We now turn our attention to the following recursive expression for the mean squared error of a general estimator.

Proposition 6.16. Let $k \geq 2$. We have the following recursion for the MSE of $K_m^{(k)}(x, y)$:

$$\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})) = \mathbb{E}\left[\mathrm{MSE}\left(\widehat{K}_m^{(k-1)}(\mathbf{SD}_1 \mathbf{x}, \mathbf{SD}_1 \mathbf{y}) | \mathbf{D}_1 \right) \right]. \qquad (6.36)$$

Proof. The result follows from a straightforward application of the law of total variance, conditioning on the matrix $\mathbf{D}_1$. Observe that

$$\begin{aligned}
\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})) &= \mathrm{Var}(\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})) \qquad (6.37) \\
&= \mathbb{E}\left[\mathrm{Var}\left(\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y}) \middle| \mathbf{D}_1 \right) \right] + \mathrm{Var}\left(\mathbb{E}\left[\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y}) \middle| \mathbf{D}_1 \right] \right) \\
&= \mathbb{E}\left[\mathrm{Var}\left(\widehat{K}_m^{(k-1)}(\mathbf{SD}_1 \mathbf{x}, \mathbf{SD}_1 \mathbf{y}) \middle| \mathbf{D}_1 \right) \right] + \mathrm{Var}\left(\mathbb{E}\left[\widehat{K}_m^{(k-1)}(\mathbf{SD}_1 \mathbf{x}, \mathbf{SD}_1 \mathbf{y}) \middle| \mathbf{D}_1 \right] \right).
\end{aligned}$$

But examining the conditional expectation in the second term, we observe

$$\mathbb{E}\left[\widehat{K}_m^{(k-1)}(\mathbf{SD}_1 \mathbf{x}, \mathbf{SD}_1 \mathbf{y}) \middle| \mathbf{D}_1 \right] = \langle \mathbf{SD}_1 \mathbf{x}, \mathbf{SD}_1 \mathbf{y} \rangle \quad \text{almost surely}, \qquad (6.38)$$

by unbiasedness of the estimator, and since $\mathbf{SD}_1$ is orthogonal almost surely, this is equal to the (constant) inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ almost surely. This conditional expectation therefore has 0 variance, and so the second term in the expression for the MSE above vanishes, which results in the statement of the proposition. □

With these intermediate propositions established, we are now in a position to prove Theorem 6.7. In order to use the recursive result of Proposition 6.16, we require the following lemma.

Lemma 6.17. For all $x, y, \in \mathbb{R}^d$, we have

$$\mathbb{E}\left[\sum_{i=1}^{d} (\mathbf{SDx})_i^2 (\mathbf{SDy})_i^2\right] = \frac{1}{d}\left(\|\mathbf{x}\|^2\|\mathbf{y}\|^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle^2 - 2\sum_{i=1}^{d} x_i^2 y_i^2\right). \tag{6.39}$$

Proof. The result follows by direct calculation. Note that

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{d} (\mathbf{SDx})_i^2 (\mathbf{SDy})_i^2\right] &= d\mathbb{E}\left[\left(\sum_{a=1}^{d} s_{1a}\delta_a x_a\right)^2 \left(\sum_{a=1}^{d} s_{1a}\delta_a y_a\right)^2\right] \\ &= d\sum_{i,j,k,l=1}^{d} s_{1i}s_{1j}s_{1k}s_{1l}x_i x_j y_k y_l \mathbb{E}\left[\delta_i\delta_j\delta_k\delta_l\right], \end{aligned} \tag{6.40}$$

where the first inequality follows since the d summands indexed by i in the initial expectation are identically distributed. Now note that the expectation $\mathbb{E}\left[\delta_i\delta_j\delta_k\delta_l\right]$ is non-zero iff $i = j = k = l$, or $i = j \neq k = l$, or $i = k \neq j = l$, or $i = l \neq k = l$; in all such cases, the expectation takes the value 1. Substituting this into the above expression and collecting terms, we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{d} (\mathbf{SDx})_i^2 (\mathbf{SDy})_i^2\right] &= \frac{1}{d}\left(\sum_{i=1}^{d} x_i^2 y_i^2 + \sum_{i\neq j} x_i^2 y_i^2 + 2\sum_{i\neq j} x_i x_j y_i y_j\right) \\ &= \frac{1}{d}\left(\sum_{i,j=1}^{d} x_i^2 y_j^2 + 2\sum_{i,j=1}^{d} x_i x_j y_i y_j - 2\sum_{i=1}^{d} x_i^2 y_i^2\right), \end{aligned} \tag{6.41}$$

from which the statement of the lemma follows immediately. □

With these preliminary results established, we may now give the proof of Theorem 6.7.

Proof of Theorem 6.7. Recall that we aim to establish the following general expression for $k \geq 1$:

$$\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})) = \tag{6.42}$$
$$\frac{1}{m}\left(\frac{d-m}{d-1}\right)\left((\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \sum_{r=1}^{k-1}\frac{(-1)^r 2^r}{d^r}(2\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \frac{(-1)^k 2^k}{d^{k-1}}\sum_{i=1}^{d} x_i^2 y_i^2\right).$$

We proceed by induction. The case $k = 1$ is verified by Proposition 6.15. For the inductive step, suppose the result holds for some $k \in \mathbb{N}$. Then observe by Proposition 6.16 and the inductive hypothesis, we have

$$\begin{aligned}\mathrm{MSE}(\widehat{K}_m^{(k+1)}(\mathbf{x},\mathbf{y})) &= \mathbb{E}\left[\mathrm{MSE}\left(\widehat{K}_m^{(k)}(\mathbf{SD}_1\mathbf{x},\mathbf{SD}_1\mathbf{y})|\mathbf{D}_1\right)\right] \\ &= \frac{1}{m}\left(\frac{d-m}{d-1}\right)\left((\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \sum_{r=1}^{k-1}\frac{(-1)^r 2^r}{d^r}(2\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2)\right. \\ &\qquad \left. + \frac{(-1)^k 2^k}{d^{k-1}}\sum_{i=1}^{d}\mathbb{E}\left[(\mathbf{SD}_1\mathbf{x})_i^2(\mathbf{SD}_1\mathbf{y})_i^2\right]\right), \end{aligned} \tag{6.43}$$

where we have used that $\mathbf{SD}_1$ is almost surely orthogonal, and therefore $\|\mathbf{SD}_1\mathbf{x}\|^2 = \|\mathbf{x}\|^2$ almost surely, $\|\mathbf{SD}_1\mathbf{y}\|^2 = \|\mathbf{y}\|^2$ almost surely, and $\langle \mathbf{SD}_1\mathbf{x}, \mathbf{SD}_1\mathbf{y}\rangle = \langle \mathbf{x},\mathbf{y}\rangle$ almost surely. Applying Lemma 6.17 to the remaining expectation and collecting terms yields the required expression for $\mathrm{MSE}(\widehat{K}_m^{(k+1)}(\mathbf{x},\mathbf{y}))$, and the proof is complete. □

6.A.2 Proof of Theorem 6.9

The proof of Theorem 6.9 follows a very similar structure to that of Theorem 6.7; we proceed by induction, and may use the results of Proposition 6.16 to set up a recursion. We first show unbiasedness of the estimator (Proposition 6.18), and then treat the base case of the inductive argument (Proposition 6.19). We prove slightly more general statements than needed for Theorem 6.9.

Proposition 6.18. The estimator $K_m^{\mathscr{H},(k)}(\mathbf{x},\mathbf{y})$ is unbiased for all $k, d \in \mathbb{N}$, $m \leq d$, and $\mathbf{x},\mathbf{y} \in \mathbb{C}^d$ with $\langle \overline{\mathbf{x}},\mathbf{y}\rangle \in \mathbb{R}$; in particular, for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$.

Proof. Following a similar argument to the proof of Proposition 6.14, note that it is sufficient to prove the claim for $k = 1$, since each **SD** block is unitary, and hence preserves the Hermitian product $\langle \overline{\mathbf{x}},\mathbf{y}\rangle$.

Next, note that the estimator can be written as a sum of identically distributed terms:

$$\widehat{K}_m^{\mathscr{H},(1)}(\mathbf{x},\mathbf{y}) = \frac{d}{m}\sum_{i=1}^{m} \mathrm{Re}\left((\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_i} \times (\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_i}\right). \tag{6.44}$$

The terms are identically distributed since the index variables J_i are marginally identically distributed, and the rows of $\mathbf{S}\mathbf{D}_1$ are marginally identically distributed (the elements of a row are i.i.d. Unif$(S^1)/\sqrt{d}$). Now note

$$\begin{aligned}\mathbb{E}\left[\mathrm{Re}\left((\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_i} \times (\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_i}\right)\right] &= \frac{1}{d}\mathbb{E}\left[\sum_{i=1}^{d} y_i\delta_i \times \sum_{j=1}^{d}\overline{x}_j\overline{\delta}_j\right] \\ &= \frac{1}{d}\mathbb{E}\left[\sum_{i=1}^{d}\overline{x}_i y_i \delta_i\overline{\delta}_i\right] + \frac{1}{d}\mathbb{E}\left[\sum_{i\neq j}\overline{x}_i y_j\overline{\delta}_i\delta_j\right] = \frac{1}{d}\langle\overline{\mathbf{x}},\mathbf{y}\rangle,\end{aligned} \tag{6.45}$$

where $\delta_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(S^1)$ for $i = 1,\ldots,d$. This immediately yields $\mathbb{E}\left[\widehat{K}_m^{\mathscr{H},(1)}(\mathbf{x},\mathbf{y})\right] = \langle\overline{\mathbf{x}},\mathbf{y}\rangle$, as required. □

We now derive the base case for our inductive proof, again proving a slightly more general statement than necessary for Theorem 6.9.

Proposition 6.19. Let $\mathbf{x},\mathbf{y} \in \mathbb{C}^d$ such that $\langle\overline{\mathbf{x}},\mathbf{y}\rangle \in \mathbb{R}$. The MSE of the single complex **SD**-block m-feature estimator $K_m^{\mathscr{H},(1)}(\mathbf{x},\mathbf{y})$ for $\langle\overline{\mathbf{x}},\mathbf{y}\rangle$ is

$$\mathrm{MSE}(\widehat{K}_m^{\mathscr{H},(1)}(\mathbf{x},\mathbf{y})) = \frac{1}{2m}\left(\frac{d-m}{d-1}\right)\left(\langle\overline{\mathbf{x}},\mathbf{x}\rangle\langle\overline{\mathbf{y}},\mathbf{y}\rangle + \langle\overline{\mathbf{x}},\mathbf{y}\rangle^2 - \sum_{i=1}^{d}|x_i|^2|y_i|^2 - \sum_{i=1}^{d}\mathrm{Re}(\overline{x}_i^2 y_i^2)\right). \tag{6.46}$$

Proof. The proof is very similar to that of Proposition 6.15. By the unbiasedness result of Proposition 6.18, the mean squared error of the estimator is simply the variance. We begin by conditioning on the random index vector $\mathbf{J}$ selected by the subsampling procedure.

$$\widehat{K}_m^{\mathscr{H},(1)}(\mathbf{x},\mathbf{y})) = \frac{1}{m}\mathrm{Re}\left(\langle\sqrt{d}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{\mathbf{J}}, \sqrt{d}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{\mathbf{J}}\rangle\right), \tag{6.47}$$

where again $\mathbf{J}$ is a vector of indices sampled uniformly from $\{1,\ldots,d\}$ without replacement, and the bar over $\mathbf{D}$ represents complex conjugation. Since the estimator is again unbiased, its MSE is equal to its variance. First conditioning on the index set $\mathbf{J}$, as for Proposition

6.15, we obtain

$$\begin{aligned}&\mathrm{Var}\left(\widehat{K}_m^{\mathscr{H},(1)}(\mathbf{x},\mathbf{y})\right) \qquad (6.48)\\ &=\frac{d^2}{m^2}\left(\mathbb{E}\left[\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right)\right]+\mathrm{Var}\left(\mathbb{E}\left[\mathrm{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right]\right)\right).\end{aligned}$$

Again, the second term vanishes as the conditional expectation is constant as a function of $\mathbf{J}$; the summands have equal expectation for fixed $\mathbf{J}$, as shown in the proof of unbiasedness. Turning attention to the conditional variance expression in the first term, we note

$$\begin{aligned}&\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right)=\\ &\sum_{p,p'=1}^{m}\sum_{i,j,k,l=1}^{d}s_{J_pi}s_{J_pj}s_{J_{p'}k}s_{J_{p'}l}\mathrm{Cov}\left(\mathrm{Re}(\overline{\delta}_i\overline{x}_i\delta_j y_j),\mathrm{Re}(\overline{\delta}_k\overline{x}_k\delta_l y_l)\right). \qquad (6.49)\end{aligned}$$

Now note that the covariance term is non-zero iff i,j are distinct, and $\{i,j\}=\{k,l\}$. We therefore obtain

$$\begin{aligned}&\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right)=\sum_{p,p'=1}^{m}\sum_{i\neq j}^{d}s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}\times\\ &\left(\mathrm{Cov}\left(\mathrm{Re}(\overline{\delta}_i\overline{x}_i\delta_j y_j),\mathrm{Re}(\overline{\delta}_i\overline{x}_i\delta_j y_j)\right)+\mathrm{Cov}\left(\mathrm{Re}(\overline{\delta}_i\overline{x}_i\delta_j y_j),\mathrm{Re}(\overline{\delta}_j\overline{x}_j\delta_i y_i)\right)\right). \qquad (6.50)\end{aligned}$$

First consider the term $\mathrm{Cov}\left(\mathrm{Re}(\overline{\delta}_i\overline{x}_i\delta_j y_j),\mathrm{Re}(\overline{\delta}_i\overline{x}_i\delta_j y_j)\right)$. The random variable $\overline{\delta}_i\overline{x}_i\delta_j y_j$ is distributed uniformly on the circle in the complex plane centred at the origin with radius $|\overline{x}_i y_j|$. Therefore the variance of its real part is

$$\mathrm{Cov}\left(\mathrm{Re}(\overline{\delta}_i\overline{x}_i\delta_j y_j),\mathrm{Re}(\overline{\delta}_i\overline{x}_i\delta_j y_j)\right)=\frac{1}{2}|\overline{x}_i y_j|^2=\frac{1}{2}x_i\overline{x}_i y_j\overline{y}_j. \qquad (6.51)$$

For the second covariance term, we perform an explicit calculation. Let $Z=e^{\mathrm{i}\theta}=\overline{\delta}_i\delta_j$. Then we have

$$\begin{aligned}&\mathrm{Cov}\left(\mathrm{Re}(\overline{\delta}_i\overline{x}_i\delta_j y_j),\mathrm{Re}(\overline{\delta}_j\overline{x}_j\delta_i y_i)\right)=\mathrm{Cov}\left(\mathrm{Re}(Z\overline{x}_i y_j),\mathrm{Re}(\overline{Z}\overline{x}_j y_i)\right)\\ &=\mathrm{Cov}\left(\cos(\theta)\mathrm{Re}(\overline{x}_i y_j)-\sin(\theta)\mathrm{Im}(\overline{x}_i y_j),\cos(\theta)\mathrm{Re}(\overline{x}_j y_i)+\sin(\theta)\mathrm{Im}(\overline{x}_j y_i)\right)\\ &=\frac{1}{2}\left(\mathrm{Re}(\overline{x}_i y_j)\mathrm{Re}(\overline{x}_j y_i)-\mathrm{Im}(\overline{x}_i y_j)\mathrm{Im}(\overline{x}_j y_i)\right), \qquad (6.52)\end{aligned}$$

with the final equality following since the angle θ is uniformly distributed on $[0,2\pi]$, and from standard trigonometric integral identities. We recognise the bracketed terms in the

final line as the real part of the product $\overline{x}_i\overline{x}_j y_i y_j$. Substituting these into the expression for the conditional variance obtained above, we have

$$\operatorname{Var}\left(\operatorname{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}\mathbf{x})_{J_p}(\mathbf{S}\mathbf{D}\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right) = \sum_{p,p'=1}^{m}\sum_{i\neq j}^{d} s_{J_p i}s_{J_p j}s_{J_{p'} i}s_{J_{p'} j}\frac{1}{2}\left(x_i\overline{x}_i y_j\overline{y}_j + \operatorname{Re}(\overline{x}_i\overline{x}_j y_i y_j)\right). \tag{6.53}$$

Now taking the expectation over the index variables $\mathbf{J}$, we note that as in the proof of Proposition 6.15, the expectation of the term $s_{J_p i}s_{J_p j}s_{J_{p'} i}s_{J_{p'} j}$ is $1/d^2$ when $p = p'$, and $-1/(d^2(d-1))$ otherwise. Therefore we obtain

$$\begin{aligned}
\operatorname{Var}\left(\widehat{K}_m^{\mathscr{H},(1)}(\mathbf{x},\mathbf{y})\right) &= \frac{d^2}{m^2}\left(\left(\frac{m}{d^2} - \frac{m(m-1)}{d^2(d-1)}\right)\frac{1}{2}\sum_{i\neq j}^{d}\left(x_i\overline{x}_i y_j\overline{y}_j + \operatorname{Re}(\overline{x}_i\overline{x}_j y_i y_j)\right)\right)\\
&= \frac{1}{2m}\left(\frac{d-m}{d-1}\right)\left(\sum_{i\neq j}^{d}\left(x_i\overline{x}_i y_j\overline{y}_j + \operatorname{Re}(\overline{x}_i\overline{x}_j y_i y_j)\right)\right)\\
&= \frac{1}{2m}\left(\frac{d-m}{d-1}\right)\left(\sum_{i,j=1}^{d}\left(x_i\overline{x}_i y_j\overline{y}_j + \operatorname{Re}(\overline{x}_i\overline{x}_j y_i y_j)\right) - \sum_{i=1}^{d}(x_i\overline{x}_i y_i\overline{y}_i + \operatorname{Re}(\overline{x}_i\overline{x}_i y_i y_i))\right)\\
&= \frac{1}{2m}\left(\frac{d-m}{d-1}\right)\left(\langle\overline{\mathbf{x}},\mathbf{x}\rangle\langle\overline{\mathbf{y}},\mathbf{y}\rangle + \langle\overline{\mathbf{x}},\mathbf{y}\rangle^2 - \sum_{i=1}^{d}(x_i\overline{x}_i y_i\overline{y}_i + \operatorname{Re}(\overline{x}_i\overline{x}_i y_i y_i))\right),
\end{aligned} \tag{6.54}$$

where in the final equality we have used the assumption that $\langle\overline{\mathbf{x}},\mathbf{y}\rangle \in \mathbb{R}$. □

We are now in a position to prove Theorem 6.9 by induction, using Proposition 6.19 as a base case, and Proposition 6.16 for the inductive step.

Proof of Theorem 6.9. Recall that we aim to establish the following general expression for $k \geq 1$, for $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$:

$$\begin{aligned}
\operatorname{MSE}(\widehat{K}_m^{\mathscr{H},(k)}(\mathbf{x},\mathbf{y})) = \frac{1}{2m}\left(\frac{d-m}{d-1}\right)\Big(&((\mathbf{x}^\top\mathbf{y})^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \\
&\sum_{r=1}^{k-1}\frac{(-1)^r 2^r}{d^r}(2(\mathbf{x}^\top\mathbf{y})^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \frac{(-1)^k 2^k}{n^{k-1}}\sum_{i=1}^{d}x_i^2 y_i^2\Big).
\end{aligned} \tag{6.55}$$

We proceed by induction. The case $k = 1$ is verified by Proposition 6.19, and by noting that in the expression obtained in Proposition 6.19, we have

$$\sum_{i=1}^{d} x_i\overline{x}_i y_i\overline{y}_i = \operatorname{Re}(\overline{x}_i\overline{x}_i y_i y_i) = \sum_{i=1}^{d} x_i^2 y_i^2. \tag{6.56}$$

For the inductive step, suppose the result holds for some $k \in \mathbb{N}$. Then observe by Proposition 6.16 and the induction hypothesis, we have, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\begin{aligned}\mathrm{MSE}(\widehat{K}_m^{\mathscr{H},(k+1)}(\mathbf{x},\mathbf{y})) &= \mathbb{E}\left[\mathrm{MSE}\left(\widehat{K}_m^{\mathscr{H},(k)}(\mathbf{SD}_1\mathbf{x},\mathbf{SD}_1\mathbf{y})|\mathbf{D}_1\right)\right] \\ &= \frac{1}{2m}\left(\frac{d-m}{d-1}\right)\Big((\langle\mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \sum_{r=1}^{k-1}\frac{(-1)^r 2^r}{d^r}(2\langle\mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) \\ &\qquad + \frac{(-1)^k 2^k}{d^{k-1}}\sum_{i=1}^{d}\mathbb{E}\left[(\mathbf{SD}_1\mathbf{x})_i^2(\mathbf{SD}_1\mathbf{y})_i^2\right]\Big), \end{aligned} \tag{6.57}$$

where we have used that $\mathbf{SD}_1$ is almost surely orthogonal, and therefore $\|\mathbf{SD}_1\mathbf{x}\|^2 = \|\mathbf{x}\|^2$ almost surely, $\|\mathbf{SD}_1\mathbf{y}\|^2 = \|\mathbf{y}\|^2$ almost surely, and $\langle\mathbf{SD_1x},\mathbf{SD}_1\mathbf{y}\rangle = \langle\mathbf{x},\mathbf{y}\rangle$ almost surely. Applying Lemma 6.17 to the remaining expectation and collecting terms yields the required expression for $\mathrm{MSE}(\widehat{K}_m^{\mathscr{H},(k+1)}(\mathbf{x},\mathbf{y}))$, and the proof is complete. □

6.A.3 Proof of Corollary 6.10

The proof follows simply by following the inductive strategy of the proof of Theorem 6.9, replacing the base case in Proposition 6.19 with the following.

Proposition 6.20. Let $\mathbf{x},\mathbf{y} \in \mathbb{C}^d$ such that $\langle\overline{\mathbf{x}},\mathbf{y}\rangle \in \mathbb{R}$. The MSE of the single complex **SD**-block m-feature estimator $K_m^{\mathscr{H},(1)}(\mathbf{x},\mathbf{y})$ using a diagonal matrix with entries $\mathrm{Unif}(\{1,-1,\mathrm{i},-\mathrm{i}\})$, rather than $\mathrm{Unif}(S^1)$ for $\langle\mathbf{x},\mathbf{y}\rangle$ is

$$\mathrm{MSE}(\widehat{K}_m^{\mathscr{H},(1)}(\mathbf{x},\mathbf{y})) = \frac{1}{2m}\left(\frac{d-m}{d-1}\right)\left(\langle\overline{\mathbf{x}},\mathbf{x}\rangle\langle\overline{\mathbf{y}},\mathbf{y}\rangle + \langle\overline{\mathbf{x}},\mathbf{y}\rangle^2 - \sum_{i=1}^{d}|x_i|^2|y_i|^2 - \sum_{i=1}^{d}\mathrm{Re}(\overline{x}_i^2 y_i^2)\right). \tag{6.58}$$

Proof. The proof of this proposition proceeds exactly as for Proposition 6.19; by following the same chain of reasoning, conditioning on the index set $\mathbf{J}$ of the subsampled rows, we arrive at

$$\begin{aligned}&\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{SD}_1\mathbf{y})_{J_p}\right)\Big|\mathbf{J}\right) = \\ &\sum_{p,p'=1}^{m}\sum_{i,j,k,l=1}^{n} s_{J_p i}s_{J_p j}s_{J_{p'}k}s_{J_{p'}l}\mathrm{Cov}\left(\mathrm{Re}(\overline{d}_i\overline{x}_i d_j y_j), \mathrm{Re}(\overline{d}_k\overline{x}_k d_l y_l)\right).\end{aligned} \tag{6.59}$$

Since we are dealing strictly with the case $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we may simplify this further to obtain

$$\operatorname{Var}\left(\operatorname{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right) = \sum_{p,p'=1}^{m}\sum_{i,j,k,l=1}^{d} s_{J_p i}s_{J_p j}s_{J_{p'}k}s_{J_{p'}l}x_i x_k y_i y_l \operatorname{Cov}\left(\operatorname{Re}(\overline{\delta}_i\delta_j), \operatorname{Re}(\overline{\delta}_k\delta_l)\right). \tag{6.60}$$

By calculating directly with the $\delta_i, \delta_j, \delta_k, \delta_l \sim \operatorname{Unif}(\{1, -1, \mathrm{i}, -\mathrm{i}\})$, we obtain

$$\operatorname{Var}\left(\operatorname{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right) = \sum_{p,p'=1}^{m}\sum_{i\neq j}^{d} s_{J_p i}s_{J_p j}s_{J_{p'}i}s_{J_{p'}j}\frac{1}{2}\left(x_i\overline{x}_i y_j\overline{y}_j + \operatorname{Re}(\overline{x}_i\overline{x}_j y_i y_j)\right), \tag{6.61}$$

exactly as in Proposition 6.19; following the rest of the argument of Proposition 6.19 yields the result. □

The proof of the corollary now follows by applying the steps of the proof of Theorem 6.9.

6.A.4 Proof of Theorem 6.11

Theorem 6.11. For each of the estimators $\widehat{K}_m^{(k)}$ and $\widehat{K}_m^{\mathscr{H},(k)}$, if uniform subsampling *with* (rather than without) replacement is used, then the MSE is worsened by a multiplicative constant of $\frac{d-1}{d-m}$.

Proof. The proof of this result is reasonably straightforward with the proofs of Theorems 6.7 and 6.9 in hand; we simply recognise where in these proofs the assumption of the sampling strategy without replacement was used. We deal first with Theorem 6.7, which deals with the MSE associated with $\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})$. The only place in which the assumption of the subsampling strategy without replacement is used is mid-way through the proof of Proposition 6.15, which quantifies $\operatorname{MSE}(\widehat{K}_m^{(1)}(\mathbf{x}, \mathbf{y}))$. Picking up the proof at the point the subsampling strategy is used, we have

$$\operatorname{MSE}(\widehat{K}_m^{(1)}(\mathbf{x}, \mathbf{y})) = \frac{d^2}{m^2}\sum_{p,p'=1}^{m}\sum_{i\neq j}^{d}\left(x_i^2 y_j^2 + x_i x_j y_i y_j\right)\mathbb{E}\left[s_{J_p i}s_{J_p j}s_{J_{p'}i}s_{J_{p'}j}\right]. \tag{6.62}$$

Now instead using subsampling strategy with replacement, note that each pair of subsampled indices J_p and $J_{p'}$ are independent. Recalling that the columns of $\mathbf{S}$ are orthogonal,

we obtain for distinct p and p' that

$$\mathbb{E}\left[s_{J_p i} s_{J_p j} s_{J_{p'} i} s_{J_{p'} j}\right] = \mathbb{E}\left[s_{J_p i} s_{J_p j}\right] \mathbb{E}\left[s_{J_{p'} i} s_{J_{p'} j}\right] = 0\,. \tag{6.63}$$

Again, for $p = p'$, we have $\mathbb{E}\left[s_{J_p i} s_{J_p j} s_{J_{p'} i} s_{J_{p'} j}\right] = 1/d^2$. Substituting the values of these expectations back into the expression for the MSE of $\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})$ then yields

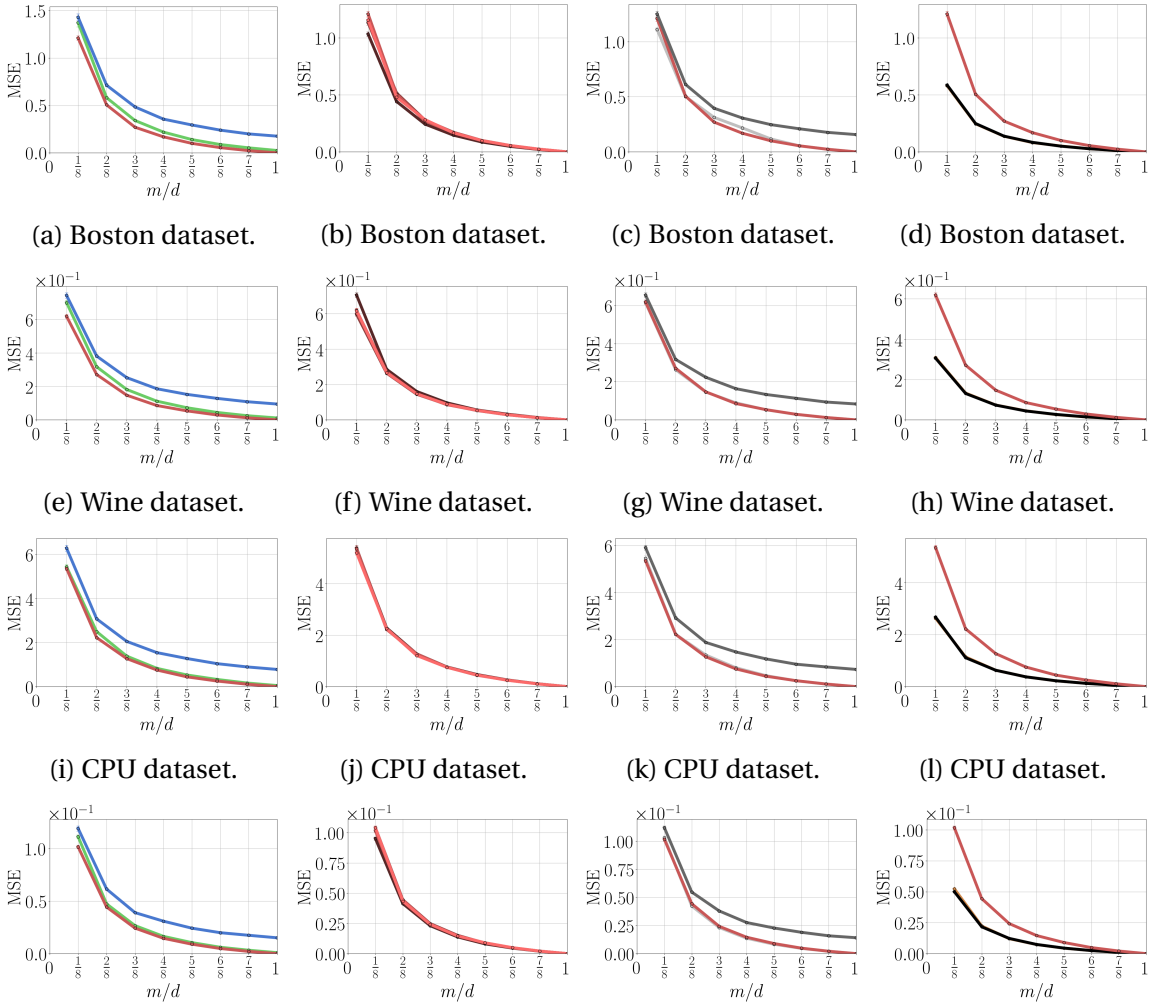
$$\begin{aligned}\mathrm{MSE}(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})) &= \frac{d^2}{m^2} \sum_{i \neq j}^{d} \left(x_i^2 y_j^2 + x_i x_j y_i y_j\right) \left(m \times \frac{1}{d^2}\right) \\ &= \frac{1}{m} \left(\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 - 2 \sum_{i=1}^{n} x_i^2 y_i^2\right), \end{aligned} \tag{6.64}$$

as required for $\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})$. The conclusion for $\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})$ now follows immediately by applying the inductive result of Proposition 6.16.

For the estimator $\widehat{K}_m^{\mathscr{H},(k)}(\mathbf{x},\mathbf{y})$, the result also immediately follows with the above calculation, as the only point in the proof of the MSE expressions for these estimators that is influenced by the subsampling strategy is in the calculation of the quantities $\mathbb{E}\left[s_{J_p i} s_{J_p j} s_{J_{p'} i} s_{J_{p'} j}\right]$; therefore, exactly the same multiplicative factor is incurred for MSE as for $\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})$. □

Appendix 6.B Additional experimental results

Here, we report additional experimental results on a range of UCI datasets, with the same experimental set up for pointwise and Gram matrix estimation as for the insurance company dataset described in Section 6.5. We give additional experimental results for pointwise estimation in Figure 6.5, and for Gram matrix estimation in Figure 6.6. Broadly, the qualitative behaviour observed in these results is as observed for the insurance company dataset, although we note that the quality of projections based on orthogonal Gaussian matrices decreases relative to those based on Hadamard–Rademacher matrices for lower-dimensional datasets.



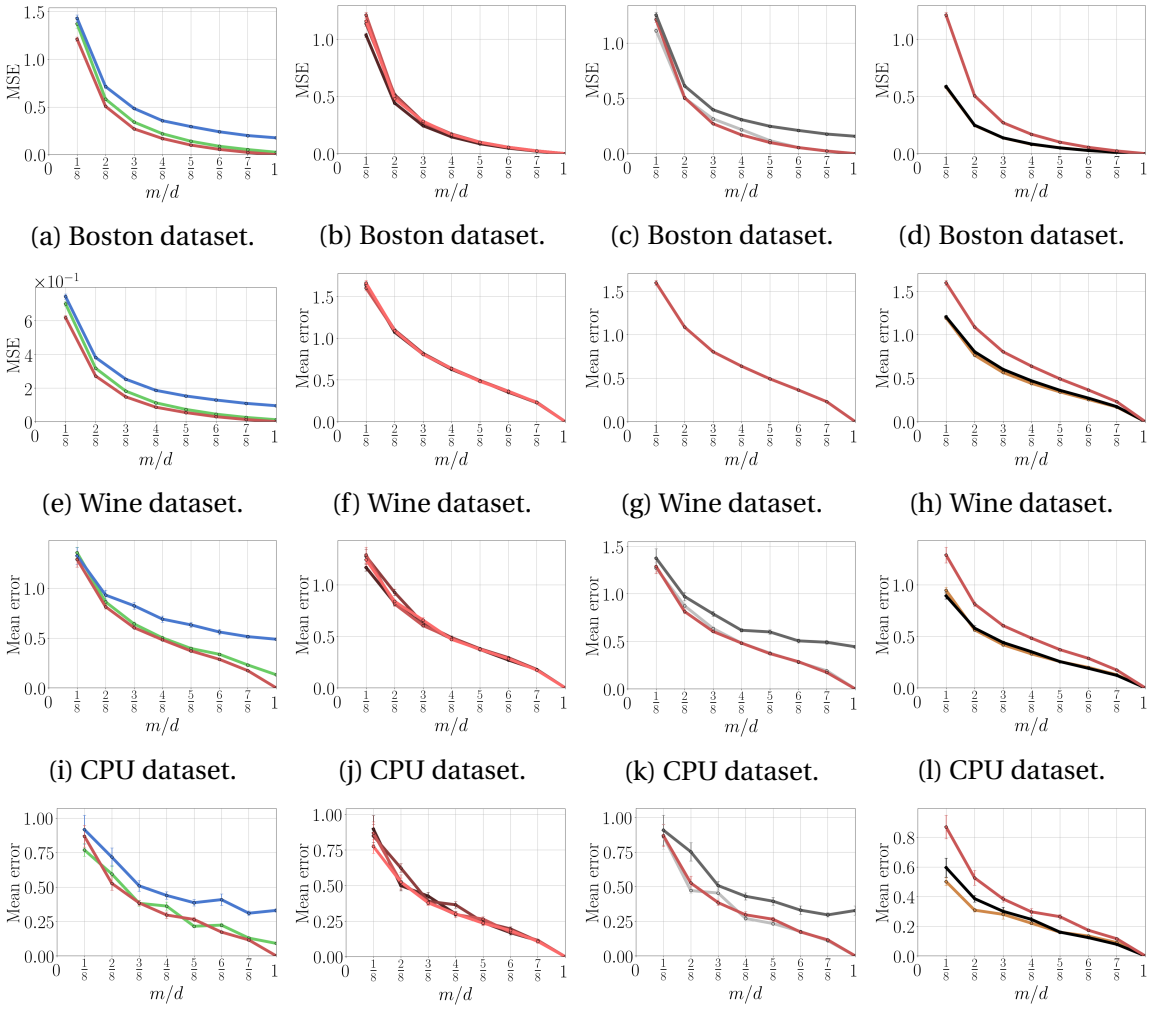
(a) Boston dataset. (b) Boston dataset. (c) Boston dataset. (d) Boston dataset.

(e) Wine dataset. (f) Wine dataset. (g) Wine dataset. (h) Wine dataset.

(i) CPU dataset. (j) CPU dataset. (k) CPU dataset. (l) CPU dataset.

(m) Parkinson dataset. (n) Parkinson dataset. (o) Parkinson dataset. (p) Parkinson dataset.

Figure 6.5. MSE for pointwise inner product reconstruction for a variety of UCI datasets. Two randomly selected datapoints from each dataset are chosen, and the kernel evaluated at these points is estimated. Plots in left-most column compare pointwise reconstruction MSE using unstructured Gaussian (blue), orthogonal Gaussian (green), and Hadamard–Rademacher (red) random matrices. Plots in centre-left column compare pointwise reconstruction MSE using Hadamard–Rademacher random matrices with $k = 1$ (dark) to $k = 4$ (light) blocks. Plots in centre-right column compare pointwise reconstruction MSE using Hadamard–Rademacher random matrices with subsampling with replacement (dark grey), without replacement (red), and taking the first m rows (light grey). Plots in right-most column compare pointwise reconstruction MSE using Hadamard–Rademacher random matrices (red), Hadamard-Hybrid with Unif(S^1) diagonals (gold), and Hadamard-Hybrid with Unif($\{1, -1, \mathrm{i}, -\mathrm{i}\}$) diagonals (black).



(a) Boston dataset. (b) Boston dataset. (c) Boston dataset. (d) Boston dataset.

(e) Wine dataset. (f) Wine dataset. (g) Wine dataset. (h) Wine dataset.

(i) CPU dataset. (j) CPU dataset. (k) CPU dataset. (l) CPU dataset.

(m) Parkinson dataset. (n) Parkinson dataset. (o) Parkinson dataset. (p) Parkinson dataset.

Figure 6.6. Gram matrix reconstruction error for a variety of UCI datasets. Plots in left-most column compare Gram matrix reconstruction error for projections using unstructured Gaussian (blue), orthogonal Gaussian (green), and Hadamard–Rademacher (red) random matrices. Plots in centre-left column compare Gram matrix reconstruction error for projections using Hadamard–Rademacher random matrices with $k = 1$ (dark) to $k = 4$ (light) blocks. Plots in centre-right column compare Gram matrix reconstruction error for projections using Hadamard–Rademacher random matrices with subsampling with replacement (dark grey), without replacement (red), and taking the first m rows (light grey). Plots in right-most column compare Gram matrix reconstruction error for projections using Hadamard–Rademacher random matrices (red), Hadamard-Hybrid with Unif(S^1) diagonals (gold), and Hadamard-Hybrid with Unif($\{1, -1, \mathrm{i}, -\mathrm{i}\}$)) diagonals (black).

Chapter 7

Variance Reduction for Random Features via Orthogonality

This chapter is based on material from the following publications:

- Choromanski, K.*, Rowland, M.*, and Weller, A. (2017). The unreasonable effectiveness of structured random orthogonal embeddings. In *Neural Information Processing Systems (NIPS)*. [*=equal contribution].
- Choromanski, K.*, Rowland, M.*, Sarlos, T., Sindhwani, V., Turner, R. E., and Weller, A. (2018a). The geometry of random features. In *Artificial Intelligence and Statistics* (AISTATS). [*=equal contribution].

More precisely, this chapter contains the results on the angular kernel from Choromanski et al. (2017), and the results on statistical accuracy of orthogonal random Fourier features from Choromanski et al. (2018a). The theoretical results in this chapter are primarily due to the author of this thesis and Krzysztof Choromanski. All experiments were designed, implemented, and run by the author of the thesis. The writing of these two papers was primarily a joint effort between the author of this thesis, Krzysztof Choromanski, and Adrian Weller, although some sections have been rewritten for this thesis.

7.1 Introduction

In this chapter, we consider random feature approximations for several classes of kernels (as described in Section 5.1), and study methods for improving the statistical and compu-

tational properties of these approximations, based on introducing orthogonal couplings between samples from the relevant Fourier distributions; the specifics of this construction are described in in Section 7.2. As described in Section 5.3, we will be interested in the effect of introducing orthogonal couplings on: (i) sampling costs; (ii) random feature map storage costs; (iii) cost of computing random features; and (iv) statistical accuracy of the kernel approximation.

The approach of using orthogonally-coupled features was first explored specifically for the *Gaussian kernel* by Yu et al. (2016). The authors observed that these orthogonal couplings, in contrast to other methods for scalable random feature maps, often demonstrated statistical improvements in the estimation of Gaussian kernel values, rather than just computational benefits. Thus, in some sense the orthogonal coupling acts as a variance reduction method. The authors gave a theoretical result guaranteeing statistical superiority of orthogonally-coupled features over i.i.d. features in asymptotically high dimensions, and demonstrated impressive empirical performance of these feature maps in kernel estimation, and also in support vector classification tasks. The authors also empirically investigated fast approximations to true orthogonally-coupled features, using Hadamard–Rademacher random matrices, as described in Section 6.2. Several questions were left open by the authors, such as for which other kernels orthogonally-coupled features can be used, and whether orthogonally-coupled features always offer statistical advantages in kernel estimation. We address both these questions in this chapter, contribute a variety of further theoretical results on the use of orthogonally coupled random features, and remark on further aspects of the theory that remain open in the discussion in Section 7.7.

We highlight the following contributions:

- In Section 7.3: Orthogonal random feature maps for the angular kernel, with corresponding theoretical guarantees of statistical superiority (measured by MSE) relative to i.i.d. random features.
- In Section 7.5: Analysis of orthogonal random Fourier features for general stationary isotropic kernels, that extends that of Yu et al. (2016) for the Gaussian kernel.
- In Section 7.6: Experimental evaluation of orthogonal random features for the angular kernel and a variety of stationary kernels.

7.2 Orthogonal random features for the Gaussian kernel

Before moving on to our original work, we first give a brief account of orthogonal random features for the Gaussian kernel, as proposed by Yu et al. (2016). The authors consider an isotropic Gaussian kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|^2\right)$. Rahimi and Recht (2007) showed that this kernel is amenable to random Fourier feature approximation, since it is stationary (see Example 5.2 for further details). In this particular case, the Fourier identity is given by

$$K(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} \exp\left(\mathrm{i}\langle \mathbf{x}-\mathbf{y}, \mathbf{w}\rangle\right) \eta(\mathrm{d}\mathbf{w}), \tag{7.1}$$

where $\eta \in \mathscr{P}(\mathbb{R}^d)$ is the $\mathcal{N}(\mathbf{0}, \sigma^{-2} I)$ distribution. To simplify notation, we take $\sigma^2 = 1$ in what follows. The random feature map that arises from Expression (7.1) thus has the form

$$\Phi_m(\mathbf{x}) = \left(\left(\frac{1}{\sqrt{m}}\cos(\langle \mathbf{w}_i, \mathbf{x}\rangle)\right)_{i=1}^m, \left(\frac{1}{\sqrt{m}}\sin(\langle \mathbf{w}_i, \mathbf{x}\rangle)\right)_{i=1}^m\right), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^d. \tag{7.2}$$

Alternatively, we may view the random feature map Φ_m as being the composition of the random matrix multiplication $\mathbf{x} \mapsto \mathbf{G}\mathbf{x}$, with application of cosine and sine non-linearities to each coordinate of the resulting vector, where $\mathbf{G} \in \mathbb{R}^{m \times d}$ has i^{th} row given by $\mathbf{w}_i$, for $i = 1, \ldots, m$. The notation $\mathbf{G}$ is chosen intentionally, as it is precisely a random matrix with i.i.d. Gaussian entries, as encountered in Sections 5.2 and 6.2. Yu et al. (2016) propose to replace $\mathbf{G}$ with a random matrix where each row marginally has the same distribution as in $\mathbf{G}$, but so that rows are orthogonal almost surely. This is precisely the Gaussian orthogonal matrix $\mathbf{G}_{\text{ort}}$ described in Definition 6.1. When $m > d$, as is often the case in random feature applications, it is not possible for all rows to be mutually orthogonal; the definition of the Gaussian orthogonal matrix is naturally extended to this case in the following manner:

Definition 7.1 (**Gaussian orthogonal matrices for $m > d$**)**.** Let $m > d$. We define the random Gaussian orthogonal matrix $\mathbf{G}_{\text{ort}}$ taking values in $\mathbb{R}^{m \times d}$ via the following construction. Let $l = \lceil m/d \rceil > 1$. The distribution of $\mathbf{G}_{\text{ort}}$ is determined by vertically stacking l independent $d \times d$ Gaussian orthogonal matrices (as defined in Definition 6.1), and removing rows from the bottom of the resulting stacked matrix if necessary to obtain an $m \times d$ matrix.

Thus, in general the random matrix $\mathbf{G}_{\text{ort}}$ taking values in $\mathbb{R}^{m \times d}$ is such that each consecutive block of rows with indices $id, \ldots, (i+1)d - 1$ has orthogonal rows almost surely, and these blocks of rows are independent from one another. The orthogonal random feature

map of Yu et al. (2016) may therefore be written

$$\Phi_m(\mathbf{x}) = \left(\left(\frac{1}{\sqrt{m}} \cos(\langle \mathbf{w}_i^{\text{ort}}, \mathbf{x} \rangle) \right)_{i=1}^{m}, \left(\frac{1}{\sqrt{m}} \sin(\langle \mathbf{w}_i^{\text{ort}}, \mathbf{x} \rangle) \right)_{i=1}^{m} \right), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^d. \quad (7.3)$$

where $\mathbf{w}_i^{\text{ort}}$ is the i^{th} row of a $\mathbf{G}_{\text{ort}}$ matrix taking values in $\mathbb{R}^{m \times d}$, for $i = 1, \ldots, m$.

The primary theoretical result given regarding the statistical performance of this estimator is given below, as stated in Yu et al. (2016).

Theorem 7.2 (Yu et al. (2016))**.** Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be an isotropic Gaussian kernel, so that $K(\mathbf{x}, \mathbf{y}) = \exp(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{y}\|^2)$, for some $\sigma^2 > 0$. Let $m \leq d$, and let $\widehat{K}_m^{\text{base}}(\mathbf{x}, \mathbf{y})$ and $\widehat{K}_m^{\text{ort}}(\mathbf{x}, \mathbf{y})$ be the Gaussian kernel estimators obtained using m i.i.d. random features and m orthogonal random features, respectively. Then, for large d, writing $z = \|\mathbf{x} - \mathbf{y}\| / \sigma$, we have

$$\frac{\text{MSE}(\widehat{K}_m^{\text{ort}}(\mathbf{x}, \mathbf{y}))}{\text{MSE}(\widehat{K}_m^{\text{base}}(\mathbf{x}, \mathbf{y}))} \approx 1 - \frac{(m-1) e^{-z^2} z^4}{d(1 - e^{-z^2})^2}. \quad (7.4)$$

We first give a brief note on the condition $m \leq d$. As discussed previously, it is often desirable to take $m > d$ in random feature methods to achieve adequately accurate approximations, so at first this condition may appear restrictive. However, since for a $\mathbf{G}_{\text{ort}}$ random matrix taking values in $\mathbb{R}^{ld \times d}$, the submatrices spanning rows $id+1$ to $(i+1)d$ (for $i = 0, \ldots, l-1$) are independent, the statistical properties of a random feature map with $m > d$ random features are easily deduced from analysing the $m \leq d$ cases. For example, the variance of a kernel estimator using ld random features is simply $1/l$ times that of an estimator using d random features. We will also often make the $m \leq d$ assumption in the analysis that follows, for these reasons.

This theorem shows that in asymptotically high dimensionality, orthogonal random feature maps yield superior mean squared error in pointwise Gaussian kernel estimation relative to their i.i.d. counterparts. This leaves many theoretical questions open, as alluded to in the discussion in Section 7.1, such as to what extent good empirical performance of orthogonal random features can be understood in non-asymptotic regimes, to what extent these findings are specific to the Gaussian kernel, and when it is in the interests of practitioners to use these feature maps. We now proceed with our original work.

7.3 The angular kernel

The first kernel we consider is the angular kernel. We recall that for two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, the angular kernel is given by

$$K^{\text{ang}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{2}{\pi} \arccos\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|} \right). \tag{7.5}$$

See the discussion around Example 5.3 for further background on the angular kernel. The angular kernel also has a representation given by

$$K^{\text{ang}}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\mathbf{w} \sim \eta} \left[\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \text{sign}(\langle \mathbf{w}, \mathbf{y} \rangle) \right], \tag{7.6}$$

for any isotropic $\eta \in \mathscr{P}(\mathbb{R}^d)$ with $\eta(\{\mathbf{0}\}) = 0$, as discussed in Example 5.3. Thus, a random feature approximation for the kernel exists, with random feature map given by

$$\Phi_m^{\text{ang}}(\mathbf{x}) = \left(\frac{1}{\sqrt{m}} \text{sign}(\langle \mathbf{w}_i, \mathbf{x} \rangle) \right)_{i=1}^m, \qquad (\mathbf{w}_i)_{i=1}^m \overset{\text{i.i.d.}}{\sim} \eta, \text{ for all } \mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}. \tag{7.7}$$

Our principal contribution in this section is to consider orthogonally coupled features for the angular kernel, and to provide an analysis of their statistical performance relative to i.i.d. features. In Section 7.6, we evaluate the performance of these features empirically.

To concisely define the orthogonal random feature map for the angular kernel, we first note that in the standard (i.i.d.) random feature map for the angular kernel, as given in Equation (7.7), we may take η to be $\mathcal{N}(\mathbf{0}, I)$. The map may thus be rewritten

$$\Phi_m^{\text{ang}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \text{sign}\left(\mathbf{G}\mathbf{x} \right), \tag{7.8}$$

where $\mathbf{G}$ is a random matrix taking values in $\mathbb{R}^{m \times d}$ with each entry an i.i.d. Gaussian (as described previously in Chapters 5 and 6), and the sign function is to be interpreted as being applied coordinate-wise to the vector $\mathbf{G}\mathbf{x}$. Note that any choice of η matching the conditions described above is valid, since the norms of the rows of the random matrix $\mathbf{G}$ do not affect the random feature map; the Gaussian distribution is simply selected here for convenience.

The idea now is to replace the unstructured matrix $\mathbf{G}$ with the random matrix $\mathbf{G}_{\text{ort}}$ taking values in $\mathbb{R}^{m \times d}$ for which all rows are orthogonal, as defined in Definition 7.1.

We may now precisely define the orthogonal random feature map for the angular kernel.

Definition 7.3 (**Othogonal random features for the angular kernel**)**.** The orthogonal random feature map with m random features for the angular kernel $K^{\text{ang}} : \mathbb{R}^d \setminus \{\mathbf{0}\} \times \mathbb{R}^d \setminus \{\mathbf{0}\} \to \mathbb{R}$ is defined by

$$\Phi_m^{\text{ang,ort}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \text{sign}\left(\mathbf{G}_{\text{ort}}\mathbf{x}\right), \tag{7.9}$$

where $\mathbf{G}_{\text{ort}}$ is a random Gaussian orthogonal matrix taking values in $\mathbb{R}^{m\times d}$, as defined in Definition 7.1.

In analogy with our analysis of random projections in Chapter 6, we define estimators of the angular kernel based on i.i.d. and orthogonal random features as follows:

$$\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\left\langle \text{sign}(\mathbf{G}\mathbf{x}), \text{sign}(\mathbf{G}\mathbf{y})\right\rangle, \tag{7.10}$$

$$\widehat{K}_m^{\text{ang,ort}}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\left\langle \text{sign}(\mathbf{G}_{\text{ort}}\mathbf{x}), \text{sign}(\mathbf{G}_{\text{ort}}\mathbf{y})\right\rangle. \tag{7.11}$$

We now proceed with a statistical analysis of these estimators. We remind the reader that a key quantity of interest will be the mean squared error of these estimators, defined by $\text{MSE}(\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})) = \mathbb{E}\left[(\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y}) - K^{\text{ang}}(\mathbf{x},\mathbf{y}))^2\right]$ for the i.i.d. estimator, and similarly for the orthogonal estimator. The expectation is taken over the distribution of the random features.

7.3.1 Statistical analysis of orthogonal random features for the angular kernel

Our main results in this section establish an exact expression for the mean squared error (MSE) of the angular kernel estimator based on i.i.d. random features for any pair of input vectors, and also show that the MSE associated with the orthogonal random features estimator is strictly better in all non-trivial circumstances. We are able to make such a precise analysis here since the non-linearity defining the random feature map, the coordinate-wise sign function, is straightforward to work with. We note this since this appears to be the exception as opposed to the rule; in Section 7.4, we will see that the trigonometric non-linearities that random Fourier features require complicate the analysis of orthogonal features significantly.

Our first result sets out the properties of the angular kernel estimator based on i.i.d. random features.

Lemma 7.4. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, the estimator $\widehat{K}_m^{\text{ang,base}}(\mathbf{x}, \mathbf{y})$ is unbiased for $K^{\text{ang}}(\mathbf{x}, \mathbf{y})$. Further, $\text{MSE}(\widehat{K}_m^{\text{ang,base}}(\mathbf{x}, \mathbf{y})) = \frac{4\theta_{\mathbf{x},\mathbf{y}}(\pi - \theta_{\mathbf{x},\mathbf{y}})}{m\pi^2}$, where $\theta_{\mathbf{x},\mathbf{y}}$ is the angle between the vectors $\mathbf{x}, \mathbf{y}$.

Proof sketch. Unbiasedness and the MSE formula follow straightforwardly from decomposing the right-hand side of Equation (7.10) into m i.i.d. terms, followed by some trigonometric calculations. □

Our main result is now the following guarantee of improved statistical performance of kernel estimation based on orthogonal random features relative to i.i.d. random features, which holds across all dimensionalities d and all non-trivial input pairs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. This generality should be contrasted with the conditions required for the result of Yu et al. (2016) in Theorem 7.2.

Theorem 7.5. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ such that $\mathbf{y}$ is not a scalar multiple of $\mathbf{x}$, the estimator $\widehat{K}_m^{\text{ang,ort}}(\mathbf{x}, \mathbf{y})$ for $K^{\text{ang}}(\mathbf{x}, \mathbf{y})$ is unbiased and satisfies:

$$\text{MSE}(\widehat{K}_m^{\text{ang,ort}}(\mathbf{x}, \mathbf{y})) < \text{MSE}(\widehat{K}_m^{\text{ang,base}}(\mathbf{x}, \mathbf{y}))\,. \tag{7.12}$$

Proof sketch. Analogously to the proof of Lemma 7.4, we decompose the right-hand side of Equation (7.11) into m terms, which despite being identically distributed, are not independent. Calculation of the MSE for the orthogonal random feature estimator therefore requires more involved trigonometric calculations. □

The condition on $\mathbf{y}$ not being a scalar multiple of $\mathbf{x}$ in Theorem 7.5 rules out exactly the cases where the angle between $\mathbf{x}$ and $\mathbf{y}$ is 0 or π. In both of these cases, both the i.i.d. and orthogonal random feature maps yield 0 MSE (exact) estimators.

7.4 Orthogonal random Fourier features

Having dealt with orthogonal random features for the angular kernel, we now shift our attention to random Fourier features for stationary kernels. Further detail on random feature maps for these kernels is given in Example 5.2. We say that a stationary kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is *normalised* if $K(\mathbf{x}, \mathbf{x}) = 1$ for any (hence all) $\mathbf{x} \in \mathbb{R}^d$. Working with normalised kernels guarantees that the measure $\eta \in \mathscr{M}(\mathbb{R}^d)$ that appears in the Bochner

identity

$$K(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \exp(\mathrm{i}\langle \mathbf{x}-\mathbf{y},\mathbf{w}\rangle)\eta(d\mathbf{w}) \tag{7.13}$$

is in fact a probability measure. This is a convenient assumption for our exposition, since it results in less cluttered notation, thus allowing us to focus our attention on the orthogonal couplings of interest. The general case, where $\eta(\mathbb{R}^d) \neq 1$, may be understood as a straightforward generalisation of the case where η is a probability measure. Assuming that K is not the zero kernel (so that η is not the zero measure), and noting that $\eta(\mathbb{R}^d) = K(\mathbf{x},\mathbf{x}) < \infty$, we write $\mu = \eta(\mathbb{R}^d)^{-1}\eta \in \mathscr{P}(\mathbb{R}^d)$ for the normalised version of η. The random feature map for η, Φ^η, may be expressed concisely in terms of that for μ, Φ^μ, as follows:

$$\Phi^\eta(\mathbf{x}) = \eta(\mathbb{R}^d)^{1/2}\Phi^\mu(\mathbf{x}), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^d. \tag{7.14}$$

To emphasise that the Fourier measure we deal with in what follows is assumed to be a probability measure, we will generally use the notation μ rather than η to represent Fourier measures associated with stationary kernels of interest.

Definition 7.6 (**Orthogonal random Fourier features**)**.** Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a stationary isotropic normalised kernel, let $\mu_K \in \mathscr{P}(\mathbb{R}^d)$ be its associated Fourier measure, and let $\rho_K \in \mathscr{P}(\mathbb{R}_{\geq 0})$ be the distribution of the norm of a vector drawn from μ. The orthogonal random feature map for K is given by

$$\Phi_m^{K,\mathrm{ort}}(\mathbf{x}) = ((\cos(\langle \mathbf{x}, \mathbf{w}_i^{\mathrm{ort}}\rangle))_{i=1}^m, (\sin(\langle \mathbf{x}, \mathbf{w}_i^{\mathrm{ort}}\rangle))_{i=1}^m), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^d, \tag{7.15}$$

where $\mathbf{w}_i^{\mathrm{ort}} \in \mathbb{R}^d$ is the i^{th} row of the random matrix $\mathbf{W}$, the distribution of which is specified as follows. We vertically stack $l = \lceil m/d \rceil$ random matrices drawn independently from Haar measure on the manifold of $d \times d$ orthogonal matrices. Each row is then scaled independently with a random variable drawn from ρ_K. This ensures that each row of $\mathbf{W}$ marginally has the distribution μ_K.

When K is an isotropic Gaussian kernel, the definition above matches that of Yu et al. (2016). We note that these orthogonal couplings have connections to randomised quasi-Monte Carlo (Dick and Pillichshammer, 2010) and other techniques for encouraging diversity in collections of samples. An important distinction is that (randomised) quasi-Monte Carlo sequences are generally designed to give good performance for estimating integrals over function classes with particular properties, such as belonging to a particular Hilbert space, or satisfying certain smoothness conditions. Here however, we

are concerned particularly with the trigonometric functions that define random Fourier features.

We will also refer to stationary isotropic kernels as RBF kernels, after *radial basis functions*, as such a kernel K can be expressed as $K(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x}-\mathbf{y})$, with ϕ a radial basis function. We summarise several RBF kernels in Table 7.1, along with the densities of their associated Fourier distributions.

Name	Positive-definite function	Fourier density
Gaussian	$\exp\left(-\frac{1}{2\sigma^2}z^2\right)$	$\frac{1}{(2\pi\sigma^{-2})^{d/2}}\exp\left(-\frac{1}{2\sigma^{-2}}\|\mathbf{w}\|_2^2\right)$
Matérn-ν	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}z\right)^{\nu}K_{\nu}\left(\sqrt{2\nu}z\right)$	$\frac{\Gamma(\nu+d/2)}{\Gamma(\nu)(2\nu\pi)^{d/2}}\left(1+\frac{1}{2\nu}\|\mathbf{w}\|^2\right)^{-\nu-p/2}$

Table 7.1. Common RBF kernels, their corresponding positive definite functions, and their Fourier transforms.

We now turn our attention to developing statistical understanding of orthogonal random feature maps. As in the analysis of Yu et al. (2016), we focus on mean squared error of the corresponding kernel evaluation estimators. We restrict to $m \leq d$ in our analysis, as discussed at the end of Section 7.2; this means that the vectors $(\mathbf{w}_i)_{i=1}^m$ appearing in the orthogonal random feature map in Expression (7.15) are all mutually orthogonal almost surely. Since in general the collections of vectors $(\mathbf{w}_i)_{i=1}^d$, $(\mathbf{w}_i)_{i=d+1}^{2d}$, ... are all independent, the mean square error analysis straightforwardly generalises to the case where $m > d$.

7.5 The variance of orthogonal random Fourier features

Having defined orthogonal random Fourier features generally, we now turn our attention to establishing the benefits of these random feature maps for several classes of RBF kernels. Suppose we are considering the estimator associated with a random feature map of an RBF kernel evaluated at two inputs $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$, and let $\mathbf{z} = \mathbf{x}-\mathbf{y}$. Our asymptotic analysis focusses on two regimes: (i) fixed dimensionality d and small $\|\mathbf{z}\|$, in Section 7.5.1; and (ii) fixed $\|\mathbf{z}\|$ and large d, in Section 7.5.2. We also provide a non-asymptotic computational analysis of the effectiveness of orthogonal random Fourier features in Section 7.5.3.

Before launching into these analyses, a first contribution is the introduction of the *charm function* associated with an RBF kernel. The charm function turns out to play a key role in our asymptotic analysis, and provides a useful heuristic in assessing the quality of orthogonal random Fourier features in non-asymptotic regimes.

Definition 7.7 (**The charm function**)**.** Consider an RBF kernel $K(\mathbf{x},\mathbf{y}) = \phi_K(\|\mathbf{x}-\mathbf{y}\|)$, with ϕ_K twice differentiable. We define the charm function Ψ_K of K as the function $\Psi_K : \mathbb{R}^d \to \mathbb{R}$ defined at the point $\mathbf{z} = \mathbf{x}-\mathbf{y}$ by

$$\Psi_K(\mathbf{z}) = \|\mathbf{z}\|^2 \frac{\mathrm{d}^2\phi_K^2(t)}{\mathrm{d}t^2}\bigg|_{t=\|\mathbf{z}\|} - \|\mathbf{z}\| \frac{\mathrm{d}\phi_K^2(t)}{\mathrm{d}t}\bigg|_{t=\|\mathbf{z}\|}. \tag{7.16}$$

We shall show that the *charm* function plays a crucial role in understanding the behaviour of orthogonal transforms for the large dimensionality regime. Indeed, in the large dimensionality regime, the superiority of orthogonal random Fourier features follows from the positivity of the charm function across the entire domain. This in turn is a consequence of the intricate connection between classes of positive definite RBF kernels not parametrised by data dimensionality and *completely monotone* functions, defined below.

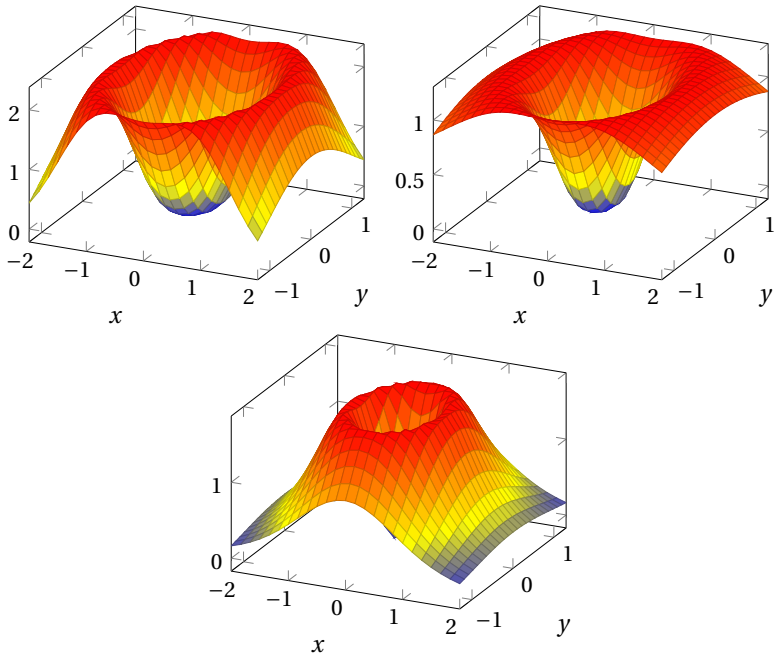


Figure 7.1. Plots of the charm function Ψ_K for $d = 2$ and several RBF kernels. Left: Gaussian kernel. Centre: kernel defined by the positive definite function $\phi(\|\mathbf{z}\|) = \left(1+\|\mathbf{z}\|^2\right)^{-1/2}$. Right: kernel defined by the positive definite function $\phi(\|\mathbf{z}\|) = \left(1+\|\mathbf{z}\|^2\right)^{-1}$. "Warmer" regions indicate larger gains from applying structured approach. Positive values of Ψ_K imply asymptotic superiority of the structured orthogonal estimator. All charm functions plotted are positive everywhere.

The benefits of using orthogonal features in comparison to i.i.d. features can be quantitatively measured by the value of the charm of the kernel at point $\mathbf{z} = \mathbf{x}-\mathbf{y}$ for large data dimensionality. Large charm values (see Figure 7.1) indicate regions where the mean squared error (defined as $\mathrm{MSE}(\widehat{K}(\mathbf{x},\mathbf{y})) = \mathbb{E}[(\widehat{K}(\mathbf{x},\mathbf{y})-K(\mathbf{x},\mathbf{y}))^2]$, where again the expectation

is over the distribution of the random features) of the orthogonal estimator is significantly smaller than for an i.i.d. estimator and thus the geometry of the charm function across the domain gives strong guidance on the accuracy gains available from using orthogonal features.

At the outset, we highlight Theorems 7.8 and 7.14 as key theoretical results.

7.5.1 Small $\|\mathbf{z}\|$ analysis

Our main result in this section compares the mean squared error (MSE) of the i.i.d. random feature estimator based on independent sampling to the MSE of the estimator applying random orthogonal feature maps for close enough input points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Throughout, we write $\mathbf{z} = \mathbf{x} - \mathbf{y}$.

Theorem 7.8. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be an RBF kernel and let $\mu_K \in \mathscr{P}(\mathbb{R}^d)$ be its associated Fourier measure. Suppose that $\mathbb{E}_{\mathbf{w}\sim\mu_K}\left[\|\mathbf{w}\|^4\right] < \infty$. Then for sufficiently small $\|\mathbf{z}\|$, we have

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x},\mathbf{y})) > \mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y}))\,. \tag{7.17}$$

Proof sketch. The proof relies on a Taylor expansion of $\mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y}))$ in $\|\mathbf{z}\|$ around 0 up to fourth order. The existence of this expansion is guaranteed by the moment condition in the theorem statement. □

The assumptions of the theorem above are satisfied for many classes of RBF kernels such as Gaussian and Matérn with smoothness parameter $\nu > 2$. In contrast to the earlier work of Yu et al. (2016), this result holds for a wide class of RBF kernels K, and in a different asymptotic regime. We shall see in Section 7.5.3 that in fact the region of "sufficiently small" $\|\mathbf{z}\|$ is often surprisingly large in practice.

7.5.2 High-dimensional analysis

In this section, we consider the same asymptotic regime as in Yu et al. (2016); fixed $\|\mathbf{z}\|$, and large d. We first present our main theoretical result[1], which under several conditions shows that the leading term in a large d expansion of the difference of mean squared

[1] The statement of this theorem has been significantly simplified relative to the corresponding result that appears in Choromanski et al. (2018a).

errors is controlled by the *charm function* Ψ_K associated with the kernel K, as introduced in Equation (7.16).

Theorem 7.9. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a stationary, isotropic, normalised kernel, defined by $K(\mathbf{x}, \mathbf{y}) = \phi(\|\mathbf{x} - \mathbf{y}\|)$, with the property that this defines a valid kernel for any $d \in \mathbb{N}$. For each $d \in \mathbb{N}$, let $\mu_d \in \mathscr{P}(\mathbb{R}^d)$ be the Fourier distribution associated with $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Let $z \in \mathbb{R}_{\geq 0}$ be fixed, and let $\mathbf{x}^{(d)}, \mathbf{y}^{(d)} \in \mathbb{R}^d$ be any sequence of vectors with $\|\mathbf{x}^{(d)} - \mathbf{y}^{(d)}\| = z$ for all d. Suppose that the following two conditions hold:

1. The MGFs M_d of μ_d satisfy $M_d(\mathbf{x}^{(d)} - \mathbf{y}^{(d)}) = o(\sqrt{d})$.
2. If $\mathbf{w}_d^{(d)} \sim \mu_d$, then $\|\mathbf{w}^{(d)}\|_2^2 / \mathbb{E}\left[\|\mathbf{w}^{(d)}\|_2^2\right]$ converges in probability to the constant 1 as $d \to \infty$.

Then we have the following limiting expressions for the difference in MSEs between i.i.d. and orthogonal random feature kernel estimation:

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x}^{(d)}, \mathbf{y}^{(d)})) - \mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x}^{(d)}, \mathbf{y}^{(d)})) = \frac{m-1}{m}\left(\frac{1}{8d}\Psi_K(\mathbf{x}^{(d)} - \mathbf{y}^{(d)}) + o(d^{-1})\right). \quad (7.18)$$

Unfortunately, the conditions on the Fourier distributions appearing in Theorem 7.9 rule out several classes of kernels of interest, such as Matérn kernels. Interestingly, we will see that the charm function still accurately describes the behaviour of such kernels in Sections 7.5.3 and 7.6. Theorem 7.9 leads to many important corollaries, as we show below. In particular, we highlight that the charm function Ψ_K associated with the kernel K is central in determining the relative performance of orthogonal random features and i.i.d. features in high dimensions, due to its place in Equation (7.18). As special cases, Theorem 7.9 implies all earlier theoretical results for orthogonal random features for the Gaussian kernel (Yu et al., 2016); we review this special case below.

Corollary 7.10. If K is an isotropic Gaussian kernel then for any fixed $\|\mathbf{z}\| \in \mathbb{R}_{>0}$ and d large enough the orthogonal random feature map outperforms the i.i.d. random feature map in MSE. This is implied by noting that the sequence of Fourier measures associated with the Gaussian kernel K across different dimensionalities are all Gaussian measures with diagonal covariance matrices (having fixed diagonal elements). Thus, the moment generating functions all take the form $\exp(\frac{1}{2}\lambda^2\|\mathbf{z}\|^2)$, where λ^2 is the bandwidth of the kernel, and Condition 1 of Theorem 7.9 is satisfied. Secondly, because different coordinates of the Fourier measure are i.i.d. with finite second moments, the weak law of large numbers implies Condition 2. It remains to observe that the charm function is positive

for the Gaussian kernel K. Straightforward calculation (taking $\lambda = 1$ for simplicity) yields $\Psi_K(\mathbf{z}) = 4\|\mathbf{z}\|^4 e^{-\|\mathbf{z}\|^2}$ (see Figure 7.1), which is clearly positive away from $\mathbf{0}$.

The fact that charm is non-negative across the entire domain for the family of Gaussian kernels is not a coincidence. In fact, the following result holds generally.

Theorem 7.11. Let $\phi : \mathbb{R} \to \mathbb{R}$ be such that for every $d \in \mathbb{N}$, $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by $K_d(\mathbf{x}, \mathbf{y}) = \phi(\|\mathbf{x} - \mathbf{y}\|)$ is a positive definite kernel. Then for each such K_d, the charm function Ψ_{K_d} is non-negative away from $\mathbf{0}$.

The result above follows straightforwardly once the following connection between positive definite functions ϕ considered above and *completely monotone* functions has been established.

Definition 7.12. A function $\varphi : [0, \infty) \to \mathbb{R}$ which is in $C[0, \infty) \cap C^\infty(0, \infty)$ and which satisfies $(-1)^r \frac{d^r \varphi}{dx^r} \geq 0$ for all $r \in \mathbb{N} \cup \{0\}$, is called completely monotone on $[0, \infty)$.

Schoenberg (1938) gave the following characterisation of completely monotone functions in terms of positive definite functions, which is the key to establishing positivity of the charm function for kernels satisfying the requirements of Theorem 7.9.

Theorem 7.13 (Schoenberg, 1938)**.** A function $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is completely monotone iff for each $d \in \mathbb{N}$, the function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ as $K(\mathbf{x}, \mathbf{y}) = \varphi(\|\mathbf{x} - \mathbf{y}\|^2)$ is positive definite.

Combining Theorem 7.9 with Theorem 7.11, we obtain the following key result.

Theorem 7.14 (**Superiority of the orthogonal transform**)**.** Under the assumptions of Theorem 7.9, for any fixed $z \in \mathbb{R}_{>0}$, for sufficiently large d, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\|\mathbf{x} - \mathbf{y}\| = z$,

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x}, \mathbf{y})) > \mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x}, \mathbf{y}))\,. \tag{7.19}$$

This concludes the development of our asymptotic theory. We emphasise the general observation that the charm function $\Psi_K(\mathbf{x} - \mathbf{y})$, describing the asymptotic effectiveness of orthogonal random Fourier features for a pair of input vectors, tends to initially grow as $\|\mathbf{x} - \mathbf{y}\|$ as increases, before decaying back to 0 (see Figure 7.1). We shall see in the next section that this behaviour is also observed in non-asymptotic regimes, and that the charm function provides a effective heuristic for understanding when orthogonal random Fourier features are useful even in relatively low-dimensional settings.

7.5.3 Non-asymptotic analysis

Complementing the theoretical asymptotic results presented above, we provide additional analysis of the behaviour of orthogonal random features in non-asymptotic regimes. The analysis centres on Proposition 7.15, which expresses the difference in MSE between i.i.d. and orthogonal random features in terms of univariate integrals, which although generally intractable, can be accurately and efficiently evaluated by deterministic numerical integration.

Proposition 7.15. For an RBF kernel K on $\mathbb{R}^d$ with Fourier measure μ_K and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, writing $\mathbf{z} = \mathbf{x} - \mathbf{y}$, we have:

$$\begin{aligned} &\mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x},\mathbf{y})) \qquad (7.20)\\ &= \frac{m-1}{m}\mathbb{E}_{R_1,R_2}\left[\frac{J_{\frac{d}{2}-1}\left(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|\right)\Gamma(d/2)}{\left(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|/2\right)^{\frac{d}{2}-1}}\right] - \frac{m-1}{m}\mathbb{E}_{R_1}\left[\frac{J_{\frac{d}{2}-1}(R_1\|\mathbf{z}\|)\,\Gamma(d/2)}{(R_1\|\mathbf{z}\|/2)^{\frac{d}{2}-1}}\right]^2, \end{aligned}$$

where R_1, R_2 are distributed i.i.d. from ρ_K, and J_α is the Bessel function of the first kind of degree α.

Firstly, in Figure 7.2, we plot the difference in MSE between i.i.d. random features and orthogonal random features for the Gaussian kernel, noting that orthogonal features provide superior MSE across a wide range of values of $\|\mathbf{z}\|$. In the same plots, we show the value of the kernel K and of the charm function Ψ_K, noting that the charm function describes the benefits of orthogonal features accurately, even in the case of low dimensions. In all plots in this section, we write ΔMSE for $\mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y}))$, so that $\Delta\mathrm{MSE} > 0$ corresponds to superior performance of orthogonal features over i.i.d. features.

Secondly, we illustrate the effectiveness of the charm function as a heuristic for the statistical efficiency of orthogonal features, even for kernels not satisfying the conditions of Theorem 7.9. In Figure 7.3, we plot difference in MSE (using Proposition 7.15), charm function and kernel function for Matérn-1/2, Matérn-5/2, and Matérn-10 kernels. In these cases, the charm function provides a good qualitative match for the behaviour of the ΔMSE curve. This suggests that the charm function provides a sensible heuristic for judging the effectiveness of orthogonal random Fourier features in general, and also that it may be possible to strengthen the theoretical results appearing in Section 7.5.2 further. Interestingly, the charm function also indicates smaller improvements for Matérn

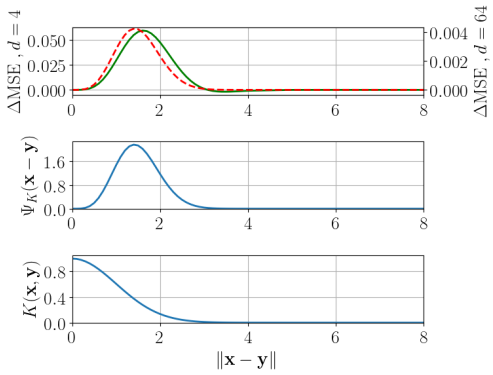


Figure 7.2. Difference between i.i.d. MSE and orthogonal MSE (top), charm function Ψ_K (middle), and kernel K (bottom) for the Gaussian kernel for a range of dimensionalities. Solid green is $d = 4$, dotted red is $d = 64$.

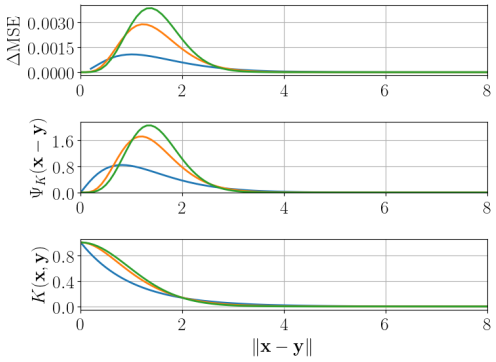


Figure 7.3. Comparison of charm function and MSE improvements for Matérn-1/2 (blue), Matérn-5/2 (yellow), and Matérn-10 (green) kernels.

kernels with heavier-tailed Fourier distributions; we will see that this is borne out in the experiments appearing in Section 7.6.

Finally, we consider an RBF kernel $K(\mathbf{x},\mathbf{y}) = \phi(\|\mathbf{x}-\mathbf{y}\|)$ which does *not* correspond to a completely monotone function. Let $d = 3$, and consider the Fourier measure μ that puts unit mass uniformly on the sphere $S^2 \subseteq \mathbb{R}^3$. As this is a finite isotropic measure on $\mathbb{R}^3$, there exists a corresponding RBF kernel K, which by performing an inverse Fourier transform can be shown to be $K(\mathbf{x},\mathbf{y}) = \sin(\|\mathbf{x}-\mathbf{y}\|)/\|\mathbf{x}-\mathbf{y}\|$. We term this the sinc kernel. Since the kernel takes on negative values for certain inputs, it does not correspond to a completely monotone function. Given the particular form of the Fourier measure, we may compute the difference in MSEs as given in Proposition 7.15 exactly, which yields

$$\mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y})) = \frac{2}{3}\left(\frac{\sin(\sqrt{2}\|\mathbf{z}\|)}{\sqrt{2}\|\mathbf{z}\|} - \frac{\sin^2(\|\mathbf{z}\|)}{\|\mathbf{z}\|^2}\right). \tag{7.21}$$

We plot this function in Figure 7.4, noting there are large regions where orthogonal features are outperformed by i.i.d. features. Thus it may not be possible to relax the requirement in Theorem 7.14 that the positive definite function ϕ_K corresponds to a completely monotone function, as in Theorem 7.13. Indeed, we also note that the requirement of asymptotically large d also cannot be dropped from the statement of Theorem 7.14. In Figure 7.2, for the Gaussian kernel in $d = 4$ dimensions, we note that there are values of $\|\mathbf{x}-\mathbf{y}\|$ between 3 and 4 for which orthogonal random features lead to worse MSE than i.i.d. random features. However, the amount by which this is the case is small relative to the gains that can be

achieved when $\|\mathbf{x}-\mathbf{y}\|$ is between approximately 0.5 and 3. These observations resolve a question that was left open by Yu et al. (2016), namely "Do orthogonal random Fourier features always improve the variance of approximation?", in the negative. However, it leaves open the (imprecise) question as to whether orthogonal random Fourier features can be shown to "never be much worse than i.i.d. features, and often much better", as we empirically observe to be the case.

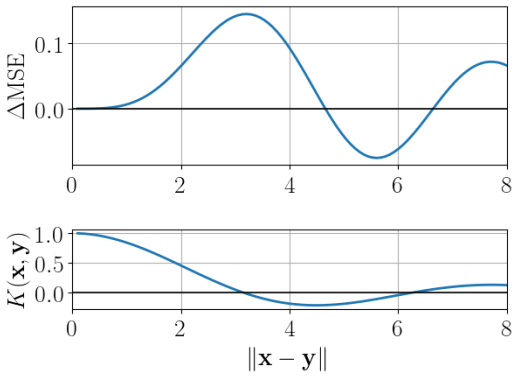
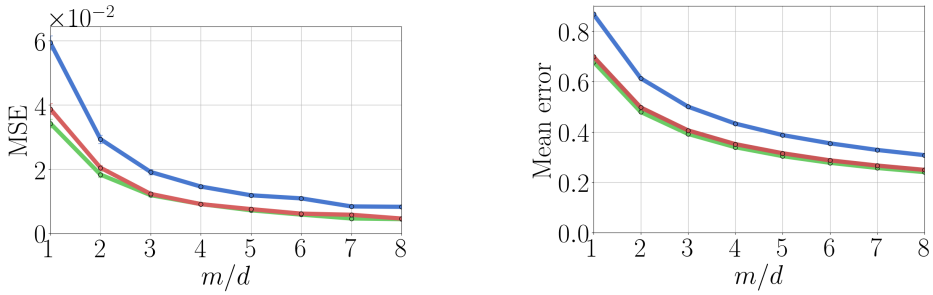


Figure 7.4. Difference in MSE for orthogonal and i.i.d. features for the sinc kernel, which does not correspond to a completely monotone positive definite function.

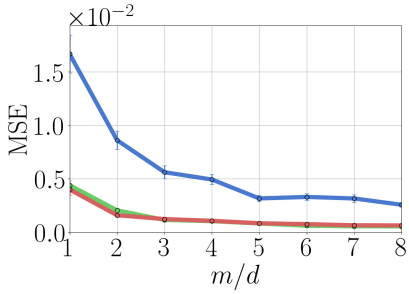
7.6 Experiments

We complement the theoretical results for pointwise kernel approximations in earlier sections with empirical studies of the effectiveness and limits of orthogonal random features in a variety of downstream applications. We also compare against structured orthogonal random features (SORF, introduced in the Gaussian case by Yu et al., 2016), where instead of directions of samples from μ_K being marginally distributed from Unif(S^{d-1}) we use the rows of a Hadamard–Rademacher random matrix, as discussed in Chapter 6 — in all cases, we use $k = 3$ blocks. As for Gaussian orthogonal matrices in Definition 7.1, when $m > d$, we vertically stack independent Hadamard–Rademacher matrices to obtain a random matrix with the required number of rows. As discussed in Chapter 6, where data dimensionality is not a power of 2, we pad with the necessary number of zero coordinates so that Hadamard–Rademacher matrices may be applied. We examine various numbers m of random features, while d is the dimensionality of the data.

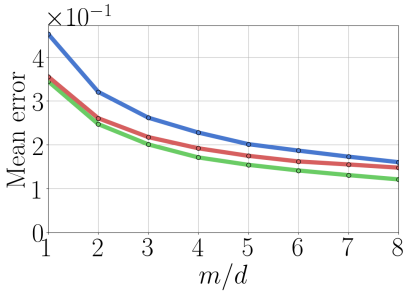


(a) Angular, pointwise

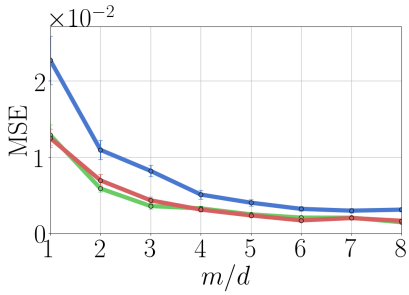
(b) Angular, Gram matrix



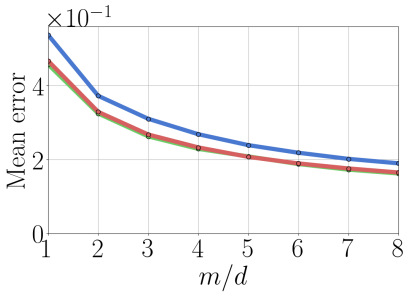
(c) Gaussian, pointwise



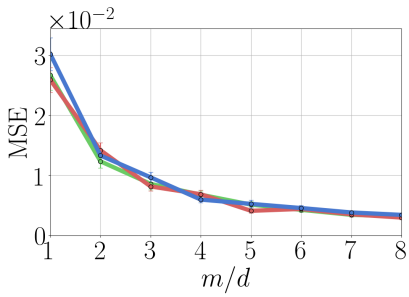
(d) Gaussian, Gram matrix



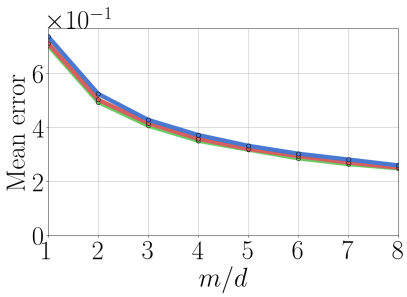
(e) Matérn-5/2, pointwise



(f) Matérn-5/2, Gram matrix



(g) Laplace, pointwise



(h) Laplace, Gram matrix

Figure 7.5. Pointwise kernel evaluation MSE (left column) and normalised Frobenius norm error for Gram matrix approximation (right column) for the UCI “wine” dataset for Gaussian (top), Matérn-5/2 (center) and Laplace (bottom) kernels. Estimators are i.i.d. random features (blue), orthogonal random features (green) and approximate Hadamard–Rademacher random features (red).

7.6.1 Pointwise kernel and Gram matrix estimation

In this experiment, we study the estimation, via random feature maps, of pointwise evaluation of kernels and of kernel Gram matrices. We use MSE as an error measure for pointwise estimation, and normalised Frobenius norm as a measure of error for Gram matrices (so that the error incurred by estimating the Gram matrix $\mathbf{X}$ with the matrix $\widehat{\mathbf{X}}$ is $\|\mathbf{X}-\widehat{\mathbf{X}}\|_{\mathrm{F}}/\|\mathbf{X}\|_{\mathrm{F}}$). Kernel bandwidths are set via the median trick (Yu et al., 2016) where relevant. We estimate pointwise kernel values (for a pair of randomly selected datapoints) and Gram matrices (on full datasets), for the angular kernel and a range of stationary isotropic kernels, on a variety of UCI regression datasets (Lichman, 2013); Figure 7.5 depicts these results for the "wine" dataset, whilst results for other datasets are presented in Appendix 7.B — qualitatively the results are similar to those presented here. We plot the estimated mean Frobenius norm error, and bootstrapped estimates of standard error of the mean error estimates; in Figure 7.5, these error bars are extremely small. Note that the orthogonal and approximate-orthogonal approaches consistently offer statistical improvements relative to i.i.d. random features under the metrics considered here. Note also that for random Fourier features, the improvement in performance is most pronounced for kernels with light-tailed Fourier distributions, as suggested by the theoretical developments in Section 7.5. Finally, note that the Laplace kernel is a special case of the Matérn kernel in Figure 7.1 (with parameter $\nu = 1/2$).

7.6.2 Gaussian processes

We consider random feature approximations to Gaussian processes (GPs) for regression, and report (i) KL divergence from the approximate predictive distribution obtained via random Fourier feature approximations to the true predictive distribution obtained by an exactly-trained GP, and (ii) predictive root mean squared error (RMSE) on test sets. Experiments were run on a variety of UCI regression datasets, using a randomly selected 80/20 train/test split. In Figures 7.2 and 7.3, results are shown for regression on the Boston housing dataset (Lichman, 2013); results for other datasets are presented in Appendix 7.B, together with timings. We use Gaussian, Matérn-5/2, and Laplace covariance kernels for the GP, and denote feature maps based on i.i.d. features, orthogonal features, and Hadamard–Rademacher approximations to orthogonal features by IID, ORF, and SORF, respectively.

Kernel	Feature map	$m/d = 1$	$m/d = 2$	$m/d = 3$	$m/d = 4$
Gaussian	IID	104.2 (10.0)	34.21 (1.5)	15.6 (0.87)	11.05 (0.73)
	ORF	**100.4 (5.6)**	**26.62 (1.5)**	**15.1 (1.1)**	**8.707 (0.42)**
	SORF	108.9 (12.0)	32.29 (2.9)	16.25 (1.3)	10.15 (0.73)
Matérn-5/2	IID	160.3 (19.0)	47.88 (2.6)	25.87 (1.3)	18.61 (1.2)
	ORF	**123.2 (6.3)**	**41.66 (1.3)**	**21.78 (0.89)**	**16.66 (0.85)**
	SORF	166.4 (21.0)	44.74 (3.1)	25.14 (0.91)	16.89 (1.1)
Laplace	IID	337.2 (19.0)	126.4 (4.1)	69.66 (3.6)	50.99 (1.7)
	ORF	299.5 (17.0)	**117.7 (3.1)**	**68.4 (2.6)**	**44.25 (1.7)**
	SORF	**298.3 (7.6)**	121.1 (2.5)	70.56 (1.9)	47.88 (1.5)

Table 7.2. Approximate GP regression results on Boston dataset. Reported numbers are average KL divergence from true posterior, along with bootstrap estimates of standard error (in parentheses).

Kernel	Feature map	$m/d = 1$	$m/d = 2$	$m/d = 3$	$m/d = 4$
Gaussian	IID	**0.54 (0.02)**	0.48 (0.01)	**0.43 (0.008)**	0.4 (0.01)
	ORF	0.59 (0.01)	**0.44 (0.008)**	0.43 (0.009)	**0.39 (0.006)**
	SORF	0.6 (0.02)	0.5 (0.02)	0.44 (0.009)	0.41 (0.008)
Matérn-5/2	IID	0.63 (0.02)	0.49 (0.008)	0.45 (0.01)	0.43 (0.006)
	ORF	**0.57 (0.02)**	0.47 (0.02)	**0.42 (0.006)**	**0.42 (0.008)**
	SORF	0.61 (0.04)	**0.47 (0.02)**	0.44 (0.01)	0.43 (0.01)
Laplace	IID	0.69 (0.04)	0.56 (0.02)	0.51 (0.01)	0.48 (0.01)
	ORF	0.65 (0.04)	0.54 (0.02)	0.51 (0.01)	0.48 (0.01)
	SORF	**0.62 (0.02)**	**0.53 (0.01)**	**0.49 (0.02)**	**0.47 (0.01)**

Table 7.3. Approximate GP regression results on Boston dataset. Reported numbers are average test RMSE, along with bootstrap estimates of standard error (in parentheses).

We remark that orthogonally-coupled features tend to obtain significantly better KL divergences to the true predictive posterior relative to i.i.d. features. This is encouraging, as it shows that the advantages of orthogonality demonstrated theoretically in earlier sections for pointwise kernel approximation are borne out in practice in downstream tasks. The results for test RMSE are less clear, with i.i.d. features occasionally outperforming orthogonally-structured features. This is perhaps not surprising, in the sense that performing random feature approximations in GP regression may be interpreted as altering the kernel used to perform regression; if the original kernel is not particularly well-suited to the dataset at hand, then close approximation of the kernel may not result in better predictive performance.

7.7 Discussion

The analysis presented in Sections 7.3 and 7.5 furthers the current understanding of orthogonality as a variance reduction mechanism in random feature methods. It was possible to make strong statements for the angular kernel (across all input pairs to the kernel, and for all dimensionalities), which is due in part to the straightforward non-linearity present in the random feature maps for this kernel. In contrast, the trigonometric non-linearities present for random Fourier features complicate analysis, and the results obtained for these random feature maps correspondingly have stricter assumptions, either principally on the input points to the kernel estimator, or the dimensionality of the space which these points occupy. Nevertheless, the charm function appears to behave as a good heuristic for judging the effectiveness of orthogonal Fourier random features in general.

Our experimental results, and non-asymptotic computational analysis utilising Proposition 7.15, indicate that the improved performance of orthogonal features holds much more generally than is currently proven theoretically, although the counterexamples described in Section 7.5.3 show that improved statistical performance from orthogonal is not necessarily guaranteed. With this said, however, the charm function and the associated guarantees of non-negativity given in Section 7.5.2 go some way to giving practitioners a toolkit for understanding orthogonal features more intuitively. An important direction for future work will be to further bridge this gap between theory and empirical observations.

Another direction we highlight, in accordance with the discussion at the end of Chapter 6, is that there is also much motivation for improving understanding of computationally efficient methods for working with, and sampling from, distributions over the orthogonal group of matrices, both improving our understanding of the effectiveness of Hadamard–Rademacher random matrices, and exploring other constructions more generally.

Appendix 7.A Proofs

7.A.1 Proofs of results in Section 7.3

Recall that the angular kernel estimator based on orthogonal features is given by

$$\widehat{K}_m^{\mathrm{ang,ort}}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\langle \mathrm{sign}(\mathbf{G}_{\mathrm{ort}}\mathbf{x}), \mathrm{sign}(\mathbf{G}_{\mathrm{ort}}\mathbf{y})\rangle\,, \tag{7.22}$$

and the angular kernel estimator based on i.i.d. features is given by

$$\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\langle \text{sign}(\mathbf{G}\mathbf{x}), \text{sign}(\mathbf{G}\mathbf{y})\rangle\,, \tag{7.23}$$

where the function sign acts on vectors coordinate-wise. In what follows, we write $\mathbf{G}_{\text{ort}}^i$ for the ith row of $\mathbf{G}_{\text{ort}}$, and $\mathbf{G}^i$ for the ith row of $\mathbf{G}$. We first demonstrate unbiasedness of the two estimators above — in fact, since $\mathbf{G}_{\text{ort}}^i$ is equal to $\mathbf{G}^i$ in distribution, unbiasedness of $\widehat{K}_m^{\text{ang,ort}}(\mathbf{x},\mathbf{y})$ follows immediately from unbiasedness of $\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})$. Thus, we compute as follows:

$$\mathbb{E}\left[\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\text{sign}\left(\mathbf{G}^i\mathbf{x}\right)\text{sign}\left(\mathbf{G}^i\mathbf{y}\right)\right] = \mathbb{E}\left[\text{sign}\left(\mathbf{G}^1\mathbf{x}\right)\text{sign}\left(\mathbf{G}^1\mathbf{y}\right)\right]\,, \tag{7.24}$$

from the fact that $\mathbf{G}^1,\ldots,\mathbf{G}^m$ are i.i.d. . Denoting the angle between $\mathbf{x}$ and $\mathbf{y}$ by θ , we note that

$$\begin{aligned}\mathbb{E}\left[\text{sign}\left(\mathbf{G}^1\mathbf{x}\right)\text{sign}\left(\mathbf{G}^1\mathbf{y}\right)\right] &= \mathbb{P}\left(\text{sign}\left(\mathbf{G}^1\mathbf{x}\right) = \text{sign}\left(\mathbf{G}^1\mathbf{y}\right)\right) - \mathbb{P}\left(\text{sign}\left(\mathbf{G}^1\mathbf{x}\right) \neq \text{sign}\left(\mathbf{G}^1\mathbf{y}\right)\right)\\ &= \left(1-\frac{\theta}{\pi}\right) - \left(\frac{\theta}{\pi}\right)\\ &= K^{\text{ang}}(\mathbf{x},\mathbf{y})\,,\end{aligned} \tag{7.25}$$

as required.

With unbiasedness of both estimators established, we now prove the remaining claim of the following lemma.

Lemma 7.4. For any $\mathbf{x},\mathbf{y}\in\mathbb{R}^d\setminus\{\mathbf{0}\}$, the estimator $\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})$ is unbiased for $K^{\text{ang}}(\mathbf{x},\mathbf{y})$. Further, $\text{MSE}(\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})) = \frac{4\theta_{\mathbf{x},\mathbf{y}}(\pi-\theta_{\mathbf{x},\mathbf{y}})}{m\pi^2}$, where $\theta_{\mathbf{x},\mathbf{y}}$ is the angle between the vectors $\mathbf{x},\mathbf{y}$.

Proof. We have already shown unbiasedness. For the MSE result, we introduce the following notation:

$$S_i = \text{sign}\left(\left\langle \mathbf{G}^i,\mathbf{x}\right\rangle\right)\text{sign}\left(\left\langle \mathbf{G}^i,\mathbf{y}\right\rangle\right)\,, \tag{7.26}$$

for $i=1,\ldots,m$. Now observe that as $\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})$ is unbiased, we have

$$\text{MSE}\left(\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})\right) = \frac{1}{m}\text{Var}\left(S_1\right)\,, \tag{7.27}$$

since $\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})$ is a sum of the i.i.d. terms $S_1,\ldots,S_m$. Now

$$\text{Var}(S_1) = 1 - \left(\frac{\pi - 2\theta_{\mathbf{x},\mathbf{y}}}{\pi}\right)^2 = \frac{4\theta_{\mathbf{x},\mathbf{y}}(\pi - \theta_{\mathbf{x},\mathbf{y}})}{\pi^2}, \tag{7.28}$$

as required for the statement of the lemma. □

Theorem 7.5. For any $\mathbf{x},\mathbf{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ such that $\mathbf{y}$ is not a scalar multiple of $\mathbf{x}$, the estimator $\widehat{K}_m^{\text{ang,ort}}(\mathbf{x},\mathbf{y})$ for $K^{\text{ang}}(\mathbf{x},\mathbf{y})$ is unbiased and satisfies:

$$\text{MSE}(\widehat{K}_m^{\text{ang,ort}}(\mathbf{x},\mathbf{y})) < \text{MSE}(\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})). \tag{7.12}$$

Proof. We have already shown unbiasedness. In addition to the notation introduced in the proof of Lemma 7.4, we write

$$S_i^{\text{ort}} = \text{sign}\left(\left\langle \mathbf{G}_{\text{ort}}^i, \mathbf{x}\right\rangle\right)\text{sign}\left(\left\langle \mathbf{G}_{\text{ort}}^i, \mathbf{y}\right\rangle\right), \tag{7.29}$$

for $i = 1,\ldots,m$. Since S_i^{ort} is equal in distribution to S_i, we have

$$\text{MSE}(\widehat{K}_m^{\text{ang,ort}}(\mathbf{x},\mathbf{y})) - \text{MSE}(\widehat{K}_m^{\text{ang,base}}(\mathbf{x},\mathbf{y})) = \frac{1}{m^2}\sum_{i\neq j}^{m}\text{Cov}\left(S_i^{\text{ort}}, S_j^{\text{ort}}\right) \tag{7.30}$$

Therefore, demonstrating the theorem is equivalent to showing, for $i \neq j$, that

$$\text{Cov}\left(S_i^{\text{ort}}, S_j^{\text{ort}}\right) < 0, \tag{7.31}$$

which is itself equivalent to showing

$$\mathbb{E}\left[S_i^{\text{ort}} S_j^{\text{ort}}\right] < \mathbb{E}\left[S_i^{\text{ort}}\right]\mathbb{E}\left[S_j^{\text{ort}}\right]. \tag{7.32}$$

Note that the variables $(S_i^{\text{ort}})_{i=1}^m$ take values in $\{\pm 1\}$. Denoting $\mathscr{A}_i = \{S_i^{\text{ort}} = -1\}$ for $i = 1,\ldots,m$, we can rewrite Expression (7.32) as

$$\mathbb{P}\left[\mathscr{A}_i^c \cap \mathscr{A}_j^c\right] + \mathbb{P}\left[\mathscr{A}_i \cap \mathscr{A}_j\right] - \mathbb{P}\left[\mathscr{A}_i \cap \mathscr{A}_j^c\right] - \mathbb{P}\left[\mathscr{A}_i^c \cap \mathscr{A}_j\right] < \left(\frac{\pi - 2\theta}{\pi}\right)^2. \tag{7.33}$$

Note that the left-hand side is equal to

$$2\left(\mathbb{P}\left[\mathscr{A}_i^c \cap \mathscr{A}_j^c\right] + \mathbb{P}\left[\mathscr{A}_i \cap \mathscr{A}_j\right]\right) - 1. \tag{7.34}$$

Plugging in the bounds of Proposition 7.16 below, and using the fact that the pair of indicators $(\mathbb{1}_{\mathscr{A}_i}, \mathbb{1}_{\mathscr{A}_j})$ is identically distributed for all pairs of distinct indices $i, j \in \{1, \dots, m\}$, thus yields the result. □

Proposition 7.16. We then have the following inequalities:

$$\mathbb{P}\left[\mathscr{A}_1 \cap \mathscr{A}_2\right] \leq \left(\frac{\theta}{\pi}\right)^2 \qquad \text{and} \qquad \mathbb{P}\left[\mathscr{A}_1^c \cap \mathscr{A}_2^c\right] \leq \left(1 - \frac{\theta}{\pi}\right)^2, \tag{7.35}$$

with both inequalities strict iff the angle between $\mathbf{x}$ and $\mathbf{y}$ is not 0 or π.

Before providing the proof of this proposition, we describe some coordinate choices we will make in order to make the calculations we require more straightforward.

We pick an orthonormal basis for $\mathbb{R}^d$ so that the first two coordinates span the $\mathbf{x}$-$\mathbf{y}$ plane, and further so that $(\mathbf{G}^1_{\text{ort}})_2$, the coordinate of $\mathbf{G}^1_{\text{ort}}$ in the second dimension, is 0. We extend this to an orthonormal basis of $\mathbb{R}^n$ so that $(\mathbf{G}^1_{\text{ort}})_3 \geq 0$, and $(\mathbf{G}^1_{\text{ort}})_i = 0$ for $i \geq 4$. Thus, in this basis, we have coordinates

$$\mathbf{G}^1_{\text{ort}} = ((\mathbf{G}^1_{\text{ort}})_1, 0, (\mathbf{G}^1_{\text{ort}})_3, 0, \dots, 0), \tag{7.36}$$

with $(\mathbf{G}^1_{\text{ort}})_1 \sim \chi_2$ and $(\mathbf{G}^1_{\text{ort}})_3 \sim \chi_{d-2}$ (by elementary calculations with multivariate Gaussian distributions). Note that the angle, ϕ, that $\mathbf{G}^1_{\text{ort}}$ makes with the $\mathbf{x}$-$\mathbf{y}$ plane is then $\phi = \arctan((\mathbf{G}^1_{\text{ort}})_3 / (\mathbf{G}^1_{\text{ort}})_1)$. Having fixed our coordinate system relative to the random variable $\mathbf{G}^1_{\text{ort}}$, the coordinates of $\mathbf{x}$ and $\mathbf{y}$ in this frame are now themselves random variables; we introduce the angle ψ to describe the angle between $\mathbf{x}$ and the positive first coordinate axis in this basis.

Now consider $\mathbf{G}^2_{\text{ort}}$. We are concerned with the direction of $((\mathbf{G}^2_{\text{ort}})_1, (\mathbf{G}^2_{\text{ort}})_2)$ in the $\mathbf{x}$-$\mathbf{y}$ plane. Conditional on $\mathbf{G}^1_{\text{ort}}$, the direction of the full vector $\mathbf{G}^2_{\text{ort}}$ is distributed uniformly on $S^{d-2}(\langle \mathbf{G}^1_{\text{ort}} \rangle^{\perp})$, the set of unit vectors orthogonal to $\mathbf{G}^1_{\text{ort}}$. Because of our particular choice of coordinates, we can therefore write

$$\mathbf{G}^2_{\text{ort}} = (-r\sin(\phi), (\mathbf{G}^2_{\text{ort}})_2, r\cos(\phi), (\mathbf{G}^2_{\text{ort}})_4, (\mathbf{G}^2_{\text{ort}})_5, \dots, (\mathbf{G}^2_{\text{ort}})_d), \tag{7.37}$$

where the $(d-1)$-dimensional vector $(r, (\mathbf{G}^2_{\text{ort}})_2, (\mathbf{G}^2_{\text{ort}})_4, (\mathbf{G}^2_{\text{ort}})_5, \dots, (\mathbf{G}^2_{\text{ort}})_d)$ has an isotropic distribution.

So the direction of $((\mathbf{G}^2_{\text{ort}})_1, (\mathbf{G}^2_{\text{ort}})_2)$ in the **x**-**y** plane follows an angular Gaussian distribution (Tyler, 1987), with covariance matrix

$$\begin{pmatrix} \sin^2(\phi) & 0 \\ 0 & 1 \end{pmatrix}. \tag{7.38}$$

With these geometrical considerations in place, we are ready to give the proof of Proposition 7.16.

Proof of Proposition 7.16. Dealing with the first inequality, we decompose the event as

$$\begin{aligned} \mathscr{A}_1 \cap \mathscr{A}_2 =& \{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\} \\ &\cup \{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle > 0\} \\ &\cup \{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle < 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\} \\ &\cup \{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle < 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle > 0\}. \end{aligned} \tag{7.39}$$

As the law of $(\mathbf{G}^1_{\text{ort}}, \mathbf{G}^2_{\text{ort}})$ is the same as that of $(\mathbf{G}^2_{\text{ort}}, \mathbf{G}^1_{\text{ort}})$ and that of $(-\mathbf{G}^1_{\text{ort}}, \mathbf{G}^2_{\text{ort}})$, it follows that all four events in the above expression have the same probability. The statement of the theorem is therefore equivalent to demonstrating the following inequality:

$$\mathbb{P}\left[\langle \mathbf{G}^1_{\text{ort}}, x\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\right] < \left(\frac{\theta}{2\pi}\right)^2. \tag{7.40}$$

We now proceed according to the coordinate choices described above. We first condition on the random angles ϕ and ψ to obtain

$$\begin{aligned} &\mathbb{P}\left[\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\right] \\ &= \int_0^{2\pi} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \mathbb{P}\left[\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0 \middle| \psi, \phi\right] \\ &= \int_0^{2\pi} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \mathbb{1}_{\{0\in[\psi-\pi/2, \psi-\pi/2+\theta]\}} \mathbb{P}\left[\langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0 \middle| \psi, \phi\right], \end{aligned} \tag{7.41}$$

where f is the density of the random angle ϕ. The final equality above follows as $\mathbf{G}^1_{\text{ort}}$ and $\mathbf{G}^2_{\text{ort}}$ are independent conditional on ψ and ϕ, and the event $\{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0\}$ is exactly the event $\{0 \in [\psi - \pi/2, \psi - \pi/2 + \theta]\}$ (by considering the geometry of the situation in the **x**-**y** plane). We can remove the indicator function from the integrand by adjusting

the limits of integration, obtaining

$$\begin{aligned}&\mathbb{P}\left[\left\langle \mathbf{G}^1_{\text{ort}},\mathbf{x}\right\rangle > 0, \left\langle \mathbf{G}^1_{\text{ort}},\mathbf{y}\right\rangle < 0, \left\langle \mathbf{G}^2_{\text{ort}},\mathbf{x}\right\rangle > 0, \left\langle \mathbf{G}^2_{\text{ort}},\mathbf{y}\right\rangle < 0\right]\\ &= \int_{\pi/2-\theta}^{\pi/2} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \mathbb{P}\left[\left\langle \mathbf{G}^2_{\text{ort}},\mathbf{x}\right\rangle > 0, \left\langle \mathbf{G}^2_{\text{ort}},\mathbf{y}\right\rangle < 0 | \psi,\phi\right]. \end{aligned} \tag{7.42}$$

We now turn our attention to the conditional probability

$$\mathbb{P}\left[\left\langle \mathbf{G}^2_{\text{ort}},\mathbf{x}\right\rangle > 0, \left\langle \mathbf{G}^2_{\text{ort}},\mathbf{y}\right\rangle < 0 | \psi,\phi\right]. \tag{7.43}$$

The event $\{\left\langle \mathbf{G}^2_{\text{ort}},\mathbf{x}\right\rangle > 0, \left\langle \mathbf{G}^2_{\text{ort}},\mathbf{y}\right\rangle < 0\}$ is equivalent to the angle t of the projection of $\mathbf{G}^2_{\text{ort}}$ into the $\mathbf{x}$-$\mathbf{y}$ plane with the first coordinate axis lying in the interval $[\psi-\pi/2, \psi-\pi/2+\theta]$. Recalling the distribution of the angle t from the geometric considerations described immediately before this proof, we obtain

$$\begin{aligned}&\mathbb{P}\left[\left\langle \mathbf{G}^1_{\text{ort}},\mathbf{x}\right\rangle > 0, \left\langle \mathbf{G}^1_{\text{ort}},\mathbf{y}\right\rangle < 0, \left\langle \mathbf{G}^2_{\text{ort}},\mathbf{x}\right\rangle > 0, \left\langle \mathbf{G}^2_{\text{ort}},\mathbf{y}\right\rangle < 0\right]\\ &= \int_{\pi/2-\theta}^{\pi/2} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi \int_{\psi-\pi/2}^{\psi-\pi/2+\theta} (2\pi\sin(\phi))^{-1}(\cos^2(t)/\sin^2(\phi)+\sin^2(t))^{-1}\mathrm{d}t. \end{aligned} \tag{7.44}$$

With $\theta \in [0,\pi/2]$, we note that the integral with respect to t can be evaluated analytically, leading us to

$$\begin{aligned}&\mathbb{P}\left[\left\langle \mathbf{G}^1_{\text{ort}},\mathbf{x}\right\rangle > 0, \left\langle \mathbf{G}^1_{\text{ort}},\mathbf{y}\right\rangle < 0, \left\langle \mathbf{G}^2_{\text{ort}},\mathbf{x}\right\rangle > 0, \left\langle \mathbf{G}^2_{\text{ort}},\mathbf{y}\right\rangle < 0\right]\\ &= \int_{\pi/2-\theta}^{\pi/2} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \frac{1}{2\pi}\left(\arctan(\tan(\psi-\pi/2+\theta)\sin(\phi)) - \arctan(\tan(\psi-\pi/2)\sin(\phi))\right)\\ &\leq \int_{\pi/2-\theta}^{\pi/2} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \frac{\theta}{2\pi}\\ &= \left(\frac{\theta}{2\pi}\right)^2, \end{aligned} \tag{7.45}$$

with the inequality being strict when $\theta \neq 0$. To deal with $\theta \in [\pi/2,\pi]$, we note that if the angle θ between $\mathbf{x}$ and $\mathbf{y}$ is obtuse, then the angle between $\mathbf{x}$ and $-\mathbf{y}$ is $\pi-\theta$ and therefore acute. Recalling from our definition that $\mathscr{A}_m = \{\text{sign}\left(\left\langle \mathbf{G}^i_{\text{ort}},\mathbf{x}\right\rangle\right)\text{sign}\left(\left\langle \mathbf{G}^i_{\text{ort}},\mathbf{y}\right\rangle\right) = -1\}$, if we denote the corresponding quantity for the pair of vectors $\mathbf{x}$, $-\mathbf{y}$ by $\bar{\mathscr{A}}_m = \{\text{sign}\left(\left\langle \mathbf{G}^i_{\text{ort}},\mathbf{x}\right\rangle\right)\text{sign}\left(\left\langle \mathbf{G}^i_{\text{ort}},-\mathbf{y}\right\rangle\right) = -1\}$, then we in fact have $\bar{\mathscr{A}}_m = \mathscr{A}_m^c$. Therefore, applying the result to the pair of vectors $\mathbf{x}$ and $-\mathbf{y}$ (which have acute angle $\pi-\theta$ between them) and

using the inclusion-exclusion principle, we obtain:

$$\begin{aligned}\mathbb{P}(\mathscr{A}_1 \cap \mathscr{A}_2) &= 1 - \mathbb{P}(\mathscr{A}_1^c) - \mathbb{P}(\mathscr{A}_2^c) + \mathbb{P}(\mathscr{A}_1^c \cap \mathscr{A}_2^c) \\ &\leq 1 - \mathbb{P}(\mathscr{A}_1^c) - \mathbb{P}(\mathscr{A}_2^c) + \left(\frac{\pi-\theta}{\pi}\right)^2 \\ &= 1 - 2\left(\frac{\pi-\theta}{\pi}\right) + \left(\frac{\pi-\theta}{\pi}\right)^2 \\ &= \left(\frac{\theta}{\pi}\right)^2, \end{aligned} \tag{7.46}$$

with the inequality strict when $\theta = \pi$, as required. The second inequality of Proposition 7.16 follows from the inclusion-exclusion principle and the first inequality:

$$\begin{aligned}\mathbb{P}\left[\mathscr{A}_1^c \cap \mathscr{A}_2^c\right] &= 1 - \mathbb{P}\left[\mathscr{A}_1\right] - \mathbb{P}\left[\mathscr{A}_2\right] + \mathbb{P}\left[\mathscr{A}_1 \cap \mathscr{A}_2\right] \\ &\leq 1 - \mathbb{P}\left[\mathscr{A}_1\right] - \mathbb{P}\left[\mathscr{A}_2\right] + \left(\frac{\theta}{\pi}\right)^2 \\ &= \left(1 - \mathbb{P}\left[\mathscr{A}_1\right]\right)\left(1 - \mathbb{P}\left[\mathscr{A}_2\right]\right) \\ &= \left(1 - \frac{\theta}{\pi}\right)^2, \end{aligned} \tag{7.47}$$

with the inequality strict when $\theta \notin \{0, \pi\}$. □

7.A.2 Proofs of results in Section 7.5.1

We begin by establishing the following result, which will be useful in the proofs of Theorems 7.8 and 7.9.

Proposition 7.17. The difference in mean squared error between the estimator $\widehat{K}_m^{\text{iid}}(\mathbf{x}, \mathbf{y})$ and $\widehat{K}_m^{\text{ort}}(\mathbf{x}, \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ is given by the following expression, in which we write $\mathbf{z} = \mathbf{x} - \mathbf{y}$ and $\widehat{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|$:

$$\begin{aligned}&\text{MSE}(\widehat{K}_m^{\text{ort}}(\mathbf{x}, \mathbf{y})) - \text{MSE}(\widehat{K}_m^{\text{iid}}(\mathbf{x}, \mathbf{y})) = \\ &\frac{m-1}{m}\left(\mathbb{E}\left[\cos(\sqrt{R_1^2 + R_2^2}\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right] - \mathbb{E}\left[\cos(R_1\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]^2\right), \end{aligned} \tag{7.48}$$

where $\mathbf{v} \sim \text{Unif}(S^{d-1})$, and $R_1, R_2 \overset{\text{i.i.d.}}{\sim} \rho_K$, where $\rho_K \in \mathscr{P}(\mathbb{R}_{\geq 0})$ is the distribution of the norm of a vector drawn from the Fourier distribution μ_K corresponding to K.

Proof. By independence of the $(\mathbf{w}_i)_{i=1}^m$ in the case of the estimator $\widehat{K}_m^{\text{iid}}(\mathbf{x},\mathbf{y})$, we have

$$\text{MSE}(\widehat{K}_m^{\text{iid}}(\mathbf{x},\mathbf{y})) = \frac{1}{m}\text{Var}(\cos(\langle \mathbf{w}_1, \mathbf{z}\rangle)). \tag{7.49}$$

Considering the analogous quantity for $\widehat{K}_m^{\text{ort}}(\mathbf{x},\mathbf{y})$, we note that it differs from the expression above by the sum of $m(m-1)$ equal covariance terms. The covariance term is of the form

$$\frac{1}{m^2}\left(\mathbb{E}\left[\cos(\langle \mathbf{w}_1^{\text{ort}}, \mathbf{z}\rangle)\cos(\langle \mathbf{w}_2^{\text{ort}}, \mathbf{z}\rangle)\right] - \mathbb{E}\left[\cos(\langle \mathbf{w}_1^{\text{ort}}, \mathbf{z}\rangle)\right]^2\right), \tag{7.50}$$

where $\mathbf{w}_1^{\text{ort}}, \mathbf{w}_2^{\text{ort}}$ are both marginally distributed according to μ, and are conditioned to be almost surely orthogonal. Consider the first term of Expression (7.50). We use the product-to-sum trigonometric identity $\cos(a)\cos(b) = \frac{1}{2}(\cos(a+b)+\cos(a-b))$ for all $a, b \in \mathbb{R}$ to obtain that

$$\mathbb{E}\left[\cos(\langle \mathbf{w}_1^{\text{ort}}, \mathbf{z}\rangle)\cos(\langle \mathbf{w}_2^{\text{ort}}, \mathbf{z}\rangle)\right] = \frac{1}{2}\left(\mathbb{E}\left[\cos(\langle \mathbf{w}_1^{\text{ort}} + \mathbf{w}_2^{\text{ort}}, \mathbf{z}\rangle)\right] + \mathbb{E}\left[\cos(\langle \mathbf{w}_1^{\text{ort}} - \mathbf{w}_2^{\text{ort}}, \mathbf{z}\rangle)\right]\right). \tag{7.51}$$

Note that $\mathbf{w}_1^{\text{ort}} + \mathbf{w}_2^{\text{ort}} \overset{d}{=} \mathbf{w}_1^{\text{ort}} - \mathbf{w}_2^{\text{ort}}$, so it is sufficient to deal with the first term in the final expression above. Since μ is rotationally invariant (as K is isotropic), each $\mathbf{w}$ drawn from μ can be decomposed as

$$\mathbf{w} = R\mathbf{v}, \tag{7.52}$$

where $\mathbf{v} \sim \text{Unif}(S^{d-1})$, and independently, R is a scalar random variable drawn from ρ_K. The key observation now is that we have a similar decomposition to Equation (7.52) for $\mathbf{w}_1^{\text{ort}} + \mathbf{w}_2^{\text{ort}}$; indeed, we have

$$\mathbf{w}_1^{\text{ort}} + \mathbf{w}_2^{\text{ort}} = \sqrt{R_1^2 + R_2^2}\,\mathbf{v}, \tag{7.53}$$

with $\mathbf{v} \sim \text{Unif}(S^{n-1})$, and independently, R_1 and R_2 are the norms of $\mathbf{w}_1^{\text{ort}}$ and $\mathbf{w}_2^{\text{ort}}$ respectively (the norm of the sum is given by this form due to almost-sure orthogonality and Pythagoras' theorem). The covariance term can therefore be written

$$\frac{1}{m^2}\left(\mathbb{E}\left[\cos\left(\sqrt{R_1^2 + R_2^2}\langle \mathbf{v}, \mathbf{z}\rangle\right)\right] - \mathbb{E}\left[\cos\left(R_1\langle \mathbf{v}, \mathbf{z}\rangle\right)\right]^2\right), \tag{7.54}$$

which completes the proof. □

We now turn our attention to the first main result of the chapter.

Theorem 7.8. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be an RBF kernel and let $\mu_K \in \mathscr{P}(\mathbb{R}^d)$ be its associated Fourier measure. Suppose that $\mathbb{E}_{\mathbf{w}\sim\mu_K}\left[\|\mathbf{w}\|^4\right] < \infty$. Then for sufficiently small $\|\mathbf{z}\|$, we have

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x},\mathbf{y})) > \mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y}))\,. \tag{7.17}$$

Proof. By Proposition 7.17, the statement of the theorem is equivalent to showing that the following term is negative:

$$\mathbb{E}\left[\cos\left(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\right)\right] - \mathbb{E}\left[\cos(R_1\|\mathbf{z}\|\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]^2. \tag{7.55}$$

We may regard this as a function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ of $\|\mathbf{z}\|$, noting that the value of the expectations does not depend on $\widehat{\mathbf{z}}$, as it appears only in the inner product with the random unit vector $\mathbf{v}$, which has an isotropic distribution. We will write $z = \|\mathbf{z}\|$ for the argument of f in what follows for convenience:

$$f(z) = \mathbb{E}\left[\cos\left(\sqrt{R_1^2+R_2^2}z\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\right)\right] - \mathbb{E}\left[\cos(R_1 z\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]^2, \qquad z \in \mathbb{R}_{\geq 0} \tag{7.56}$$

Observe trivially that $f(0) = 0$. We will show that f is decreasing in a neighbourhood around 0, from which the statement of the theorem immediately follows.

First observe that f is well defined on all of $\mathbb{R}_{\geq 0}$, since the expectations are of bounded, measurable functions of random variables. A priori, it is not clear that f is differentiable, but we will see that by the dominated convergence theorem, if the random variable R has a finite k^{th} moment, for some $k \in \mathbb{N}$, then f is k times differentiable everywhere, and moreover, the k^{th} derivative is continuous. Specifically, recall the following corollary of the dominated convergence theorem.

Proposition 7.18. Let $\mu \in \mathscr{P}(\mathbb{R})$ be a Borel probability measure, and let $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be such that:

- $x \mapsto g(t,x)$ is in $L^1(\mu)$ for all t;
- $t \mapsto g(t,x)$ is differentiable for all x;
- For some function $h \in L^1(\mu)$, we have

$$\left|\frac{\partial g}{\partial t}(t,x)\right| \leq h(x) \qquad \forall\, t,x \in \mathbb{R}. \tag{7.57}$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{X\sim\mu}\left[g(t,X)\right]=\mathbb{E}\left[\frac{\partial g}{\partial t}(t,X)\right]. \tag{7.58}$$

By the assumption that R has a finite 4^{th} moment, we have $R,R^2,R^3,R^4\in L^1(\mu)$, and we may therefore use these as the dominating functions in Proposition 7.18 to establish fourth-order differentiability of the expectation $\mathbb{E}\left[\cos(R_1 z\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]$. We note further that since $\sqrt{R_1^2+R_2^2}\leq R_1+R_2$ almost-surely, we may use $(R_1+R_2)^k$ for $k=1,\ldots,4$ as dominating functions in Proposition 7.18 to establish the fourth-order differentiability of the expectation $\mathbb{E}\left[\cos\left(\sqrt{R_1^2+R_2^2}z\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\right)\right]$. We therefore derive that f is 4 times differentiable, and we obtain the following (by Taylor's theorem with Lagrange's remainder):

$$f(h)=f(0)+hf^{(1)}(0)+\frac{h^2}{2!}f^{(2)}(0)+\frac{h^3}{3!}f^{(3)}(0)+\frac{h^4}{4!}f^{(4)}(s). \tag{7.59}$$

Direct computation leveraging Proposition 7.18 yields

$$\begin{aligned}
f^{(1)}(z)=&-\mathbb{E}\left[\sqrt{R_1^2+R_2^2}\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\sin\left(z\sqrt{R_1^2+R_2^2}\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\right)\right]+\\
&2\mathbb{E}\left[\cos(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\sin(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right],\\
f^{(2)}(z)=&-\mathbb{E}\left[(R_1^2+R_2^2)\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^2\cos\left(z\sqrt{R_1^2+R_2^2}\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\right)\right]-2\mathbb{E}\left[R_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\sin(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]^2\\
&+2\mathbb{E}\left[\cos(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^2\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^2\cos(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right],\\
f^{(3)}(z)=&\mathbb{E}\left[(R_1^2+R_2^2)^{3/2}\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^3\sin\left(z\sqrt{R_1^2+R_2^2}\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\right)\right]\\
&-4\mathbb{E}\left[R_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\sin(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^2\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^2\cos(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\\
&-2\mathbb{E}\left[R_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\sin(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^2\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^2\cos(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\\
&-2\mathbb{E}\left[\cos(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^3\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^3\sin(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right],\\
f^{(4)}(z)=&\mathbb{E}\left[(R_1^2+R_2^2)^2\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^4\cos\left(z\sqrt{R_1^2+R_2^2}\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\right)\right]\\
&-6\mathbb{E}\left[R_1^2\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^2\cos(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]^2\\
&+6\mathbb{E}\left[R_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\sin(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^3\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^3\sin(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\\
&+2\mathbb{E}\left[R_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\sin(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^3\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^3\sin(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\\
&-2\mathbb{E}\left[\cos(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^4\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^4\cos(zR_1\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right].
\end{aligned} \tag{7.60}$$

Directly substituing $z = 0$ into these expressions, we obtain

$$f(0) = f'(0) = f^{(2)}(0) = f^{(3)}(0) = 0\,, \qquad f^{(4)}(0) = \mathbb{E}\left[R_1^2\right]^2 \left(2\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^4\right] - 6\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2\right]^2\right). \quad (7.61)$$

To establish the sign of $f^{(4)}(0)$, we compute the expectations $\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^4\right]$, $\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2\right]$ directly. Firstly, note that $\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle$ can be written $\cos(\theta)$, where θ is the angle a uniformly random direction makes with a fixed direction in $\mathbb{R}^d$. By considering hyperspherical coordinates, the density of the angle on the interval $[0, \pi]$ is deduced to be

$$\frac{\sin^{d-2}(\theta)}{\int_0^\pi \sin^{d-2}(\theta')\mathrm{d}\theta'}\,. \quad (7.62)$$

Therefore, we have

$$\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2\right] = \frac{\int_0^\pi \cos^2(\theta)\sin^{d-2}(\theta)\mathrm{d}\theta}{\int_0^\pi \sin^{d-2}(\theta)\mathrm{d}\theta} = \frac{\sqrt{\pi}\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})} - \sqrt{\pi}\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2}+1)}}{\sqrt{\pi}\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})}} = \frac{1}{d}\,, \quad (7.63)$$

$$\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^4\right] = \frac{\int_0^\pi \cos^4(\theta)\sin^{n-2}(\theta)\mathrm{d}\theta}{\int_0^\pi \sin^{d-2}(\theta)\mathrm{d}\theta} = \frac{\sqrt{\pi}\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})} - 2\sqrt{\pi}\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2}+1)} + \sqrt{\pi}\frac{\Gamma(\frac{d+3}{2})}{\Gamma(\frac{d}{2}+2)}}{\sqrt{\pi}\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})}} = \frac{3}{d(d+2)}\,, \quad (7.64)$$

which yields $f^{(4)}(0) = 6\mathbb{E}\left[R_1^2\right]^2 (\frac{1}{d(d+2)} - \frac{1}{d^2}) < 0$.

Finally, again by applying the dominated convergence theorem to each expectation in the expression above for $f^{(4)}(z)$, we obtain that this function is continuous. Hence, we have:

$$f(h) = \frac{h^4}{4!} f^{(4)}(s)\,, \quad (7.65)$$

for some $s \in (0, h)$, and by continuity of $f^{(4)}$, for sufficiently small h, the right-hand side above is negative, completing the proof. □

7.A.3 Proof of Theorem 7.9

Theorem 7.9. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a stationary, isotropic, normalised kernel, defined by $K(\mathbf{x}, \mathbf{y}) = \phi(\|\mathbf{x} - \mathbf{y}\|)$, with the property that this defines a valid kernel for any $d \in \mathbb{N}$. For each $d \in \mathbb{N}$, let $\mu_d \in \mathscr{P}(\mathbb{R}^d)$ be the Fourier distribution associated with $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

Let $z \in \mathbb{R}_{\geq 0}$ be fixed, and let $\mathbf{x}^{(d)}, \mathbf{y}^{(d)} \in \mathbb{R}^d$ be any sequence of vectors with $\|\mathbf{x}^{(d)} - \mathbf{y}^{(d)}\| = z$ for all d. Suppose that the following two conditions hold:

1. The MGFs M_d of μ_d satisfy $M_d(\mathbf{x}^{(d)} - \mathbf{y}^{(d)}) = o(\sqrt{d})$.
2. If $\mathbf{w}_d^{(d)} \sim \mu_d$, then $\|\mathbf{w}^{(d)}\|_2^2 / \mathbb{E}\left[\|\mathbf{w}^{(d)}\|_2^2\right]$ converges in probability to the constant 1 as $d \to \infty$.

Then we have the following limiting expressions for the difference in MSEs between i.i.d. and orthogonal random feature kernel estimation:

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x}^{(d)}, \mathbf{y}^{(d)})) - \mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x}^{(d)}, \mathbf{y}^{(d)})) = \frac{m-1}{m}\left(\frac{1}{8d}\Psi_K(\mathbf{x}^{(d)} - \mathbf{y}^{(d)}) + o(d^{-1})\right). \quad (7.18)$$

Proof. To reduce notational clutter, we will drop dependence on the dimension parameter d where it does not cause confusion.

Trigonometric manipulations. We begin with the observation of Proposition 7.17:

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x}, \mathbf{y})) - \mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x}, \mathbf{y})) = \quad (7.66)$$
$$\frac{m-1}{m}\left(\mathbb{E}\left[\cos(\sqrt{R_1^2 + R_2^2}\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right] - \mathbb{E}\left[\cos(R_1\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]^2\right).$$

By writing the second term as $\mathbb{E}[\cos(R_1\|\mathbf{z}\|\langle \mathbf{v}_1, \widehat{\mathbf{z}}\rangle)]\,\mathbb{E}[\cos(R_2\|\mathbf{z}\|\langle \mathbf{v}_2, \widehat{\mathbf{z}}\rangle)]$, with $\mathbf{v}_1, \mathbf{v}_2 \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(S^{d-1})$ and independently $R_1, R_2 \overset{\text{i.i.d.}}{\sim} \rho_d$, where $\rho_d \in \mathscr{P}(\mathbb{R}_{\geq 0})$ is the distribution of the length of a vector drawn from μ_d, we may use the product-to-sum cosine identity to rewrite the term as

$$\mathbb{E}\left[\frac{1}{2}\cos(\langle R_1\mathbf{v}_1 + R_2\mathbf{v}_2, \mathbf{z}\rangle) + \frac{1}{2}\cos(\langle R_1\mathbf{v}_1 - R_2\mathbf{v}_2, \mathbf{z}\rangle)\right]. \quad (7.67)$$

Now note that $R_1\mathbf{v}_1 + R_2\mathbf{v}_2$ and $R_1\mathbf{v}_1 - R_2\mathbf{v}_2$ are both equal in distribution to $\sqrt{R_1^2 + R_2^2 + 2R_1R_2\langle \mathbf{v}_1, \mathbf{v}_2\rangle}\mathbf{v}$, with $\mathbf{v} \sim \mathrm{Unif}(S^{d-1})$ independent of $R_1, R_2, \mathbf{v}_1, \mathbf{v}_2$. Introducing the notation $u = \langle \mathbf{v}_1, \mathbf{v}_2\rangle$, we may rewrite the right-hand side of Equation (7.66) (without the $m^{-1}(m-1)$ factor) as

$$\left(\mathbb{E}\left[\cos\left(\sqrt{R_1^2 + R_2^2}\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle\right)\right] - \mathbb{E}\left[\cos\left(\sqrt{R_1^2 + R_2^2 + 2R_1R_2u}\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle\right)\right]\right). \quad (7.68)$$

Power series expansion. The next stage of the argument uses the bound on moment generating functions assumed in the theorem, and the dominated convergence theorem

to manipulate power series expansions of the cosine functions appearing above. More specifically, we note that

$$\cos\left(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|\langle\mathbf{v},\mathbf{z}\rangle\right)=\sum_{k=0}^{\infty}\frac{(-1)^k}{(2k)!}\left(R_1^2+R_2^2\right)^k\langle\mathbf{v},\mathbf{z}\rangle^{2k}. \tag{7.69}$$

With the aim of applying the dominated convergence theorem to argue that we may exchange expectation with the infinite sum above, we observe that the following random variable dominates all partial sums

$$\sum_{k=0}^{\infty}\frac{1}{(2k)!}\left(R_1^2+R_2^2\right)^k\langle\mathbf{v},\mathbf{z}\rangle^{2k}. \tag{7.70}$$

This is dominated by $\exp\left(\sqrt{R_1^2+R_2^2}\langle\mathbf{v},\mathbf{z}\rangle\right)$, which is integrable by the assumptions on the MGFs laid out in the theorem statement. Thus, by the dominated convergence theorem, we may rewrite Expression (7.68) as

$$\sum_{k=0}^{\infty}\frac{(-1)^k}{(2k)!}\mathbb{E}\left[\langle\mathbf{v},\mathbf{z}\rangle^{2k}\right]\left(\mathbb{E}\left[(R_1^2+R_2^2)^k\right]-\mathbb{E}\left[(R_1^2+R_2^2+2R_1R_2u)^k\right]\right). \tag{7.71}$$

Note that u is independent of R_1,R_2, and as u is the inner product of two independent Unif(S^{d-1}) random vectors, we have $\mathbb{E}\left[u^{2l+1}\right]=0$ for all $l\in\mathbb{N}$, $\mathbb{E}\left[u^2\right]=d^{-1}$ and $\mathbb{E}\left[u^{2l}\right]=O(d^{-l})$ for each $l\in\mathbb{N}$. Thus, performing a binomial expansion of the right-most term in Expression (7.71), we obtain

$$\sum_{k=0}^{\infty}\frac{(-1)^k}{(2k)!}\mathbb{E}\left[\langle\mathbf{v},\mathbf{z}\rangle^{2k}\right]\left(-d^{-1}\binom{k}{2}\mathbb{E}\left[(R_1^2+R_2^2)^{k-2}(2R_1R_2)^2\right]-\right.$$
$$\left.O(d^{-2})\sum_{\substack{j\geq 4\\ \text{even}}}^{k}\binom{k}{j}\mathbb{E}\left[(R_1^2+R_2^2)^{k-j}(2R_1R_2)^j\right]\right). \tag{7.72}$$

Dealing with the second term. We now show that of the two terms that emerge by multiplying out the brackets in Expression (7.72), the right term decays to 0 faster than $\Theta(d^{-1})$. Note that it suffices to show that

$$\sum_{k=0}^{\infty}\frac{(-1)^k}{(2k)!}\mathbb{E}\left[\langle\mathbf{v},\mathbf{z}\rangle^{2k}\right]\sum_{\substack{j\geq 4\\ \text{even}}}^{k}\binom{k}{j}\mathbb{E}\left[(R_1^2+R_2^2)^{k-j}(2R_1R_2)^j\right]=o(d). \tag{7.73}$$

To show this, we observe that by applying the triangle inequality to the sum over k and the inequality $2R_1R_2 \leq R_1^2 + R_2^2$, we obtain

$$\left| \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} \mathbb{E}\left[\langle \mathbf{v}, \mathbf{z} \rangle^{2k} \right] \sum_{\substack{j \geq 4 \\ \text{even}}}^{k} \binom{k}{j} \mathbb{E}\left[(R_1^2 + R_2^2)^{k-j} (2R_1R_2)^j \right] \right| \leq \sum_{k=0}^{\infty} \frac{1}{(2k)!} 2^k \mathbb{E}\left[\left(\sqrt{R_1^2 + R_2^2} \langle \mathbf{v}, \mathbf{z} \rangle \right)^{2k} \right]. \tag{7.74}$$

With an aim of applying the MGF bound in the assumptions of the theorem, we next observe that by convexity of $x \mapsto x^k$ for each $k \in \mathbb{N} \cup \{0\}$ and Jensen's inequality, we have $\mathbb{E}\left[(R_1^2 + R_2^2 + 2R_1R_2u)^k \right] \geq \mathbb{E}\left[(R_1^2 + R_2^2)^k \right]$. Substituting this bound into the right-hand side of Inequality (7.74), we obtain the following upper bound for the left-hand side of (7.74):

$$\sum_{k=0}^{\infty} \frac{1}{(2k)!} \mathbb{E}\left[2^k \left(R_1^2 + R_2^2 + 2R_1R_2u \right)^k \langle \mathbf{v}, \mathbf{z} \rangle^{2k} \right] = \mathbb{E}\left[\exp\left(\langle R_1\mathbf{v}_1 + R_2\mathbf{v}_2, \sqrt{2}\mathbf{z} \rangle \right) \right] = M_d(\sqrt{2}\mathbf{z})^2, \tag{7.75}$$

which by the assumptions of the theorem is $o(d)$, meaning that overall, the second term of Expression (7.72) goes as $o(d^{-1})$, as we set out to show.

Dealing with the first term. Now, since $R_1^2/\mathbb{E}[R_1^2]$ converges in probability to the constant 1 as $d \to \infty$, there exist functions $\varepsilon, \delta : \mathbb{N} \to \mathbb{R}_{\geq 0}$ such that $\varepsilon, \delta = o(1)$, and which satisfy $\mathbb{P}(|R_1^2/\mathbb{E}[R_1^2] - 1| > \varepsilon(d)) < \delta(d)$. We thus observe that

$$\begin{aligned}
&\left| \mathbb{E}\left[(R_1^2 + R_2^2)^{k-2} (2R_1R_2)^2 \right] - \mathbb{E}\left[(R_1^2 + R_2^2)^k \right] \right| \\
=&\mathbb{E}\left[(R_1^2 + R_2^2)^{k-2} (R_1^2 - R_2^2)^2 \right] \\
=&\delta(d)\mathbb{E}\left[(R_1^2 + R_2^2)^k \right] + (1 - \delta(d))(2\varepsilon(d)\mathbb{E}[R_1^2])^2 \mathbb{E}\left[(R_1^2 + R_2^2)^{k-2} \right] \\
=&o(1)\mathbb{E}\left[(R_1^2 + R_2^2)^k \right].
\end{aligned} \tag{7.76}$$

Thus, the first term of Expression (7.72) may be written

$$\begin{aligned}
&d^{-1} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} \binom{k}{2} \mathbb{E}\left[\langle \mathbf{v}, \mathbf{z} \rangle^{2k} \right] \mathbb{E}\left[(R_1^2 + R_2^2)^k \right] (1 + o(1)) \\
=&d^{-1} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} \binom{k}{2} \mathbb{E}\left[\langle \mathbf{v}, \mathbf{z} \rangle^{2k} \right] \mathbb{E}\left[(R_1^2 + R_2^2)^k \right] + o(d^{-1}).
\end{aligned} \tag{7.77}$$

By calculations mimicking those above, we have $\mathbb{E}\left[(R_1^2+R_2^2)^k\right] = \mathbb{E}\left[(R_1^2+R_2^2+2R_1R_2u)^k\right](1+2^k o(1))$, so that overall, we have

$$\begin{aligned}&\mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x},\mathbf{y}))\\ &\quad= \frac{m-1}{m}\left(-d^{-1}\sum_{k=0}^{\infty}\frac{(-1)^k}{(2k)!}\binom{k}{2}\mathbb{E}\left[(R_1^2+R_2^2+2R_1R_2u)^k\langle\mathbf{v},\widehat{\mathbf{z}}\rangle\|\mathbf{z}\|^{2k}\right]+o(d^{-1})\right). \quad (7.78)\end{aligned}$$

Recovering the charm function. We observe that $\binom{k}{2} = \frac{2k(2k-1)-2k}{8}$. Using this observation, and swapping the order of differentiation, integration and summation (by the dominated convergence theorem again), and factoring the power series back into cosines, we thus obtain

$$\begin{aligned}&\mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}_m^{\mathrm{iid}}(\mathbf{x},\mathbf{y}))\\ &\quad= \frac{m-1}{m}\left(-\frac{d^{-1}}{8}\left(\|\mathbf{z}\|^2\frac{\mathrm{d}^2}{\mathrm{d}x^2}\phi_K^2(x)\Big|_{x=\|\mathbf{z}\|} - \|\mathbf{z}\|\frac{\mathrm{d}}{\mathrm{d}x}\phi_K^2(x)\Big|_{x=\|\mathbf{z}\|}\right)+o(d^{-1})\right)\\ &\quad= \frac{m-1}{m}\left(-\frac{d^{-1}}{8}\Psi_K(\mathbf{z})+o(d^{-1})\right), \quad (7.79)\end{aligned}$$

as required. □

7.A.4 Proof of Theorem 7.11

Theorem 7.11. Let $\phi:\mathbb{R}\to\mathbb{R}$ be such that for every $d\in\mathbb{N}$, $K_d:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ defined by $K_d(\mathbf{x},\mathbf{y}) = \phi(\|\mathbf{x}-\mathbf{y}\|)$ is a positive definite kernel. Then for each such K_d, the charm function Ψ_{K_d} is non-negative away from $\mathbf{0}$.

Proof. Note that by Schoenberg's characterisation of positive-definite kernels (Theorem 7.13) we know that $\phi(z) = \xi(z^2)$ for some completely monotone function ξ. Thus we obtain

$$\frac{\mathrm{d}\phi^2(x)}{\mathrm{d}x} = 4x\xi(x^2)\frac{\mathrm{d}\xi(y)}{\mathrm{d}y}\Big|_{y=x^2}, \quad (7.80)$$

and

$$\frac{\mathrm{d}^2\phi^2(x)}{\mathrm{d}x^2} = 4\left[\xi(x^2)\frac{\mathrm{d}\xi(y)}{\mathrm{d}y}\Big|_{y=x^2} + 2x^2\left(\frac{\mathrm{d}\xi(y)}{\mathrm{d}y}\Big|_{y=x^2}\right)^2 + 2x^2\xi(x^2)\frac{\mathrm{d}^2\xi(y)}{\mathrm{d}y^2}\Big|_{y=x^2}\right]. \quad (7.81)$$

Therefore we obtain

$$\Psi_K(z) = 8z^4 \left[\left(\frac{\mathrm{d}\xi(y)}{\mathrm{d}y}\Big|_{y=z^2} \right)^2 + \xi(z^2) \frac{\mathrm{d}^2\xi(y)}{\mathrm{d}y^2}\Big|_{y=z^2} \right]. \tag{7.82}$$

This completes the proof, since every non-trivial completely monotone function is non-negative and strictly convex. □

7.A.5 Proof of Proposition 7.15

Proposition 7.15. For an RBF kernel K on $\mathbb{R}^d$ with Fourier measure μ_K and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, writing $\mathbf{z} = \mathbf{x} - \mathbf{y}$, we have:

$$\begin{aligned} &\mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_m(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_m(\mathbf{x},\mathbf{y})) \qquad (7.20)\\ &= \frac{m-1}{m}\mathbb{E}_{R_1,R_2}\left[\frac{J_{\frac{d}{2}-1}\left(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|\right)\Gamma(d/2)}{\left(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|/2\right)^{\frac{d}{2}-1}} \right] - \frac{m-1}{m}\mathbb{E}_{R_1}\left[\frac{J_{\frac{d}{2}-1}\left(R_1\|\mathbf{z}\|\right)\Gamma(d/2)}{\left(R_1\|\mathbf{z}\|/2\right)^{\frac{d}{2}-1}} \right]^2, \end{aligned}$$

where R_1, R_2 are distributed i.i.d. from ρ_K, and J_α is the Bessel function of the first kind of degree α.

Proof. This result follows first by recalling the result of Proposition 7.17, namely that:

$$\begin{aligned} &\mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_m(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_m(\mathbf{x},\mathbf{y})) = \qquad (7.83)\\ &\qquad \frac{m-1}{m}\left(\mathbb{E}\left[\cos(R_1\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]^2 - \mathbb{E}\left[\cos\left(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|\,\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle \right) \right] \right). \end{aligned}$$

Note that both expectations appearing in the expression above have the form $\mathbb{E}[\cos(A\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)]$, for some non-negative scalar random variable A. We rewrite this as a nested conditional expectation over the uniform direction, as $\mathbb{E}[\mathbb{E}[\cos(A\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)|A]]$. Next, we recall from the proof of Proposition 7.17 that the random variable $\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle$ may be written $\cos(\theta)$, for a random angle θ distributed on $[0,\pi]$ with density

$$\frac{\sin^{d-2}(\theta)}{\int_0^\pi \sin^{d-2}(\theta')\mathrm{d}\theta'}. \tag{7.84}$$

Therefore, we can write:

$$\mathbb{E}\left[\mathbb{E}\left[\cos(A\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)|A\right]\right] = \mathbb{E}_A\left[\frac{\int_0^\pi \cos(A\cos(\theta))\sin^{d-2}(\theta)\mathrm{d}\theta}{\int_0^\pi \sin^{d-2}(\theta')\mathrm{d}\theta'}\right]. \tag{7.85}$$

For the integral in the denominator of the fraction, we recall that

$$\int_0^\pi \sin^{d-2}(\theta)\mathrm{d}\theta = \frac{\pi\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})}. \tag{7.86}$$

For the integral in the numerator, we use the Poisson-Bessel identity:

$$J_\nu(w) = \left(\frac{w}{2}\right)^\nu \frac{2}{\sqrt{\pi}\Gamma(\nu-\frac{1}{2})}\int_0^{\pi/2}\cos(w\cos(t))\sin^{2\nu}(t)\mathrm{d}t. \tag{7.87}$$

Substituting these expressions into Equation (7.83) yields the statement of the proposition. □

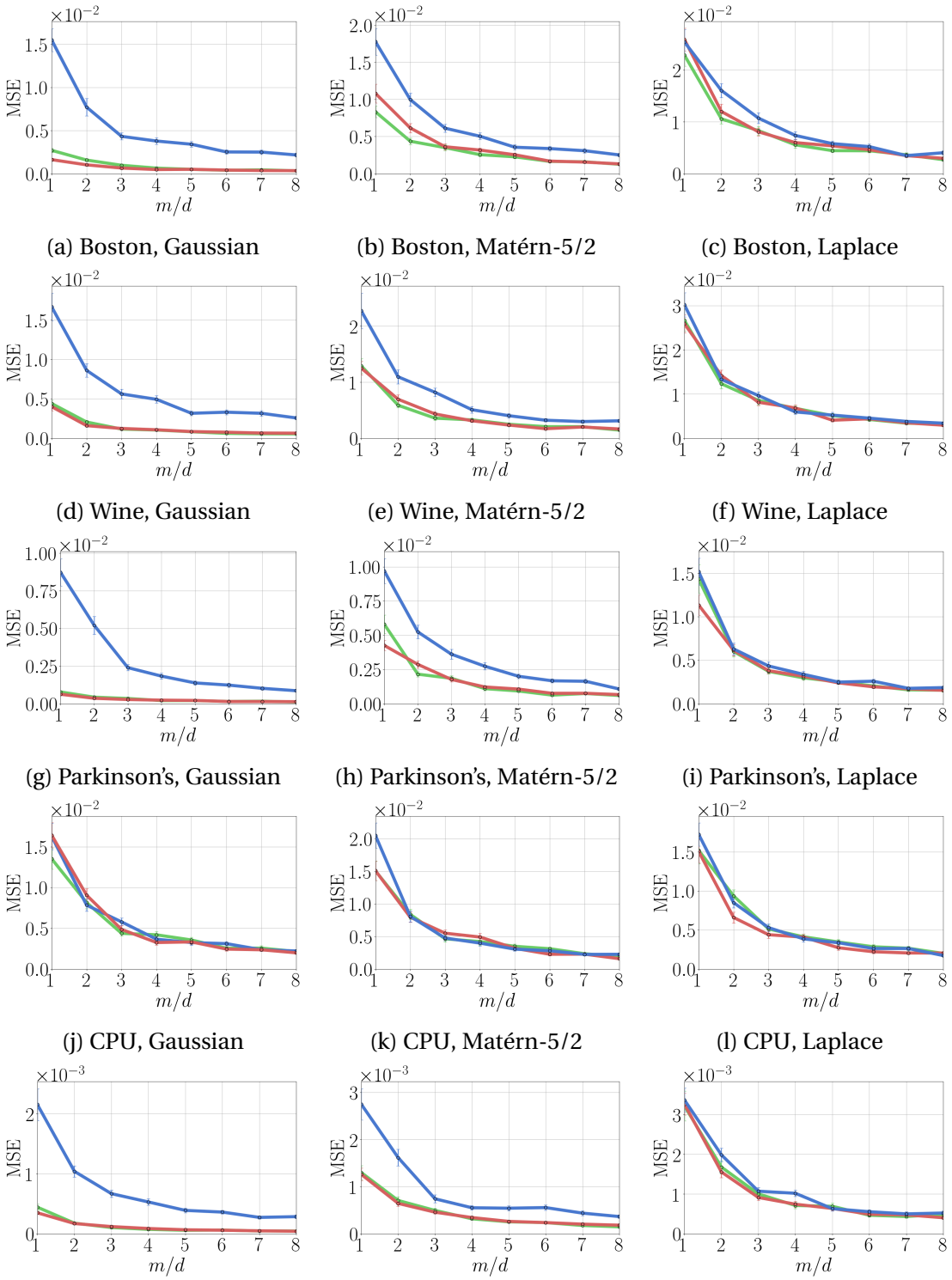
Appendix 7.B Additional experimental results

7.B.1 Pointwise kernel and Gram matrix estimation

Here, we provide results for the pointwise kernel and Gram matrix estimation experiments described in Section 7.6.1 for a larger range of UCI regression datasets. The results for pointwise estimation are given in Figure 7.6, and the results for Gram matrix estimation are displayed in Figure 7.7; see the caption for full details. We broadly observe similar qualitative behaviour as described in Section 7.6.

7.B.2 Gaussian process regression experiments

In this section we give full results for the Gaussian process regression experiments described in Section 7.6.2 for several other UCI regression datasets. We report KL divergence against predictions obtained from exact inference (i.e. GP regression without random feature approximation), RMSE prediction error, and wall-clock runtimes; we report the mean and (a bootstrapped estimate of the standard error of this estimate in parentheses) of each of these quantities across 10 runs of the experiment. Experiments were run on a cluster without full control of other processes running on the cluster; timing results



(a) Boston, Gaussian (b) Boston, Matérn-5/2 (c) Boston, Laplace

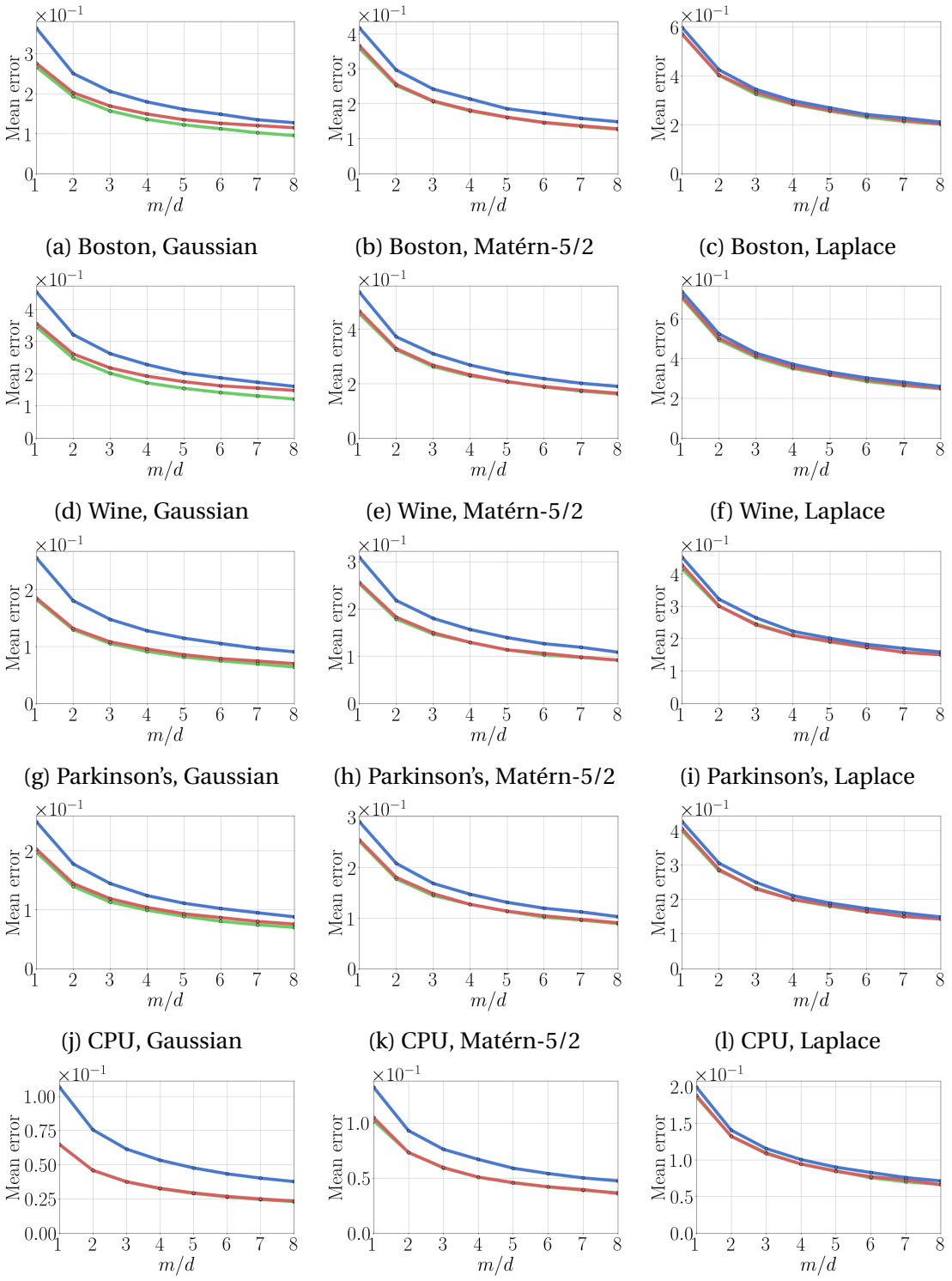
(d) Wine, Gaussian (e) Wine, Matérn-5/2 (f) Wine, Laplace

(g) Parkinson's, Gaussian (h) Parkinson's, Matérn-5/2 (i) Parkinson's, Laplace

(j) CPU, Gaussian (k) CPU, Matérn-5/2 (l) CPU, Laplace

(m) Insurance Company, Gaussian (n) Insurance Company, Matérn-5/2 (o) Insurance Company, Laplace

Figure 7.6. MSE for pointwise kernel estimation for a variety of UCI datasets and kernels. Two randomly selected datapoints from each dataset are chosen, and the kernel evaluated at these points is estimated. Estimators are i.i.d. random features (blue), orthogonal random features (green) and approximate Hadamard–Rademacher random features (red). In several plots, the red and green curves lie on top of one another.



(a) Boston, Gaussian (b) Boston, Matérn-5/2 (c) Boston, Laplace

(d) Wine, Gaussian (e) Wine, Matérn-5/2 (f) Wine, Laplace

(g) Parkinson's, Gaussian (h) Parkinson's, Matérn-5/2 (i) Parkinson's, Laplace

(j) CPU, Gaussian (k) CPU, Matérn-5/2 (l) CPU, Laplace

(m) Insurance Company, Gaussian (n) Insurance Company, Matérn-5/2 (o) Insurance Company, Laplace

Figure 7.7. Normalised Frobenius norm error for Gram matrix estimation for a variety of UCI datasets and kernels. Estimators are i.i.d. random features (blue), orthogonal random features (green) and approximate Hadamard–Rademacher random features (red). In several plots, the red and green curves lie on top of one another.

Kernel	Features	$m/d = 1$	$m/d = 2$	$m/d = 3$	$m/d = 4$	$m/d = 5$	$m/d = 6$	$m/d = 7$
Gaussian	IID	10640.0 (360.0)	3541.0 (130.0)	**1723.0 (44.0)**	1111.0 (21.0)	770.6 (22.0)	575.0 (14.0)	456.5 (12.0)
	ORF	**10620.0 (310.0)**	**3252.0 (90.0)**	1778.0 (40.0)	**1078.0 (23.0)**	**737.1 (18.0)**	**551.9 (8.5)**	**421.6 (5.1)**
	SORF	11200.0 (470.0)	3386.0 (120.0)	1801.0 (58.0)	1153.0 (43.0)	805.3 (13.0)	598.7 (14.0)	479.8 (12.0)
Matérn-5/2	IID	14100.0 (500.0)	4806.0 (130.0)	2574.0 (38.0)	1784.0 (41.0)	1276.0 (15.0)	979.7 (19.0)	791.1 (18.0)
	ORF	**13750.0 (400.0)**	4653.0 (120.0)	2578.0 (48.0)	**1684.0 (20.0)**	1227.0 (21.0)	**946.1 (13.0)**	**773.0 (13.0)**
	SORF	14460.0 (580.0)	**4649.0 (82.0)**	**2482.0 (62.0)**	1693.0 (21.0)	**1202.0 (18.0)**	954.4 (14.0)	773.4 (7.8)
Laplace	IID	32220.0 (820.0)	12710.0 (150.0)	8092.0 (110.0)	**5571.0 (64.0)**	4354.0 (55.0)	3593.0 (53.0)	3069.0 (52.0)
	ORF	31460.0 (710.0)	**12530.0 (140.0)**	**7861.0 (63.0)**	5720.0 (58.0)	4400.0 (54.0)	**3581.0 (53.0)**	2991.0 (36.0)
	SORF	**31170.0 (700.0)**	12550.0 (100.0)	8001.0 (160.0)	5735.0 (40.0)	**4317.0 (28.0)**	3593.0 (41.0)	**2983.0 (53.0)**

Table 7.4. Approximate GP regression results on cpu dataset. Reported numbers are average KL divergence from true posterior, along with bootstrap estimates of standard error (in parentheses).

Kernel	Features	$m/d = 1$	$m/d = 2$	$m/d = 3$	$m/d = 4$	$m/d = 5$	$m/d = 6$	$m/d = 7$
Gaussian	IID	**0.69 (0.02)**	0.57 (0.01)	0.51 (0.01)	0.48 (0.007)	0.44 (0.007)	**0.43 (0.007)**	**0.41 (0.008)**
	ORF	0.7 (0.01)	**0.56 (0.02)**	**0.5 (0.01)**	**0.47 (0.007)**	**0.44 (0.009)**	0.43 (0.006)	0.42 (0.006)
	SORF	0.72 (0.01)	0.59 (0.01)	0.52 (0.01)	0.5 (0.01)	0.47 (0.007)	0.45 (0.007)	0.43 (0.005)
Matérn-5/2	IID	0.7 (0.01)	0.57 (0.02)	0.49 (0.008)	0.45 (0.008)	0.44 (0.005)	0.4 (0.008)	0.4 (0.006)
	ORF	**0.7 (0.01)**	**0.54 (0.01)**	0.47 (0.009)	**0.43 (0.009)**	0.43 (0.01)	0.4 (0.008)	0.38 (0.01)
	SORF	0.71 (0.03)	0.54 (0.01)	**0.47 (0.01)**	0.44 (0.006)	**0.39 (0.009)**	**0.4 (0.007)**	**0.38 (0.01)**
Laplace	IID	0.76 (0.03)	0.56 (0.01)	**0.48 (0.01)**	0.44 (0.008)	0.42 (0.01)	**0.36 (0.007)**	0.36 (0.008)
	ORF	0.74 (0.01)	0.53 (0.02)	0.48 (0.01)	**0.41 (0.01)**	**0.4 (0.01)**	0.37 (0.007)	**0.35 (0.008)**
	SORF	**0.7 (0.02)**	**0.52 (0.01)**	0.48 (0.01)	0.44 (0.01)	0.42 (0.01)	0.38 (0.01)	0.36 (0.008)

Table 7.5. Approximate GP regression results on cpu dataset. Reported numbers are average test RMSE, along with bootstrap estimates of standard error (in parentheses).

should therefore be interpreted cautiously. We emphasise also that a fully-optimised fast Hadamard transform was not used in these experiments, and that the runtime of SORF methods may therefore be an underestimate of the achievable runtimes for these methods.

We observe that the structured methods, ORF and SORF, typically outperform on KL measures. On RMSE, there is little consistent advantage. On timing, we expect that the fast Hadamard transform for rapidly computing matrix multiplications will enable SORF to perform best when dimensionality is high. We do not observe that here, which we believe is due to the use of highly optimised code for (regular) dense matrix multiplication.

Kernel	Features	$m/d = 1$	$m/d = 2$	$m/d = 3$	$m/d = 4$	$m/d = 5$	$m/d = 6$	$m/d = 7$
Gaussian	IID	0.0408 (0.0008)	0.0753 (0.0009)	0.112 (0.00037)	0.156 (0.00058)	0.2 (0.00088)	0.256 (0.0013)	0.305 (0.00074)
	ORF	0.0434 (0.00026)	0.0805 (0.00088)	0.125 (0.0012)	0.171 (0.00066)	0.218 (0.00071)	0.282 (0.00052)	0.334 (0.001)
	SORF	0.0413 (0.00081)	0.0748 (0.00082)	0.115 (0.00039)	0.158 (0.00083)	0.204 (0.0011)	0.263 (0.00098)	0.312 (0.0012)
Matérn-5/2	IID	0.041 (0.00084)	0.0737 (0.00072)	0.115 (0.0006)	0.158 (0.0008)	0.201 (0.00084)	0.259 (0.0015)	0.312 (0.0015)
	ORF	0.0449 (0.00031)	0.0805 (0.00089)	0.127 (0.0008)	0.173 (0.00064)	0.218 (0.0011)	0.286 (0.00081)	0.337 (0.00062)
	SORF	0.0415 (0.0016)	0.0748 (0.001)	0.117 (0.00074)	0.162 (0.0011)	0.205 (0.00072)	0.265 (0.0012)	0.316 (0.0013)
Laplace	IID	0.0427 (0.00057)	0.0801 (0.0006)	0.121 (0.00086)	0.169 (0.00085)	0.214 (0.00069)	0.272 (0.0016)	0.325 (0.00078)
	ORF	0.0458 (0.0003)	0.0847 (0.00054)	0.133 (0.0011)	0.183 (0.0007)	0.23 (0.00072)	0.297 (0.001)	0.349 (0.00074)
	SORF	0.0431 (0.00021)	0.0801 (0.00055)	0.122 (0.0013)	0.169 (0.0011)	0.217 (0.00086)	0.29 (0.012)	0.326 (0.0016)

Table 7.6. Approximate GP regression results on cpu dataset. Reported numbers are average runtime (in seconds), along with bootstrap estimates of standard error (in parentheses).

Kernel	Features	$m/d = 1$	$m/d = 2$	$m/d = 3$	$m/d = 4$	$m/d = 5$	$m/d = 6$	$m/d = 7$
Gaussian	IID	30940.0 (820.0)	11470.0 (180.0)	6677.0 (100.0)	4362.0 (59.0)	3083.0 (25.0)	2415.0 (34.0)	1906.0 (28.0)
	ORF	**28010.0 (550.0)**	**11020.0 (120.0)**	**6313.0 (110.0)**	**4192.0 (47.0)**	**3023.0 (57.0)**	**2337.0 (22.0)**	**1847.0 (31.0)**
	SORF	31700.0 (680.0)	12520.0 (300.0)	7011.0 (210.0)	4635.0 (72.0)	3330.0 (57.0)	2514.0 (39.0)	1979.0 (33.0)
Matérn-5/2	IID	62860.0 (1200.0)	26610.0 (240.0)	16740.0 (160.0)	12030.0 (130.0)	9344.0 (50.0)	**7519.0 (110.0)**	**6424.0 (51.0)**
	ORF	**57660.0 (250.0)**	**26110.0 (190.0)**	**16390.0 (110.0)**	**12020.0 (100.0)**	**9344.0 (100.0)**	7584.0 (120.0)	6474.0 (39.0)
	SORF	60670.0 (760.0)	26500.0 (260.0)	16810.0 (140.0)	12300.0 (160.0)	9557.0 (71.0)	7708.0 (42.0)	6495.0 (97.0)
Laplace	IID	198300.0 (1500.0)	96220.0 (810.0)	62890.0 (850.0)	47220.0 (550.0)	38770.0 (460.0)	31470.0 (450.0)	27440.0 (400.0)
	ORF	**197400.0 (1700.0)**	**95470.0 (570.0)**	**62140.0 (440.0)**	**46260.0 (550.0)**	**37030.0 (360.0)**	**31290.0 (230.0)**	**27180.0 (290.0)**
	SORF	198600.0 (2400.0)	96500.0 (960.0)	64060.0 (740.0)	46880.0 (530.0)	38490.0 (400.0)	32190.0 (390.0)	27260.0 (330.0)

Table 7.7. Approximate GP regression results on wine dataset. Reported numbers are average KL divergence from true posterior, along with bootstrap estimates of standard error (in parentheses).

Kernel	Features	$m/d = 1$	$m/d = 2$	$m/d = 3$	$m/d = 4$	$m/d = 5$	$m/d = 6$	$m/d = 7$
Gaussian	IID	0.86 (0.005)	0.82 (0.002)	0.82 (0.002)	0.81 (0.002)	0.8 (0.002)	0.8 (0.002)	0.8 (0.002)
	ORF	**0.85 (0.003)**	**0.82 (0.002)**	**0.81 (0.003)**	**0.8 (0.003)**	**0.8 (0.002)**	**0.8 (0.002)**	**0.79 (0.002)**
	SORF	0.86 (0.004)	0.83 (0.002)	0.82 (0.002)	0.81 (0.002)	0.8 (0.002)	0.8 (0.002)	0.8 (0.001)
Matérn-5/2	IID	0.87 (0.005)	0.83 (0.003)	0.81 (0.001)	0.81 (0.002)	0.8 (0.002)	0.8 (0.002)	0.8 (0.002)
	ORF	**0.84 (0.003)**	0.82 (0.003)	**0.81 (0.003)**	0.81 (0.002)	**0.8 (0.002)**	0.8 (0.002)	0.79 (0.002)
	SORF	0.86 (0.005)	**0.82 (0.004)**	0.81 (0.002)	**0.8 (0.002)**	0.8 (0.002)	**0.8 (0.001)**	**0.79 (0.002)**
Laplace	IID	**0.88 (0.01)**	0.84 (0.006)	**0.82 (0.002)**	0.82 (0.004)	0.81 (0.002)	0.8 (0.003)	0.8 (0.002)
	ORF	0.89 (0.01)	**0.83 (0.004)**	0.82 (0.003)	**0.81 (0.002)**	0.8 (0.003)	**0.8 (0.001)**	0.8 (0.003)
	SORF	0.88 (0.007)	0.84 (0.004)	0.82 (0.004)	0.81 (0.003)	**0.8 (0.002)**	0.8 (0.003)	**0.8 (0.002)**

Table 7.8. Approximate GP regression results on wine dataset. Reported numbers are average test RMSE, along with bootstrap estimates of standard error (in parentheses).

Kernel	Features	$m/d = 1$	$m/d = 2$	$m/d = 3$	$m/d = 4$	$m/d = 5$	$m/d = 6$	$m/d = 7$
Gaussian	IID	0.0183 (0.00064)	0.0399 (0.00069)	0.0529 (0.00049)	0.0693 (0.0006)	0.0897 (0.001)	0.11 (0.00085)	0.126 (0.00099)
	ORF	0.0198 (0.00045)	0.0419 (0.00058)	0.0567 (0.00077)	0.0747 (0.00068)	0.097 (0.0007)	0.118 (0.00094)	0.137 (0.00091)
	SORF	0.0194 (8.6e-05)	0.0411 (0.00026)	0.055 (0.00095)	0.0715 (0.00076)	0.0911 (0.00063)	0.112 (0.00059)	0.132 (0.0008)
Matérn-5/2	IID	0.0188 (0.00024)	0.038 (0.00072)	0.0553 (0.00045)	0.0732 (0.00066)	0.0933 (0.00091)	0.113 (0.00073)	0.131 (0.00076)
	ORF	0.02 (0.00018)	0.0407 (0.00064)	0.0588 (0.00087)	0.0781 (0.00081)	0.0974 (0.00074)	0.122 (0.00092)	0.143 (0.00067)
	SORF	0.0192 (0.00041)	0.0402 (0.00048)	0.0574 (0.0008)	0.0746 (0.00057)	0.0944 (0.00064)	0.113 (0.0012)	0.135 (0.00071)
Laplace	IID	0.0198 (0.00018)	0.0394 (0.00064)	0.0586 (0.00047)	0.0773 (0.00041)	0.0959 (0.00072)	0.117 (0.00094)	0.14 (0.00066)
	ORF	0.0205 (3.5e-05)	0.0415 (0.00051)	0.0625 (0.00061)	0.0833 (0.0008)	0.104 (0.00057)	0.129 (0.00073)	0.152 (0.0012)
	SORF	0.0202 (0.00014)	0.0411 (0.00038)	0.0592 (0.00067)	0.0785 (0.00067)	0.101 (0.00093)	0.12 (0.00086)	0.143 (0.0011)

Table 7.9. Approximate GP regression results on wine dataset. Reported numbers are average runtime (in seconds), along with bootstrap estimates of standard error (in parentheses).

Chapter 8

Conclusions

8.1 Contributions

In this thesis, we have studied two fundamental questions: (i) how can we understand the performance of Sherali–Adams polytopes using higher-order cluster consistency conditions? and (ii) how can we understand the effectiveness of orthogonality conditions in Monte Carlo sampling? We briefly review our high-level contributions across these topics:

- Two new hybrid conditions for tightness of Sherali–Adams relaxations, based on a primal perturbation argument, and an argument using graph minor theory and polyhedral geometry. These two conditions represent theoretical steps towards understanding the effectiveness of Sherali–Adams relaxations on real-world instances of graphical models.
- A study of uprooting and rerooting transformations for general binary graphical models, which may be used in conjunction with a wide variety of techniques for approximate inference, along with analysis of the interactions of these transformations with Sherali–Adams relaxations for MAP inference, in particular leading to a proof of the unique universal rootedness of $\mathbb{L}_3$.
- Quantitative descriptions of the behaviour of Hadamard–Rademacher random matrices and several generalisations thereof for random projections, casting light on the effectiveness of these random matrices that has recently been observed in a variety of domains.

- Analysis and empirical evaluation of orthogonal random features for the angular kernel, and for a variety of stationary kernels, including the introduction of the charm function, which succinctly summarises the degree of effectiveness of orthogonal features.

8.2 Future work

There are several high-level themes in this thesis which naturally present questions for further work. We list several of these themes and questions below.

- **Exactness guarantees for marginal and MAP inference.** Our study of the tightness of Sherali–Adams relaxations has used a variety of techniques, and centred on hybrid conditions for exactness of Sherali–Adams approximations to MAP inference. Sherali–Adams relaxations also feature prominently in variational approaches to approximate marginal inference, most notably in Bethe-Kikuchi approximations. A natural question for further exploration is to what extent these exactness guarantees can be transferred over to marginal inference. More generally, it will be interesting to see to what extent the techniques developed in this thesis are applicable more generally to understanding hybrid conditions for exact inference (Cooper and Živný, 2011, 2017).
- **Orthogonality as a variance reduction technique.** Variance reduction has been studied as long as Monte Carlo methods themselves, and its importance is only growing for modern statistics and machine learning applications, owing to the recent proliferation of high-dimensional models and large datasets. Geometric couplings, such as the orthogonality conditions studied here, are coming to the forefront as variance-reduction techniques in a wide range of applications, including LSH (Andoni et al., 2015), dimensionality reduction (Choromanski et al., 2017), kernel approximation (Choromanski et al., 2018a; Yu et al., 2016), and black-box policy optimisation (Choromanski et al., 2018b), and there are many other applications which are natural candidates for such methods, such as variational inference in generative models (Kingma and Welling, 2014; Rezende et al., 2014) and policy gradient methods in methods in reinforcement learning (Mnih et al., 2016). Understanding the benefits of orthogonality in these new domains forms an interesting direction for future work.
- **Further study of methods for approximate orthogonal matrix sampling.** Tightly connected to the previous point, the statistical efficiency of orthogonally-coupled samples in many applications naturally leads to the question of how sampling, and computation with these samples, can be done in a computationally efficient manner. In this thesis,

we have obtained quantitative descriptions of the behaviour of Hadamard–Rademacher random matrices. There remain important theoretical and practical questions regarding the properties of this particular sampling mechanism, and there are many other sampling mechanisms which warrant further exploration, such as Kac's random walk (Pillai and Smith, 2018).

Bibliography

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147 – 169.

Ailon, N. and Chazelle, B. (2006). Approximate nearest neighbors and the fast Johnson–Lindenstrauss transform. In *Symposium on Theory of Computing (STOC)*.

Ailon, N. and Chazelle, B. (2009). The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322.

Andoni, A., Indyk, P., Laarhoven, T., Razenshteyn, I., and Schmidt, L. (2015). Practical and optimal LSH for angular distance. In *Neural Information Processing Systems (NIPS)*.

Arnborg, S., Proskurowski, A., and Corneil, D. (1990). Forbidden minors characterization of partial 3-trees. *Discrete Mathematics*, 80(1).

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Aung, A., Ng, B. P., and Rahardja, S. (2009). Conjugate symmetric sequency-ordered complex hadamard transform. *Trans. Sig. Proc.*, 57(7):2582–2593.

Barahona, F. (1982). On the computational complexity of Ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241.

Barahona, F. (1983). The max-cut problem on graphs not contractible to K_5. *Operations Research Letters*, 2(3):107–111.

Barahona, F., Grötschel, M., Jünger, M., and Reinelt, G. (1988). An application of combinatorial optimization to statistical physics and circuit layout design. *Operations Research*, 36(3):493–513.

Bartolucci, F. and Forcina, A. (2000). A likelihood ratio test for MTP_2 within binary variables. *The Annals of Statistics*, 28(4):1206–1218.

Batra, D., Nowozin, S., and Kohli, P. (2011). Tighter relaxations for MAP-MRF inference: A local primal-dual gap based separation algorithm. In *Artificial Intelligence and Statistics (AISTATS)*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Bodlaender, H. (1998). A partial k-arboretum of graphs with bounded treewidth. *Theoretical Computer Science*, 209(1-2):1–45.

Bojarski, M., Choromanska, A., Choromanski, K., Fagan, F., Gouy-Pailler, C., Morvan, A., Sakr, N., Sarlos, T., and Atif, J. (2017). Structured adaptive and random spinners for fast machine learning computations. In *Artificial Intelligence and Statistics (AISTATS)*.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239.

Chandrasekaran, V., Srebro, N., and Harsha, P. (2008). Complexity of inference in graphical models. In *Uncertainty in Artificial Intelligence (UAI)*.

Choromanski, K., Rowland, M., Sarlos, T., Sindhwani, V., Turner, R. E., and Weller, A. (2018a). The geometry of random features. In *Artificial Intelligence and Statistics (AISTATS)*.

Choromanski, K., Rowland, M., Sindhwani, V., Turner, R. E., and Weller, A. (2018b). Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning (ICML)*.

Choromanski, K., Rowland, M., and Weller, A. (2017). The unreasonable effectiveness of structured random orthogonal embeddings. In *Neural Information Processing Systems (NIPS)*.

Choromanski, K. and Sindhwani, V. (2016). Recycling randomness with structure for sublinear time kernel expansions. In *International Conference on Machine Learning (ICML)*.

Cook, W. J., Cunningham, W. H., Pulleyblank, W. R., and Schrijver, A. (1998). *Combinatorial Optimization*. John Wiley & Sons.

Cooper, M. C. and Živný, S. (2011). Hybrid tractability of valued constraint problems. *Artif. Intell.*, 175(9-10):1555–1569.

Cooper, M. C. and Živný, S. (2017). Hybrid tractable classes of constraint problems. In *The Constraint Satisfaction Problem: Complexity and Approximability*, pages 113–135.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. In *Machine Learning*, pages 273–297.

Dabney, W., Rowland, M., Bellemare, M. G., and Munos, R. (2017). Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Dasgupta, A., Kumar, R., and Sarlós, T. (2010). A sparse Johnson-Lindenstrauss transform. In *Symposium on Theory of Computing (STOC)*, pages 341–350.

Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Deza, M. and Laurent, M. (1997). *Geometry of Cuts and Metrics*. Springer, 1st edition.

Dick, J. and Pillichshammer, F. (2010). *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press.

Diestel, R. (2010). *Graph Theory*. Springer, fourth edition.

Eaton, F. and Ghahramani, Z. (2009). Choosing a variable to clamp. In *Artificial Intelligence and Statistics (AISTATS)*.

Erdogdu, M. A., Deshpande, Y., and Montanari, A. (2017). Inference in graphical models via semidefinite programming hierarchies. In *Neural Information Processing Systems (NIPS)*.

Esary, J. D., Proschan, F., and Walkup, D. W. (1967). Association of random variables, with applications. *Ann. Math. Statist.*, 38(5):1466–1474.

Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N., and Zwiernik, P. (2017). Total positivity in Markov structures. *Ann. Statist.*, 45(3):1152–1184.

Fortuin, C. M., Kasteleyn, P. W., and Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Comm. Math. Phys.*, 22(2):89–103.

Gales, M. and Young, S. (2007). The application of hidden Markov models in speech recognition. *Found. Trends Signal Process.*, 1(3):195–304.

Geelen, J., Gerards, B., and Whittle, G. (2014). Solving Rota's conjecture. *Notices of the AMS*, 61(7):736–743.

Goemans, M. X. and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145.

Guenin, B. (2001). A characterization of weakly bipartite graphs. *Journal of Combinatorial Theory, Series B*, 83(1):112–168.

Haberman, S. J. (1973). Log-linear models for frequency data: Sufficient statistics and likelihood equations. *Ann. Statist.*, 1(4):617–632.

Hammer, P. L., Hansen, P., and Simeone, B. (1984). Roof duality, complementation and persistency in quadratic 0-1 optimization. *Math. Program.*, 28:121–155.

Hernández-Lobato, J. M., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., and Turner, R. E. (2016). Black-box alpha divergence minimization. In *International Conference on Machine Learning (ICML)*.

Heskes, T., Albers, K., and Kappen, B. (2003). Approximate inference and constrained optimization. In *Uncertainty in Artificial Intelligence (UAI)*.

Hinrichs, A. and Vybíral, J. (2011). Johnson–Lindenstrauss lemma for circulant matrices. *Random Structures & Algorithms*, 39(3):391–398.

Holland, P. W. and Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, 14(4):1523–1543.

Honeine, P. and Richard, C. (2010). The angular kernel in machine learning for hyperspectral data classification. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*.

Huynh, T. (2009). *The linkage problem for group-labelled graphs*. PhD thesis, University of Waterloo.

Izmailov, P., Novikov, A., and Kropotov, D. (2018). Scalable gaussian processes with billions of inducing inputs via tensor train decomposition. In *Artificial Intelligence and Statistics (AISTATS)*.

Jaimovich, A., Elidan, G., Margalit, H., and Friedman, N. (2006). Towards an integrated protein–protein interaction network: A relational Markov network approach. *Journal of Computational Biology*, 13(2):145–164.

Jerrum, M. and Sinclair, A. (1993). Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comput.*, 22(5):1087–1116.

Joachims, T. (2006). Training linear SVMs in linear time. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Johnson, W. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206.

Kappes, J. H., Andres, B., Hamprecht, F. A., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B. X., Kröger, T., Lellmann, J., Komodakis, N., Savchynskyy, B., and Rother, C. (2015). A comparative study of modern inference techniques for structured discrete energy minimization problems. *Int. J. Comput. Vision*, 115(2):155–184.

Karlin, S. and Rinott, Y. (1981). Total positivity properties of absolute value multinormal variables with applications to confidence interval estimates and related probabilistic inequalities. *Ann. Statist.*, 9(5):1035–1049.

Karp, R. (1972). Reducibility among combinatorial problems. In Miller, R. and Thatcher, J., editors, *Complexity of Computer Computations*, pages 85–103.

Kingma, D. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*.

Kohli, P. and Rother, C. (2012). Higher-order models in computer vision. In Lezoray, O. and Grady, L., editors, *Image Processing and Analysing Graphs: Theory and Practice*, chapter 3. CRC Press.

Kohli, P. and Torr, P. H. S. (2007). Dynamic graph cuts for efficient inference in Markov random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2079–2088.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Kolmogorov, V., Thapper, J., and Živný, S. (2015). The power of linear programming for general-valued CSPs. *SIAM Journal on Computing*, 44(1):1–36.

Komodakis, N. and Paragios, N. (2008). Beyond loose LP-relaxations: Optimizing MRFs by repairing cycles. In *European Conference on Computer Vision (ECCV)*.

Kondor, R. and Barbosa, M. S. (2010). Ranking with kernels in Fourier space. In *Conference on Learning Theory (COLT)*.

Korte, B. and Vygen, J. (2007). *Combinatorial Optimization: Theory and Algorithms*. Springer, 4th edition.

Land, A. H. and Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica*, 28(3):pp. 497–520.

Laurent, M. (2003). A comparison of the sherali-adams, lovász-schrijver, and lasserre relaxations for 0–1 programming. *Math. Oper. Res.*, 28(3):470–496.

Lauritzen, S., Uhler, C., and Zwiernik, P. (2018). Maximum likelihood estimation in Gaussian models under total positivity. *The Annals of Statistics*, page to appear.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50:157–224.

Le, Q., Sarlós, T., and Smola, A. (2013). Fastfood - approximating kernel expansions in loglinear time. In *International Conference on Machine Learning (ICML)*.

Lee, J., Sohl-Dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. (2018). Deep neural networks as Gaussian processes. In *International Conference on Learning Representations (ICLR)*.

Lehman, A. (1990). On the width-length inequality and degenerate projective planes. In *Polyhedral Combinatorics*, pages 101–105. American Mathematical Society.

Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Lichtenstein, D. (1982). Planar formulae and their uses. *SIAM Journal on Computing*, 11(2):329–343.

MacKay, D. J. and Neal, R. M. (1995). Good codes based on very sparse matrices. In *Cryptography and Coding*. Springer.

MacKay, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press.

Malioutov, D. M., Johnson, J. K., and Willsky, A. S. (2006). Walk-sums and belief propagation in gaussian graphical models. *J. Mach. Learn. Res.*, 7:2031–2064.

Marchetti, G. M. and Wermuth, N. (2016). Palindromic bernoulli distributions. *Electron. J. Statist.*, 10(2):2435–2460.

Matthews, A. G. D. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations (ICLR)*.

Mezard, M. and Montanari, A. (2009). *Information, Physics, and Computation*. Oxford University Press.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning (ICML)*.

Mooij, J. (2010). libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173.

Murray, I. and Ghahramani, Z. (2004). Bayesian learning in undirected graphical models: Approximate MCMC algorithms. In *Uncertainty in Artificial Intelligence (UAI)*.

Ormoneit, D. and Sen, S. (2002). Kernel-based reinforcement learning. *Machine Learning*, 49(2-3):161–178.

Padberg, M. (1989). The boolean quadric polytope: Some characteristics, facets and relatives. *Math. Program.*, 45(1):139–172.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pillai, N. S. and Smith, A. (2018). On the mixing time of Kac's walk and other high-dimensional Gibbs samplers with constraints. *Ann. Probab.*, 46(4):2345–2399.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Neural Information Processing Systems (NIPS)*.

Rasmussen, C. and Williams, C. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*.

Robertson, N. and Seymour, P. (1986). Graph minors. II. algorithmic aspects of tree-width. *Journal of Algorithms*, 7(3):309 – 322.

Robertson, N. and Seymour, P. (1995). Graph minors. XIII. The disjoint paths problem. *Journal of Combinatorial Theory, Series B*, 63(1):65–110.

Robertson, N. and Seymour, P. (2004). Graph minors. XX. Wagner's conjecture. *Journal of Combinatorial Theory, Series B*, 92(2):325–357.

Rowland, M., Bellemare, M. G., Dabney, W., Munos, R., and Teh, Y. W. (2018). An analysis of categorical distributional reinforcement learning. In *Artificial Intelligence and Statistics (AISTATS)*.

Rowland, M., Pacchiano, A., and Weller, A. (2017). Conditions beyond treewidth for tightness of higher-order LP relaxations. In *Artifical Intelligence and Statistics (AISTATS)*.

Sanders, D. (1993). *Linear algorithms for graphs of tree-width at most four*. PhD thesis, Georgia Tech.

Schiex, T., Fargier, H., and Verfaillie, G. (1995). Valued constraint satisfaction problems: Hard and easy problems. In *International Joint Conference on Artificial Intelligence (ICJAI)*.

Schmidt, L., Sharifi, M., and Moreno, I. (2014). Large-scale speaker identification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1650–1654. IEEE.

Schneider, R. (1993). *Convex bodies : the Brunn-Minkowski theory*. Cambridge University Press.

Schoenberg, I. (1938). Metric Spaces and Completely Monotone Functions. *The Annals of Mathematics*, 39(4):811–841.

Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press.

Sherali, H. and Adams, W. (1990). A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3(3):411–430.

Shwe, M. A., Middleton, B., Heckerman, D. E., Henrion, M., Horvitz, E. J., Lehmann, H. P., and Cooper, G. F. (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. *Methods of information in medicine*, 30(4):241–255.

Solin, A. (2016). *Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression*. PhD thesis, Aalto University.

Sontag, D. (2010). *Approximate Inference in Graphical Models using LP Relaxations*. PhD thesis, Massachusetts Institute of Technology.

Sontag, D., Choe, D. K., and Li, Y. (2012). Efficiently searching for frustrated cycles in MAP inference. In *Uncertainty in Artificial Intelligence (UAI)*.

Sontag, D., Meltzer, T., Globerson, A., Jaakkola, T., and Weiss, Y. (2008). Tightening LP relaxations for MAP using message passing. In *Uncertainty in Artificial Intelligence (UAI)*.

Stone, L. D., Corwin, T. L., and Barlow, C. A. (1999). *Bayesian Multiple Target Tracking*. Artech House, 1st edition.

Sundaram, N., Turmukhametova, A., Satish, N., Mostak, T., Indyk, P., Madden, S., and Dubey, P. (2013). Streaming similarity search over one billion tweets using parallel locality-sensitive hashing. *Proceedings of the VLDB Endowment*, 6(14):1930–1941.

Sutherland, D. and Schneider, J. (2015). On the error of random Fourier features. In *Uncertainty in Artificial Intelligence (UAI)*.

Thapper, J. and Živný, S. (2016). The complexity of finite-valued CSPs. *Journal of the ACM*, 63(4).

Tripuraneni, N., Rowland, M., Ghahramani, Z., and Turner, R. E. (2017). Magnetic Hamiltonian Monte Carlo. In *International Conference on Machine Learning (ICML)*.

Tyler, D. (1987). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika*, 74(3):579–589.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). Graph kernels. *J. Mach. Learn. Res.*, 11:1201–1242.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theor.*, 13(2):260–269.

Vybíral, J. (2011). A variant of the Johnson–Lindenstrauss lemma for circulant matrices. *Journal of Functional Analysis*, 260(4):1096–1105.

Wainwright, M., Jaakkola, T., and Willsky, A. (2002). Map estimation via agreement on (hyper)trees: Message-passing and linear programming approaches. *IEEE Transactions on Information Theory*, 51:3697–3717.

Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Wainwright, M. J. and Jordan, M. I. (2004). Treewidth-based conditions for exactness of the Sherali-Adams and Lasserre relaxations. Technical report, University of California, Berkeley.

Wainwright, M. J., Ravikumar, P., and Lafferty, J. (2006). High-dimensional graphical model selection using l1-regularized logistic regression. In *Neural Information Processing Systems (NIPS)*.

Watanabe, Y. (2011). Uniqueness of belief propagation on signed graphs. In *Neural Information Processing Systems (NIPS)*.

Weller, A. (2015). Revisiting the limits of MAP inference by MWSS on perfect graphs. In *Artificial Intelligence and Statistics (AISTATS)*.

Weller, A. (2016a). Characterizing tightness of LP relaxations by forbidding signed minors. In *Uncertainty in Artifical Intelligence (UAI)*.

Weller, A. (2016b). Uprooting and rerooting graphical models. In *International Conference on Machine Learning (ICML)*.

Weller, A. and Domke, J. (2016). Clamping improves TRW and mean field approximations. In *Artificial Intelligence and Statistics (AISTATS)*.

Weller, A. and Jebara, T. (2014). Clamping variables and approximate inference. In *Neural Information Processing Systems (NIPS)*.

Weller, A., Rowland, M., and Sontag, D. (2016). Tightness of LP relaxations for almost balanced models. In *Artificial Intelligence and Statistics (AISTATS)*.

Williams, C. K. I. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Neural Information Processing Systems (NIPS)*.

Yedidia, J., Freeman, W., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Information Theory*, 51:2282–2312.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2001). Bethe free energy, Kikuchi approximations and belief propagation algorithms. Technical report, Mitsubishi Electric Research Laboratories.

Yu, F., Suresh, A., Choromanski, K., Holtmann-Rice, D., and Kumar, S. (2016). Orthogonal random features. In *Neural Information Processing Systems (NIPS)*.

Zhang, H. and Cheng, L. (2013). New bounds for circulant Johnson–Lindenstrauss embeddings. *Communications in Mathematical Sciences*, 12.

Zhang, X., Yu, F. X., Guo, R., Kumar, S., Wang, S., and Chang, S. (2015). Fast orthogonal projection based on Kronecker product. In *International Conference on Computer Vision (ICCV)*.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. S. (2015). Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*.