A Theory of Generative Models and Robustness via Regularization

Hisham Husain

A thesis submitted for the degree of
Doctor of Philosophy
The Australian National University

May 2021

Draft Copy – 14 May 2021

Except where otherwise indicated, this thesis is my own original work. This thesis is based on the following publications

- **Husain, H**., Nock, R., Williamson, R.C. (2019). A Primal-Dual Link between GANs and Autoencoders. In NeurIPS2019
- **Husain, H**., Balle, B., Cranko, Z., Nock, R. (2020). Local Differential Privacy for Sampling. In AISTATS2020.
- **Husain, H**. (2020). Distributional Robustness with IPMs with Links to GANs and Regularization. In NeurIPS2020
- **Husain, H**., Ciosek, K., Tomioka, R. (2021). Regularized Policies are Reward Robust. In AISTATS2021.

I contributed to majority of the development of research ideas, writing and implementation of the above publications. Other works completed during the candidature not included in this thesis are

- Knoblauch, J., **Husain, H**., Diethe, T. (2020). Optimal Continual Learning has Perfect Memory and is NP-Hard. in ICML2020.
- Mhammedi, Z., **Husain, H**. (2020). Risk-Monotonicity in Statistical Learning. ArXiv
- Soen, A., **Husain, H**., Nock, R. (2020). Data Preprocessing to Mitigate Bias with Fair Boosted Mollifiers. ArXiv

Hisham Husain
14 May 2021

to my mother, Adiba.

Acknowledgments

First of all, I would like to acknowledge my primary supervisor Prof. Richard Nock. There is nothing more than I could ask for in a supervisor than what Richard provided for me during my candidature. He provided tremendous guidance that aided me for both scientific and personal purposes. I want to thank the other members who have served on my supervisory panel at any point, including Professor Robert C. Williamson, Dr. Borja de Pigem Balle, and Professor Richard Hartley who have given me a significant amount of research guidance over the years.

I acknowledge the financial support given by the Australian Government Research Training Program (RTP) and CSIRO Data61 for providing me with office space to conduct my doctoral studies.

During the past years, I was fortunate to undertake two research internships in the lovely city of Cambridge, UK at two world-leading research labs: Amazon and Microsoft Research.

I want to thank Dr. Tom Diethe for hosting me at Amazon and exposing me to unique research projects. I would also like to thank my fellow interns who have become long-term collaborators from my time at Amazon, including Jeremias Knoblauch, Dr. Hans Kersting and Dr. Shuai Tang.

I would then like to thank Dr. Ryota Tomioka and Dr. Kamil Ciosek for hosting me during my time at Microsoft Research. My sincere gratitude goes to acknowledge Ryota for his fantastic mentorship and advice, which continued beyond the duration of the internship and Kamil for exposing me to the world of Reinforcement learning. I want to thank my fellow interns, Luisa Zintgraf and Tabish Rashid, who made the experience enjoyable.

I want to acknowledge my collaborator Dr. Wittawat Jitkrittum for hosting me as a visitor at the Max Planck Institute and giving me research advice.

I would then like to thank other collaborators and friends who have made the journey significantly easier including Zakaria Mhammedi, Alexander Soen, Dr. Gianpaolo Gioiosa, Vikrant Ashvinkumar, Anton Jurisevic, Jimmy Shi, Dr. Zac Cranko and Dustin Bates.

Finally, I want to thank my family, who always encouraged me to pursue my studies.

Abstract

Regularization in Machine Learning (ML) is a central technique with great practical significance, whose motivations depend on the learning setting. For example, the popular entropic regularization scheme for Reinforcement Learning (RL) is used to aid in disambiguating optimal policies. On the other hand, Generative Adversarial Networks (GANs) employ regularization for computational purposes, avoiding instability in training. Despite the widespread use and multi-faceted motivation of regularization, extensive evidence has suggested that regularization is a crucial method towards empirical success. Therefore, it is natural that a formal study of regularization would present results that aid in closing the gap between theory and practice.

In this thesis, we study a range of different learning problems from the unifying perspective of regularization and uncover various results that contribute to our understanding of machine learning methods. First, we focus on generative modelling and discover a primal-dual relationship between two pioneering methods in the literature of generative modelling, namely Generative Adversarial Networks (GANs) and Autoencoders. The discovery not only explicates a bridge between existing results but proves to be helpful in algorithmic guidance. The study on generative models is then extended to build a boosting-based model that can generate samples compliant with local differential privacy. We then focus on machine learning robustness, where one is interested in understanding the susceptibility of a model in the face of adversarial threats. We show that regularization is intimately connected to distributional robustness, which subsumes existing results and extends them to a great deal of generality, including applications to the unsupervised learning setting. We continue this narrative to the RL setting and similarly expose the robustifying benefits of using regularization, which sheds light on the widely-used entropy-regularized schemes, amongst others. In summary, this thesis's study of regularization contributes substantially within the literature of generative modelling, machine learning robustness, and RL while touching upon additional domains such as privacy and boosting.

Contents

Chapter 1

Introduction

The learning process in Machine Learning (ML) typically occurs as a well-understood optimization problem such as minimizing an error term or maximizing a reward. By denoting Θ to be set of models and $L_n : \Theta \to \mathbb{R}$ be a loss function, then minimizing error amounts to finding

$$\min_{\theta \in \Theta} \left(L_n(\theta) + \lambda \Omega(\theta) \right).$$

The loss function $L_n(\theta)$ depends on the training data, and the term Ω (with weight $\lambda > 0$) is a penalty term, such as a norm, which is typically included to prevent overfitting. The inclusion of Ω is referred to as *regularization*, whose purpose more generally is to incorporate additional information to an ill-posed problem by adding a form of control to the problem. This notion exists beyond the realm of machine learning, such as Tikhonov regularization, which has been widely applied for numerical methods [Calvetti and Reichel, 2003]. In the above example, one can interpret the optimization problem as a trade-off between fitting the data (through the term L) and model complexity (through the term Ω).

Regularization plays a crucial role in the success of machine learning methods and their empirical performances. The details of implementation and motivation, however, vary depending on the specific learning task of interest. We now give examples of two concrete machine learning problem domains and how regularization can manifest as an appealing, practical consideration.

In Reinforcement Learning (RL), one is given an environment and a reward function, where the goal is to learn a policy that can navigate the environment to achieve the maximum reward [Sutton and Barto, 2018]. In these problems, many different policies achieve optimal reward; however, we are often interested in finding those possessing exploratory behaviour. This is especially interesting in environments that are understudied or unexplored, and domain investigation is of interest. For example, in medical applications, a clinician would be interested in discovering different treatment types (policies) to understand the patient and problem better [Masood and Doshi-Velez, 2019]. Including entropy in the optimization task is a form of regularization, and a typical remedy for these problems, since it favours policies with diversified behaviour and, therefore, disambiguates optimal policies from this standpoint.

On the other hand, a popular method for generative modelling, Generative Adversarial Networks (GANs), minimizes the Jensen-Shannon divergence approximated by a set of discriminators $\mathcal{D}$ [Goodfellow et al., 2014a]. Theoretically, the strength of this approximation improves for more extensive choices of $\mathcal{D}$, often implemented by deep neural networks. However, this set is restricted or regularised in practice to avoid instabilities in training and has found great success [Fedus et al., 2017], despite the approximation being further removed from the divergence. Compared to the RL setting, the motivation for regularization here is primarily computational.

Despite the multi-faceted motivation for practicality, regularization in the above examples of GANs and RL is heuristic and *deviates* from the formally motivated original learning optimization procedure. Considering the wide-scale use and practical significance of regularization, a theoretical investigation on the ramifications would unveil foundational results for machine learning that naturally bridge the gap between theory and practice. We further exemplify the advantage of studying regularization along two axes:

(1) There are many choices for regularization, which are often decided at the practitioner's discretion. While having more options seem bountiful, they can also be a curse. The choice itself is overwhelming for the practitioner, given the sheer number of different schemes available without knowing the precise formal relationships. A prime example of this is the generative modelling literature, which has witnessed various GANs and Autoencoder methods instantiated by different choices of regularizers. In particular, the Wasserstein Autoencoder (WAE) [Tolstikhin et al., 2017] uses a penalty function Ω whose choice is plentiful and highly understudied. An in-depth study of regularization, which can connect different models or provide reinterpretations, would clarify the effect of each choice and naturally serve a great purpose to the practitioner.

(2) Among many different domains of ML, the addition of regularization has witnessed empirical benefits. While this may not be surprising, it is remarkable from a formal perspective. Recalling the above example on RL, the reward maximization problem is based on Markov Decision Processes (MDP), which is directly motivated through the axioms of utility theory [Russell and Norvig, 2002]; however, adding entropic regularization to this objective has no such axiomatic provocation. Moreover, it also turns out that entropic-regularised RL consistently outperforms its standard counterpart on the original MDP. Similarly, regularization in GANs has also illustrated superior performance. A theoretical study outlining the benefits of regularization would significantly improve the transparency of these methods.

1.1 Thesis Contribution

The thesis studies a range of different learning problems from the unifying perspective of regularization. We first focus on generative models since various regularization schemes are largely understudied in light of (1). We then study regularization

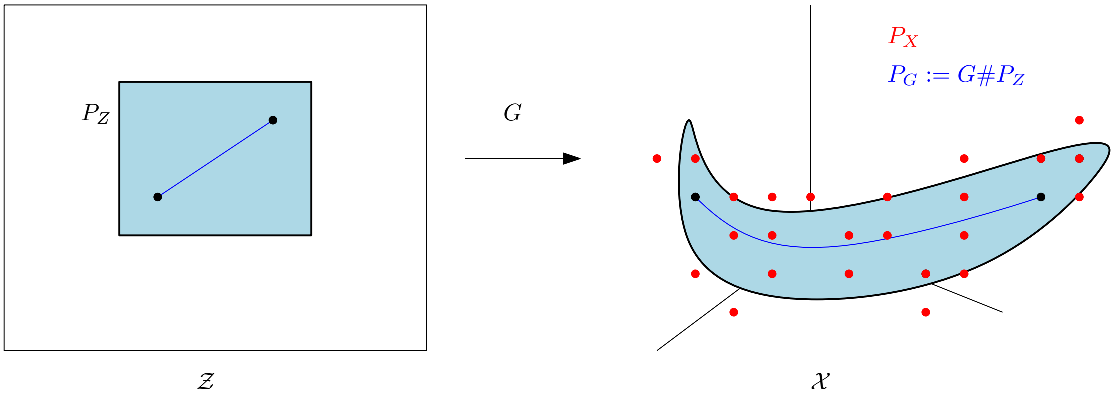


Figure 1.1: Implicit generative models are given a distribution P_X and seek to find a function $G : \mathcal{Z} \to \mathcal{X}$ so that the distribution of P_Z pushed through G matches P_X.

more generally in other settings such as supervised learning and reinforcement learning, discovering benefits of robustness which naturally addresses (2). We find several other results that contribute to machine learning more generally. We summarize the contributions into two categories: generative models and robustness.

1.1.1 Generative Models

Generative modelling aims to produce a model that mimics a given dataset P_X, referred to as training data, in the sense of matching its probability distribution and can be viewed as solving the classical density estimation problem. A subclass of these methods referred to as *implicit* generative models [Diggle and Gratton, 1984; Mohamed and Lakshminarayanan, 2016] approaches this problem by learning a function $G : \mathcal{Z} \to \mathcal{X}$ that transforms a simple distribution P_Z (such as a unit Gaussian or uniform distribution) in a latent space $\mathcal{Z}$ to match the given dataset P_X. We illustrate this pictorially in Figure 1.1, where $G\#P_Z$ denotes the model distribution. The function G is typically parameterized by a neural network, and the training process involves ensuring that G minimizes $D(G\#P_Z, P_X)$ where D is a dissimilarity (or distance) measure between distributions. The main advantage is that one can easily sample from this distribution by first sampling from P_Z followed by applying G; however, such models' densities are not available in closed form. Therefore, our primary understanding of the model distribution is characterized entirely by its samples. This drawback limits the transparency of what is being learned and makes the training process strenuous since computing $D(G\#P_Z, P_X)$ must be sample-based and, therefore, incurs additional approximation error. Despite this, implicit models have achieved tremendous success, impacting and resurging several applications, which include and are not limited to unpaired domain translation [Zhu et al., 2017], data imputation [Yoon et al., 2018] and experimental calibration [Amodio and Krishnaswamy, 2018].

The literature on implicit generative modelling exhibits a large body of different models, each differing in their choice of the dissimilarity measure D. Two examples are Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [Kingma and Welling, 2013]. Both methods have differing backgrounds and formulations, which naturally leads to each having contrasting benefits and downfalls. In GANs, the dissimilarity D is chosen based on how well a set of discriminators $\mathcal{D}$ can distinguish between $G\#P_Z$ and P_X, which consequently becomes a game between the model G and discriminators $\mathcal{D}$. VAEs, on the other hand, represent the data in a latent space with the use of an encoder, which in comparison to GANs, runs into more minor computational stability issues when training. Additionally, the encoder has use-cases outside of generative modelling, unlike the discriminator from GANs; for example, one can understand the data and domain by analyzing the latent structure learned. Unfortunately, it has been observed that GANs have consistently produced results with a higher degree of precision relative to those attained by VAEs. However, a model similar to the VAE, known as the Wasserstein Autoencoder (WAE), was proposed, achieving performance similar to that of GANs, posing as a candidate for unifying benefits of both models. The WAE brings along with it some slight ambiguities, such as a regularization term whose choices are plentiful and whose ramifications are unknown. It is not clear which choice one should employ and whether the similarity to GANs is a mere coincidence.

Chapter 3 provides a study on the regularization mechanisms behind these models and discovers an equivalence between GANs and WAE. The connection itself explains the similarity in results between WAE and GANs, and therefore provides a form of clarification as mentioned in (1). However, in much more generality, the result lays the groundwork and builds the start of a taxonomy for understanding this vast literature on generative modelling. Using the bridge formed by this equivalence, we derive generalization bounds for WAEs, which have been studied and exclusively existed for GANs [Zhang et al., 2017]. Our analysis of generalization is relevant to the finite data scheme, where most of these models are operated on, and the absence of such results in the literature has been noted as a particular shortcoming of previous theoretical studies [Li and Malik, 2018].

Practically speaking, the result derives a specific choice of regularization that connects WAE to GANs, which is then naturally encouraged. One of the equivalence conditions is to enforce a Lipschitzness constraint, a form of regularization, over the discriminators in GANs, which has been previously utilized in the GAN objective and consequently demonstrated great empirical success [Zhou et al., 2019; Farnia and Tse, 2018].

One application of generative models is Differential Privacy (DP), where one is interested in generating data that resembles the training data without retaining sensitive and confidential information. However, to tackle this, one needs a strong understanding of the model, and it is challenging for implicit models due to the density's unavailability. In Chapter 4, we contribute to filling the gap by proposing a generative model, although not implicit, that can comply with privacy constraints.

Inspired by the use of discriminators in GANs, we develop a boosting scheme that learns a density compliant with local differential privacy for sampling. The method itself introduces a general framework for DP referred to as *mollification*, where one can privatize any density model and is of independent interest to the literature of privacy. The specific algorithm we present is efficient and uses boosting to achieve DP by way of mollification. A crucial highlight of our method is due to the post-processing property of DP, meaning that privacy and sensitive information are not retained for any task employing samples produced by this generative model. Therefore, our method has use-cases for various methods that we exemplify in the setting of natural language generation.

We then present guarantees from the perspective of utility and show that the density learned converges to the information-theoretic model in Kullback-Leibler (KL) divergence with increasing rounds of boosting. Moreover, we show guarantees on the learned density's mode-coverage abilities - an area where GANs have been notably lacking [Goodfellow, 2016, Section 5.1.1].

1.1.2 Robustness

The premise that regularization, or reducing model complexity, improves performance on unseen data points is well-embraced in ML, as it can be traced to Occam's razor. This fundamental principle roughly states that we should prefer the simplest ones among all hypotheses consistent or sufficiently accurate. From a formal perspective, this rationale has been conceptualized in statistical learning theory [Vapnik, 1999], where one is interested in bounding the difference in population ("true data") and empirical ("observed data") loss. These results are referred to as *generalization bounds* and commonly depend on a complexity measure such as the Vapnik-Chervonkis (VC) dimension [Vapnik and Chervonenkis, 1974] or Rademacher complexity [Bartlett and Mendelson, 2002]. In a nutshell, the results suggest that regularized or reduced model classes will have a reduced gap on their population and empirical losses. These bounds have been improved and significantly tightened in the probabilistic setting by replacing the complexity with a measure of deviation from a prior belief [McAllester, 1999; Guedj, 2019].

An alternative measure of utility beyond generalization is to analyze the *robustness* of a model. In particular, an area of study known as *adversarial* robustness has been receiving increasing popularity due to the discovery that small perturbations can change the decision-making abilities of deep neural networks [Goodfellow et al., 2014b; Madry et al., 2017]. This phenomenon is a critical concern since these models are often heavily relied upon in several applications, including healthcare and autonomous vehicles. A promising strategy to tackle this rigidity of models is to consider how robust ML models are to shifts in the distribution - a field of study recollected as Distributionally Robust Optimization (DRO) [Scarf, 1957]. In particular, we are interested in learning a model which performs well under an adversary who shifts the training data.

Naturally, the resulting formulation is intractable due to the added adversarial complication. In the specific context of ML, DRO has considered restricted adversaries who shift distributions with respect to a divergence or distance measure, reducing to a more computationally feasible problem. The majority of existing results consider the Wasserstein distance as the divergence and show that DRO in this setting is reduced to Lipschitz regularization - a popular choice of regularization with a large body of empirical validation [Blanchet and Murthy, 2019; Blanchet et al., 2019; Cranko et al., 2018; Shafieezadeh-Abadeh et al., 2019]. There exist other choices of divergences with similar findings such as the f-divergences [Duchi et al., 2013, 2016] and the well-regarded Maximum Mean Discrepancy (MMD) [Staib and Jegelka, 2019], which found connections to variance and Hilbert space regularization, respectively.

In this thesis, Chapter 5 extends and generalizes these results by considering DRO derived by Integral Probability Metrics (IPMs) - a family of divergences that include both the Wasserstein distance and MMD. We discover a general connection between DRO based on IPMs and various regularization schemes, which subsume existing results and boast new advancements. For example, we find that regularization schemes such as those appearing in manifold regularization [Belkin et al., 2006] and generalized variance are reductions of DRO formulations and thus provide a robustness reinterpretation. These new insights parallel the premise described in (1) as we work towards clarifying the role of regularization bridging objectives; however, they also provide guarantees and benefits as described in (2). The existing results deriving this connection have come in the form of inequalities; however, Chapter 5 derives a necessary and sufficient condition for equality. This condition not only comments on new choices of IPMs but also applies to and, therefore, improves upon existing results in MMD and Wasserstein distances.

The results are then applied to GANs, which to the best of our knowledge, is the first study of the distributional robustness of GANs. In particular, we study how robust the distribution learned by a GAN is to shifts in the training data. The main finding is similar to the previous contribution on generative modelling, where one can gain robustification benefits by regularizing the discriminator set, heavily touches upon understanding empirical benefits of regularization motivated in (2). Therefore, the results in unison provide a robustness perspective for much existing work, which have restricted and regularized their discriminators, such as Sobelov-, MMD- and Fisher-GANs [Li et al., 2017; Arbel et al., 2018; Mroueh and Sercu, 2017; Mroueh et al., 2017].

In Chapter 6, the compelling narrative between regularization and robustness is then extended to the RL setting, which, as previously stated, is an area of study where regularization is extensively applied yet poorly understood formally. While we motivated the popular choice of entropic-regularization earlier, schemes involving other forms of regularization such as q-Tsallis or penalties based on prior information also exist.

At the technical level, the main result shows that general policy maximization can be interpreted as a one-player game that involves an adversary altering the set-up so that the best achievable reward is as small as possible. The result, therefore, applies to a wide variety of schemes with policy maximization including and not limited to regularized MDPs, minimization of a divergence such as in Generative Adversarial

Imitation Learning (GAIL) [Ho and Ermon, 2016] and even reward-free paradigms such as pure entropic maximization [Hazan et al., 2019].

The concrete takeaway for practice is that the policy learned in this fashion is optimal for the reward learned in the dual adversarial game. When applied to the extant entropy-based schemes, this forms a robustness guarantee that provides validation for empirical claims of success with such models. The conclusion drawn from this result aligns naturally with both (1) and (2) since we can reinterpret the role of entropy and describe the robustifying benefits.

To derive the main results, we encounter several supplementary results that are of independent interest to RL. An example of this includes a study on deep Q-learning [Watkins and Dayan, 1992] - an area of practical importance where one solves the RL problem using a supervised learning strategy. In particular, we find a novel link between deep Q learning and implicit regularization of the policy. Therefore, we can further clarify the effect of policy regularization in light of a particular method, deep Q learning, which parallels the notions and motivation set up in (1). Additionally, one can then use the above findings of the robustification benefits of regularization to derive guarantees for deep Q learning.

1.2 Thesis Structure and Publications

This thesis is structured as a compilation of four different papers. Since a significant portion of the results are based on convex analysis, we dedicate Chapter 2 to guide the reader and revise some facts on dualities and divergences between probability distributions and to set some recurring notations. Each chapter will begin with short text that links their respective content to the overarching narrative, followed by the respective publication formatted as they appear in proceedings. Chapter 7 will then draw together results with final remarks and future work. The chapter-publication correspondence is as follows:

- Chapter 3: A Primal-Dual Link between GANs and Autoencoders.
  H. Husain, R. Nock and R. Williamson. NeurIPS2019.

- Chapter 4: Local Differential Privacy for Sampling.
  H. Husain, B. Balle, Z. Cranko, and R. Nock. AISTATS2020.

- Chapter 5: Distributional Robustness with IPMs and Links to Regularization and GANs.
  H. Husain. NeurIPS2020.

- Chapter 6: Regularized Policies are Reward Robust.
  H. Husain, K. Ciosek and R. Tomioka. AISTATS2021.

For each of the above publications, I contributed to the majority of the development of research ideas, solutions, writing and implementation of all experiments involved.

Chapter 2

Preliminaries

The purpose of this chapter is to make the reader familiar with some elementary results in convex analysis and some central divergences between probability distributions. The discussions here will make the proofs of the main theorems throughout the thesis easier to digest.

2.1 Convex Analysis

First, we will revise relevant results in elementary convex analysis, which are readily found in [Penot, 2012]. We will use $\mathbb{R}$ and $\mathbb{N}$ to denote the real and natural numbers respectively, additionally defining $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ and $\mathbb{N}_* := \mathbb{N} \setminus \{0\}$. Each chapter will be defining some Polish space, which we will use $\mathcal{Z}$ here, with a σ-algebra Σ. More often than not, Σ will be the Borel σ-algebra. An important space is then the set of all bounded and finitely signed additive measures, given below.

Definition 1 (Bounded-Additive measures) *For any σ-algebra Σ over $\mathcal{Z}$, we define the set of bounded-additive measures, denoted by $\mathscr{B}(\mathcal{Z})$, as all measures $\mu : \Sigma \to \overline{\mathbb{R}}$ such that $\mu(\emptyset) = 0$ where $\emptyset \in \Sigma$ is the empty set and*

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n), \tag{2.1}$$

for any disjoint sequence $\{A_i\}_{i=1}^{\infty}$.

In particular, this set is of interest since $\mathscr{B}(\mathcal{Z})$ is a Banach space, where the norm is the variation of the measure [Dunford and Schwartz, 1988]. Another important Banach space we will consider is the set of bounded and measurable functions (with respect to Σ), which we denote by $\mathscr{F}(\mathcal{Z}, \mathbb{R})$. We recall a classical result which states that $\mathscr{B}(\mathcal{Z})$ and $\mathscr{F}(\mathcal{Z}, \mathbb{R})$ are continuous dual spaces of each other [Hildebrandt, 1934; Fichtenholz and Kantorovitch, 1934], where the dual pairing operator is given by

$$\langle h, \mu \rangle = \int_{\mathcal{Z}} h(z) d\mu(z), \tag{2.2}$$

for any $h \in \mathscr{F}(\mathcal{Z}, \mathbb{R})$ and $\mu \in \mathscr{B}(\mathcal{Z})$. We then denote an especially interesting subset

$\mathscr{P}(\mathcal{Z}) \subset \mathscr{B}(\mathcal{Z})$ as the set of *probability* measures which are all measures μ such that $\mu(\mathcal{Z}) = 1$.

Example 1 *Consider the finite choice of $\mathcal{Z} = \{1, \ldots, k\}$ for some $k \in \mathbb{N}_*$. In this setting, both $\mathscr{F}(\mathcal{Z}, \mathbb{R})$ and $\mathscr{B}(\mathcal{Z})$ are isomorphic to $\mathbb{R}^k$, noting that Euclidean spaces are self-dual. Moreover, $\langle \cdot, \cdot \rangle$ corresponds to the standard inner product operation on Euclidean spaces and $\mathscr{P}(\mathcal{Z})$ are all vectors whose entries are non-negative and sum to 1.*

As one can imagine, this duality becomes vastly more interesting when $\mathcal{Z}$ is an infinite set. Throughout the remainder of this section, we will denote by X to be either of $\mathscr{B}(\mathcal{Z})$ or $\mathscr{F}(\mathcal{Z}, \mathbb{R})$ with $X^\star$ denoting the continuous dual. We now introduce elementary notions of convex analysis.

Definition 2 (Convex function) *A function $F : X \to \mathbb{R}$ is convex if for any $x, x' \in X$ and $t \in [0,1]$, it holds $F(tx + (1-t)x') \leq tF(x) + (1-t)F(x')$.*

We say that a convex function is proper if there exists a $x \in X$ such that $F(x) < \infty$ and $F(x') > -\infty$ for all $x' \in X$ and define the *domain* of F as

$$\operatorname{dom}(F) := \{x \in X : F(x) < \infty\}.$$

We now define the *Legendre-Fenchel* conjugate of a function F.

Definition 3 (Rockafellar [1968]) *For any function $F : X \to (-\infty, \infty]$, we define the conjugate $F^\star : X^\star \to (-\infty, \infty]$ as*

$$F^\star(x^\star) = \sup_{x \in \operatorname{dom}(F)} \left(\langle x, x^\star \rangle - F(x) \right)$$

and the double conjugate $F^{\star\star} : X \to (-\infty, \infty]$ as

$$F^{\star\star}(x) = \sup_{x^\star \in \operatorname{dom}(F^\star)} \left(\langle x, x^\star \rangle - F^\star(x^\star) \right)$$

The Legendre-Fenchel conjugate, or often called the convex conjugate, allows us to represent a function in its dual space. This conjugate operation becomes interesting in the scenario when the double conjugate recovers itself, as one can expect with a *dual*. We give the conditions on a function in order to satisfy this.

Theorem 1 (Zalinescu [2002] Theorem 2.3.3) *If X is a Hausdorff locally convex space, and $F : X \to (-\infty, \infty]$ is a proper convex lower semi-continuous function then $F^{\star\star} = F$.*

Therefore, if $F : X \to (-\infty, \infty]$ satisfies the above conditions then we have

$$F(x) = \sup_{x^\star \in \operatorname{dom}(F^\star)} \left(\langle x, x' \rangle - F^\star(x^\star) \right). \tag{2.3}$$

This rewriting of F is a useful strategy when deriving alternative forms of function and will be commonly used in the proof techniques of this thesis. In the next section, we will illustrate how one can use this trick and derive a current result. Similar to integral convolutions in Fourier analysis, there exists a convolution that preserves interesting properties in convex analysis.

Definition 4 (Infimal convolution) *For any $F, H : X \to (-\infty, \infty]$, we define the infimal convolution, $F \overline{\star} H : X \to (-\infty, \infty]$*

$$(F \overline{\star} H)(x) = \inf_{x' \in X} \left(F(x') + H(x - x') \right).$$

The infimal convolution is a central operation - the properties of which have been extensively studied in [Strömberg, 1994]. The operation can be viewed as the analogue of addition in the dual space due to the following result.

Lemma 1 ([Penot, 2012], Proposition 3.43) *For any pair of functions $F, H : X \to (-\infty, \infty]$, it holds that*

$$(F + H)^{\star}(x) = \left(F^{\star} \overline{\star} H^{\star} \right)(x)$$

The above essentially asserts that the conjugate of the addition is equal to the convolution of their conjugates.

2.2 Divergences

Distortion measures between probability distributions form the backbone to many results presented in this thesis. This section will discuss three main families of distortion measures and their existing relationships. We first introduce the notion of a divergence.

Definition 5 (Divergence) *A divergence $D : \mathscr{P}(\mathcal{Z}) \times \mathscr{P}(\mathcal{Z}) \to \mathbb{R}$ satisfies $D(\mu, \nu) \geq 0$ with $D(\mu, \nu) = 0 \impliedby \mu = \nu$.*

A divergence can be thought of a a *distance* over probability measures however it need not be one since a distance formally requires satisfaction of the triangle inequality, symmetry and $D(\mu, \nu) = 0 \iff \mu = \nu$. The first family of divergence we introduce depends on a convex function f.

Definition 6 (f-divergence) *For any convex lower semicontinuous function $f : \mathbb{R} \to (-\infty, \infty]$ such that $f(1) = 0$, the f-divergence between any two probability distributions is $\mu, \nu \in \mathscr{P}(\mathcal{Z})$ is*

$$D_f(\mu, \nu) = \int_{\mathcal{Z}} f\left(\frac{d\mu}{d\nu} \right) d\nu, \tag{2.4}$$

if $\mu \ll \nu$ and $+\infty$ otherwise.

Table 2.1: Example of some f-divergences and their corresponding conjugate functions $f^\star$.

Name	$f(t)$	$f^\star(t)$
Kullback-Leibler	$t\log t$	$\exp(t-1)$
Pearson χ^2	$(t-1)^2$	$\frac{t^2}{4}+t$
GAN divergence	$t\log t-(t+1)\log(t+1)$	$\log(1-\exp(t))$
Reverse KL	$-\log t$	$\log\left(-\frac{1}{t}\right)$
Squared Hellinger	$\left(\sqrt{t}-1\right)^2$	$\frac{t}{1-t}$

The f-divergence [Csiszár, 1964; Ali and Silvey, 1966; Csiszár, 1967] is a well-regarded choice of discrepancy between distributions, and includes the popular Kullback-Leibler (KL) divergence, instantianted with $f(t) = t\log(t)$. Other well-regarded choices of f and their corresponding divergences are detailed in Table 2.1. It is important to note that one requires absolute continuity of μ with respect to ν, which is often recollected as a drawback of f-divergences, since in practice finitely supported distributions will be excluded. Moreover, the f-divergence is not symmetric i.e. $D(\mu,\nu) \neq D(\nu,\mu)$, nor does it necessarily satisfy the triangle inequality. Typical methods of estimating f-divergences include first approximating the density ratio $\frac{d\mu}{d\nu}$ [Sugiyama et al., 2012].

The f-divergence admits a dual form which has been leveraged for estimation such as in [Ruderman et al., 2012; Nowozin et al., 2016], however is also useful for theoretical purposes as we will demonstrate in Chapter 3.

Lemma 2 ([Nguyen et al., 2010]) *For any proper convex lower semicontinuous function $f:\mathbb{R}\to(-\infty,\infty]$ with $f(1)=0$, the f-divergence between any two probability distributions $\mu,\nu\in\mathscr{P}(\mathcal{Z})$ satisfies*

$$D_f(\mu,\nu) = \sup_{h\in\mathscr{F}(\mathcal{Z},\mathrm{dom}(f^\star))}\left(\mathbb{E}_\mu[h]-\mathbb{E}_\nu[f^\star(h)]\right), \tag{2.5}$$

where $f^\star(t)=\sup_{t'\in\mathrm{dom}(f)}\left(tt'-f(t')\right)$ is the Legendre-Fenchel conjugate.

This is typically referred to as the *variational form* of the f-divergence and has strong links to class probability estimation [Reid et al., 2011].

Example 2 *We will show now how convex analysis described above can be used to derive the above variational form. Letting $X=\mathbb{R}$ and noting that $X^\star=\mathbb{R}$, we apply self-conjugacy to*

$f = f^{\star\star}$, which yields

$$\begin{aligned}
D_f(\mu, \nu) &= \int_{\mathcal{Z}} f\left(\frac{d\mu}{d\nu}\right) d\nu, \\
&= \int_{\mathcal{Z}} f^{\star\star}\left(\frac{d\mu}{d\nu}\right) d\nu, \\
&= \int_{\mathcal{Z}} \sup_{h \in \mathrm{dom}(f^\star)} \left(h \cdot \frac{d\mu}{d\nu} - f^\star(h)\right) d\nu, \\
&\overset{(*)}{=} \sup_{h \in \mathscr{F}(\mathcal{Z}, \mathrm{dom}(f^\star))} \int_{\mathcal{Z}} \left(h \cdot \frac{d\mu}{d\nu} - f^\star(h)\right) d\nu, \\
&= \sup_{h \in \mathscr{F}(\mathcal{Z}, \mathrm{dom}(f^\star))} \left(\mathbb{E}_\mu[h] - \mathbb{E}_\nu[f^\star(h)]\right),
\end{aligned}$$

where the step $(*)$ requires more technalities such as decomposibility of the space $\mathscr{F}(\mathcal{Z}, \mathrm{dom}(f^\star))$ [Rockafellar and Wets, 2009].

We now move to the second divergence.

Definition 7 (Integral Probability Metric) *For a set $\mathcal{F} \subseteq \mathscr{F}(\mathcal{Z}, \mathbb{R})$, the Integral Probability Metric (IPM) between $\mu, \nu \in \mathscr{P}(\mathcal{Z})$ is*

$$d_{\mathcal{F}}(\mu, \nu) = \sup_{h \in \mathcal{F}} \left(\mathbb{E}_\mu[h] - \mathbb{E}_\nu[h]\right). \tag{2.6}$$

The identification of *integral probability metric* was used in [Müller, 1997] however the IPM had appeared as *probability metrics with ζ-structure* earlier in [Zolotarev, 1984]. The set $\mathcal{F}$ characterizes the strength of the IPM as a dissimilarity measure between distributions. For example, if $\mathcal{F}$ only contains constant functions then $d_{\mathcal{F}}(\mu, \nu) = 0$ for all μ, ν. If $-\mathcal{F} = \mathcal{F}$ then $d_{\mathcal{F}}$ is symmetric and additional technical assumptions on $\mathcal{F}$ are required to ensure $d_{\mathcal{F}}(\mu, \nu) = 0 \iff \mu = \nu$. Furthermore, $d_{\mathcal{F}}$ will satisfy the triangle-inequality for any choice of $\mathcal{F}$. The Total Variation distance is an instance of the IPM when selecting $\mathcal{F}$ to be all functions bounded above and below by 1, which is also an f-divergence for the choice of $f(t) = |t-1|$. The conditions on f and $\mathcal{F}$ such that D_f and $d_{\mathcal{F}}$ relate to each other have been extensively pursued in [Sriperumbudur et al., 2009]. A popular and practical choice in ML for $\mathcal{F}$ is the set of functions in a Reproducing Kernel Hilbert Space with norm 1, which corresponds to the Maximum Mean Discrepancy (MMD) [Gretton et al., 2012]. In particular, the MMD is available in closed form for finitely supported distributions, making it a suitable choice in practice. To introduce the final family of divergences, we define the set of couplings between $\mu, \nu \in \mathscr{P}(\mathcal{Z})$ as

$$\Pi(\mu, \nu) = \{\pi \in \mathscr{P}(\mathcal{Z} \times \mathcal{Z}) : \pi(A \times \mathcal{Z}) = \mu(A), \pi(\mathcal{Z} \times A) = \nu(A), A \in \Sigma\}$$

We are now ready to introduce the divergence.

Definition 8 (Wasserstein distance) *For any cost $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, the 1-Wasserstein distance between two probability measures $\mu, \nu \in \mathscr{P}(\mathcal{Z})$ is*

$$W_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{Z} \times \mathcal{Z}} c(z, z') d\pi(z, z'). \tag{2.7}$$

The Wasserstein distance comes from the study of optimal transport [Villani, 2008] and the specific instance in Definition 8 is known as the 1-Wasserstein distance. Unlike the f-divergence, the Wasserstein distance does not require absolute continuity and is typically considered antidotal for cases where the f-divergences are not suitable. Similar to the f-divergence, the Wasserstein distance admits a well-renowned dual form, based on its linear programming interpretation, which coincides with the IPM when c is a metric.

Lemma 3 (Rubinstein-Kantorovich Duality) *Let $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a metric. It then holds that*

$$W_c(\mu, \nu) = \sup_{h \in \mathcal{H}_c} \left(\mathbb{E}_\mu[h] - \mathbb{E}_\nu[h] \right), \tag{2.8}$$

where $\mathcal{H}_c = \{h \in \mathscr{F}(\mathcal{Z}, \mathbb{R}) : h(z, z') \leq c(z, z'), \forall z, z' \in \mathcal{Z}\}$ is the set of 1-Lipschitz functions with respect to c.

Other choices of IPMs and Wasserstein distances will be discussed when used in Chapters 3 and 5.

2.3 Summary

This chapter revised elementary content in convex analysis, such as the Legendre-Fenchel dual, which will be instrumental for proving the main results of this thesis. Well-known choices of divergences were then revised, which will serve helpful, particularly for digesting the content of the following two chapters on generative models.

Chapter 3

A Primal-Dual Link between GANs and Autoencoders

This paper initiates the study of regularization for generative models. In this work, the Fenchel dual form of the restricted GAN setting is utilized and then shown connections to Autoencoders. The motivation is type (1) where we would like to understand how different regularizers affect models. Therefore, the main contribution shows a primal-dual relationship between GANs and Autoencoders, whose implications are then drawn out for theoretical purposes, along with explaining certain practicalities.

A Primal-Dual link between GANs and Autoencoders

Hisham Husain[‡,†] **Richard Nock**[†,‡,♣] **Robert C. Williamson**[‡,†]

[‡]The Australian National University, [†]Data61, [♣]The University of Sydney

firstname.lastname@{data61.csiro.au,anu.edu.au}

Abstract

Since the introduction of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAE), the literature on generative modelling has witnessed an overwhelming resurgence. The impressive, yet elusive empirical performance of GANs has lead to the rise of many GAN-VAE hybrids, with the hopes of GAN level performance and additional benefits of VAE, such as an encoder for feature reduction, which is not offered by GANs. Recently, the Wasserstein Autoencoder (WAE) was proposed, achieving performance similar to that of GANs, yet it is still unclear whether the two are fundamentally different or can be further improved into a unified model. In this work, we study the f-GAN and WAE models and make two main discoveries. First, we find that the f-GAN and WAE objectives partake in a primal-dual relationship and are equivalent under some assumptions, which then allows us to explicate the success of WAE. Second, the equivalence result allows us to, for the first time, prove generalization bounds for Autoencoder models, which is a pertinent problem when it comes to theoretical analyses of generative models. Furthermore, we show that the WAE objective is related to other statistical quantities such as the f-divergence and in particular, upper bounded by the Wasserstein distance, which then allows us to tap into existing efficient (regularized) optimal transport solvers. Our findings thus present the first primal-dual relationship between GANs and Autoencoder models, comment on generalization abilities and make a step towards unifying these models.

1 Introduction

Implicit probabilistic models [1] are defined to be the pushforward of a simple distribution P_Z over a latent space $\mathcal{Z}$ through a map $G : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{X}$ is the space of the input data. Such models allow easy sampling, but the computation of the corresponding probability density function is intractable. The goal of these methods is to match $G\#P_Z$ to a target distribution P_X by minimizing $D(P_X, G\#P_Z)$, for some discrepancy $D(\cdot,\cdot)$ between distributions. An overwhelming number of methods have emerged after the introduction of Generative Adversarial Networks [2, 3] and Variational Autoencoders [4] (GANs and VAEs), which have established two distinct paradigms: Adversarial (networks) training and Autoencoders respectively. Adversarial training involves a set of functions $\mathcal{D}$, referred to as *discriminators*, with an objective of the form

$$D(P_X, G\#P_Z) = \max_{d \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_X}[a(d(x))] - \mathbb{E}_{x \sim G\#P_Z}[b(d(x))] \right\}, \tag{1}$$

for some functions $a : \mathbb{R} \to \mathbb{R}$ and $b : \mathbb{R} \to \mathbb{R}$. Autoencoder methods are concerned with finding a function $E : \mathcal{X} \to \mathcal{Z}$, referred to as an *encoder*, whose goal is to reverse G, and learn a feature space with the objective

$$D(P_X, G\#P_Z) = \min_{E} \left\{ \mathcal{R}(G, E) + \Omega(E) \right\}, \tag{2}$$

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

where $\mathcal{R}(G, E)$ is the *reconstruction* loss and acts to ensure G and E reverse each other and $\Omega(E)$ is a regularization term. Much work on Autoencoder methods has focused upon the choice of Ω.

In practice, the two methods demonstrate contrasting abilities in their strengths and limitations, which have resulted in differing directions of progress. Indeed, there is a lack of theoretical understanding of how these frameworks are parametrized and it is not clear whether the methods are fundamentally different. For example, Adversarial training based methods have empirically demonstrated high performance when it comes to producing realistic looking samples from P_X. However, GANs often have problems in convergence and stability of training [5]. Autoencoders, on the other hand, deal with a more well behaved objective and learn an encoder in the process, making them useful for feature representation. However in practice, Autoencoder based methods have reported shortfalls, such as producing blurry samples for image based datasets [6]. This has motivated researchers to adapt Autoencoder models by borrowing elements from Adversarial networks in the hopes of GAN level performance whilst learning an encoder. Examples include replacing Ω with Adversarial objectives [7, 8] or replacing the reconstruction loss with an adversarial objective [9, 10]. Recently, the Wasserstein Autoencoder (WAE) [6] has been shown to subsume these two methods with an Adversarial based Ω, and has demonstrated performance similar to that of Adversarial methods.

Understanding the connection between the two paradigms is important for not only the practical purposes outlined above but for the inheritance of theoretical analyses from one another. For example, when it comes to directions of progress, Adversarial training methods now have theoretical guarantees on generalization performance [11], however no such theoretical results have been obtained to date for autoencoders. Indeed, generalization performance is a pressing concern, since both techniques implicitly assume the samples represent the target distribution [12] and eventually leads to memorizing training data.

In this work, we study the two paradigms and in particular focus on the f-GANs [3] for Adversarial training and Wasserstein Autoencoders (WAE) for Autoencoders, which generalize the original GAN and VAE models respectively. We prove that the f-GAN objective with Lipschitz (with respect to a metric c) discriminators is equivalent to the WAE objective with cost c. In particular, we show that the WAE objective is an upper bound; schematically we get

$$\boxed{f\text{-GAN} \leq \text{WAE}}$$

and discuss the tightness of this bound. Our result is a generalization of the Kantorovich-Rubinstein duality and thus suggests a primal-dual relationship between Adversarial and Autoencoder methods. Consequently we show, to the best of our knowledge, the first generalization bounds for autoencoders. Furthermore, using this equivalence, we show that the WAE objective is related to key statistical quantities such as the f-divergence and Wasserstein distance, which allows us to tap into efficient (regularized) OT solvers.

The main contributions can be summarized as the following:

▷ (Theorem 8) Establishes an equivalence between Adversarial training and Wasserstein Autoencoders, showing conditions under which the f-GAN and WAE coincide. This further justifies the similar performance of WAE to GAN based methods. When the conditions are not met, we have an inequality, which allows us to comment on the behavior of the methods.

▷ (Theorem 9, 10 and 14) Show that the WAE objective is related to other statistical quantities such as f-divergence and Wasserstein distance.

▷ (Theorem 13) Provide generalization bounds for WAE. In particular, this focuses on the empirical variant of the WAE objective, which allows the use of Optimal Transport (OT) solvers as they are concerned with discrete distributions. This allows one to employ efficient (regularized) OT solvers for the estimation of WAE, f-GANs and the generalization bounds.

2 Preliminaries

2.1 Notation

We will use $\mathcal{X}$ to denote the input space (a Polish space), typically taken to be a Euclidean space. We use $\mathcal{Z}$ to denote the latent space, also taken to be Euclidean. We use $\mathbb{N}_*$ to denote the natural numbers without 0: $\mathbb{N} \setminus \{0\}$. We denote by $\mathscr{P}$ the set of probability measures over $\mathcal{X}$, and elements of this set will be referred to as *distributions*. If $P \in \mathscr{P}(\mathcal{X})$ happens to be absolutely continuous with respect to the Lebesgue measure then we will use $dP/d\lambda$ to refer to the *density* function (Radon-Nikodym derivative with respect to the Lebesgue measure). For any $T \in \mathscr{F}(\mathcal{X}, \mathcal{Z})$, for any measure $\mu \in \mathscr{P}(\mathcal{X})$, the pushforward measure of μ through T denoted $T\#\mu \in \mathscr{P}(\mathcal{Z})$ is such that $T\#\mu(A) = \mu(T^{-1}(A))$ for any measurable set $A \subset \mathcal{Z}$. The set $\mathscr{F}(\mathcal{X}, \mathbb{R})$ refers to all measurable functions from $\mathcal{X}$ into the set $\mathbb{R}$. We will use functions to represent conditional distributions over a space $\mathcal{Z}$ conditioned on elements $\mathcal{X}$, for example $P \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))$ so that for any $x \in \mathcal{X}$, $P(x) = P(\cdot|x) \in \mathscr{P}(\mathcal{Z})$. For any $P \in \mathscr{P}(\mathcal{X})$, the *support* of P is $\text{supp}(P) = \{x \in \mathcal{X} : \text{ if } x \in N_x \text{ open } \implies P(N_x) > 0\}$. In any metric space $(\mathcal{X}, c)$, for any set $S \subseteq \mathcal{X}$, we define the *diameter* of S to be $\text{diam}_c(S) = \sup_{x,x' \in S} c(x, x')$. Given a metric c over $\mathcal{X}$, for any $f \in \mathscr{F}(\mathcal{X}, \mathbb{R})$, $\text{Lip}_c(f)$ denotes the Lipschitz constant of f with respect to c and $\mathcal{H}_c = \{g \in \mathscr{F}(\mathcal{X}, \mathbb{R}) : \text{Lip}_c(g) \leq 1\}$. For some set $S \subseteq \mathbb{R}$, $\mathbf{1}_S$ corresponds to the convex *indicator function*, ie. $\mathbf{1}_S(x) = 0$ if $x \in S$ and $\mathbf{1}_S(x) = \infty$ otherwise. For any $x \in \mathcal{X}$, $\delta_x : \mathcal{X} \to \{0, 1\}$ corresponds to the *characteristic function*, with $\delta_x(0) = 1$ if $x = 0$ and $\delta_x(0) = 0$ if $x \neq 0$.

2.2 Background

2.2.1 Probability Discrepancies

Probability discrepancies are central to the objective of finding the best fitting model. We introduce some key discrepancies and their notation, which will appear later.

Definition 1 (f-Divergence) *For a convex function $f : \mathbb{R} \to (-\infty, \infty]$ with $f(1) = 0$, for any $P, Q \in \mathscr{P}(\mathcal{X})$ with P absolutely continuous with respect to Q, the f-Divergence between P and Q is*

$$D_f(P, Q) := \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ,$$

with $D_f(P, Q) = \infty$ if P is note absolutely continuous with respect to Q.

An example of a method to compute the f-divergence is to first compute dP/dQ and estimate the integral empirically using samples from Q.

Definition 2 (Integral Probability Metric) *For a fixed function class $\mathcal{F} \subseteq \mathscr{F}(\mathcal{X}, \mathbb{R})$, the Integral Probability Metric (IPM) based on $\mathcal{F}$ between $P, Q \in \mathscr{P}(\mathcal{X})$ is defined as*

$$\text{IPM}_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X}} f(x) dP(x) - \int_{\mathcal{X}} f(x) dQ(x) \right\}.$$

If we have that $-\mathcal{F} = \mathcal{F}$ then $\text{IPM}_{\mathcal{F}}$ forms a metric over $\mathscr{P}(\mathcal{X})$ [13]. A particular IPM we will make use of is Total Variation (TV): $\text{TV}(P, Q) = \text{IPM}_{\mathcal{V}}(P, Q)$ where $\mathcal{V} = \{h \in \mathscr{F}(\mathcal{X}, \mathbb{R}) : |h| \leq 1\}$. We also note that when $f(x) = |x - 1|$ then $\text{TV} = D_f$ and thus TV is both an IPM and an f-divergence.

Definition 3 *For any $P, Q \in \mathscr{P}(\mathcal{X})$, define the* set of couplings *between P and Q to be*

$$\Pi(P, Q) = \left\{ \pi \in \mathscr{P}(\mathcal{X} \times \mathcal{X}) : \int_{\mathcal{X}} \pi(x, y) dx = P, \int_{\mathcal{X}} \pi(x, y) dy = Q \right\}.$$

For a cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the Wasserstein distance *between P and Q is*

$$W_c(P, Q) := \inf_{\pi \in \Pi(P,Q)} \left\{ \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y) \right\}.$$

The Wasserstein distance can be regarded as an infinite linear program and thus admits a dual form, and in the case of c being a metric, belongs to the class of IPMs. We summarize this fact the following lemma [14].

Lemma 4 (Wasserstein Duality) *Let $(\mathcal{X}, c)$ be a metric space, and suppose $\mathcal{H}_c$ is the set of all 1-Lipschitz functions with respect to c. Then for any $P, Q \in \mathscr{P}(\mathcal{X})$, we have*

$$\begin{aligned} W_c(P, Q) &= \sup_{h \in \mathcal{H}_c} \left\{ \int_{\mathcal{X}} h(x) dP(x) - \int_{\mathcal{X}} h(x) dQ(x) \right\} \\ &= \text{IPM}_{\mathcal{H}_c}(P, Q). \end{aligned}$$

2.3 Generative Models

In both GAN and VAE models, we have a latent space $\mathcal{Z}$ (typically taken to be $\mathbb{R}^d$, with d being small) and a prior distribution $P_Z \in \mathscr{P}(\mathcal{Z})$ (e.g. unit variance Gaussian). We have a function referred to as the generator $G : \mathcal{Z} \to \mathcal{X}$, which induces the *generated* distribution, denoted by $P_G \in \mathscr{P}(\mathcal{X})$, as the pushforward of P_Z through G: $P_G = G\#P_Z$. The true data distribution will be referred to as $P_X \in \mathscr{P}(\mathcal{X})$. The common goal between the two methods is to find a generator G such that the samples generated by pushing forward P_Z through G ($G\#P_Z$) are close to the true data distribution (P_X). More formally, one can cast this as an optimization problem by finding the best G such that $D(P_G, P_X)$ is minimized, where $D(\cdot, \cdot)$ is some discrepancy between distributions. Both methods (as we outline below) utilize their own discrepancies between P_X and P_G, which offer their own benefits and weaknesses.

2.3.1 Wasserstein Autoencoder

Let $E : \mathcal{X} \to \mathscr{P}(\mathcal{Z})$ denote a probabilistic *encoder* [1], which maps each point x to a conditional distribution $E(x) \in \mathscr{P}(\mathcal{Z})$, denoted as the *posterior* distribution. The pushforward of P_X through E: $E\#P_X$, will be referred to as the *aggregated posterior*.

Definition 5 (Wasserstein Autoencoder [6]) *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, $\lambda > 0$ and $\Omega : \mathscr{P}(\mathcal{Z}) \times \mathscr{P}(\mathcal{Z}) \to \mathbb{R}_{\geq 0}$ with $\Omega(P, P) = 0$ for all $P \in \mathscr{P}(\mathcal{Z})$. The Wasserstein Autoencoder objective is*

$$\mathrm{WAE}_{c,\lambda\cdot\Omega}(P_X, G) = \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x) + \lambda \cdot \Omega(E\#P_X, P_Z) \right\}.$$

We remark that there are various choices of c and $\lambda \cdot \Omega$. [6] select these by tuning λ and selecting different measures of discrepancies between probability distortions for Ω.

2.3.2 f-Generative Adversarial Network

Let $d : \mathcal{X} \to \mathbb{R}$ denote a *discriminator* function.

Definition 6 (f-GAN [3]) *Let $f : \mathbb{R} \to (-\infty, \infty]$ denote a convex function with property $f(1) = 0$ and $\mathcal{D} \subset \mathscr{F}(\mathcal{X}, \mathbb{R})$ a set of discriminators. The f-GAN model minimizes the following objective for a generator $G : \mathcal{Z} \to \mathcal{X}$*

$$\mathrm{GAN}_f(P_X, G; \mathcal{D}) := \sup_{d \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_X}[d(x)] - \mathbb{E}_{z \sim P_Z}[f^*(d(G(z)))] \right\}, \qquad (3)$$

where $f^\star(x) = \sup_y \{x \cdot y - f(y)\}$ is the convex conjugate of f.

There are two knobs in this method, namely $\mathcal{D}$, the set of discriminators, and the convex function f. The objective in (3) is a variational approximation to D_f [3]; if $\mathcal{D} = \mathscr{F}(\mathcal{X}, \mathbb{R})$, then $\mathrm{GAN}_f(P_X, G; \mathcal{D}) = D_f(P_X, P_G)$ [15]. In the case of $f(x) = x\log(x) - (x+1)\log(x+1) + 2\log 2$, we recover the original GAN [2].

3 Related Work

Current attempts at building a taxonomy for generative models have largely been within each paradigm or the proposal of hybrid methods that borrow elements from the two. We first review major and relevant advances in each paradigm, and then move on to discuss results that are close to the technical contributions of our work.

The line of Autoencoders begin with $\Omega = 0$, which is the original autoencoder concerned only with reconstruction loss. VAE then introduced a non-zero Ω, along with implementing Gaussian encoders [4]. This was then replaced by an adversarial objective [7], which is sample based and consequently allows arbitrary encoders. In the spirit of unification, Adversarial Autoencoders (AAE) [8] proposed Ω to be a discrepancy between the pushforward of the target distribution through the encoder ($E\#P_X$)

[1] We remark that this is not standard notation in the VAE and Variational Inference literature.

and the prior distribution (P_Z) in the latent space, which was then showed to be equivalent to the VAE Ω minus a mutual information term [16]. Independently, InfoVAE [17] proposed a similar objective, which was subsequently shown to be equivalent to adding mutual information. [6] reparametrized the Wasserstein distance into an Autoencoder objective (WAE) where the Ω term generalizes AAE, and has reported performance comparable to that of Adversarial methods. Other attempts also include adjusting the reconstruction loss to be adversarial as well [9, 10]. Another work that focuses on WAE is the Sinkhorn Autoencoders (SAE) [18], which select Ω to be the Wasserstein distance and show that the overall objective is an upper bound to the Wasserstein distance between P_X and P_G.

[19] discussed the two paradigms and their unification by interpretting GANs from the perspective of variational inference, which allowed a connection to VAE, resulting in a GAN implemented with importance weighting techniques. While this approach is the closest to our work in forming a link, their results apply to standard VAE (and not other AE methods such as WAE) and cannot be extended to all f-GANs. [20] introduced the notion of an Adversarial divergence, which subsumed mainstream adversarial based methods. This also led to the formal understanding of how the selected discriminator set $\mathcal{D}$ affects the final generator G learned. However, this approach is silent with regard to Autoencoder based methods. [11] established the tradeoff between the Rademacher complexity of the discriminator class $\mathcal{D}$ and generalization performance of G, with no results present for Autoencoders. These theoretical advances in Adversarial training methods are inherited by Autoencoders as a consequence of the equivalence presented in our work.

One key point in the proof of our equivalence is the use of a result that decomposes the GAN objective into an f-divergence and an IPM for a restricted class of discriminators (which we used for Lipschitz functions). This decomposition is used in [21] and applied to linear f-GANs, showing that the adversarial training objective decomposes into a mixture of maximum likelihood and moment matching. [22] used this decomposition with Lipschitz discriminators like our work, however does not make any extension or further progress to establish the link to WAE. Indeed, GANs with Lipschitz discriminators have been independently studied in [23], which suggest that one should enforce Lipschitz constraints to provide useful gradients.

4 f-Wasserstein Autoencoders

We define a new objective, that will help us in the proof of the main theorems of this paper.

Definition 7 (f-Wasserstein Autoencoder) *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\lambda > 0$, $f : \mathbb{R} \to (-\infty, \infty]$ be a convex function (with $f(1) = 0$), the f-Wasserstein Autoencoder (f-WAE) objective is*

$$\overline{W}_{c,\lambda\cdot f}(P_X, G) = \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{W_c(P_X, (G \circ E)\#P_X) + \lambda D_f(E\#P_X, P_Z)\} \tag{4}$$

In the proof of the main result, we will show that the f-WAE objective is indeed the same as the WAE objective when using the same cost c and selecting the regularizer to be $\lambda \cdot \Omega = D_{\lambda f} = \lambda D_f$. The only difference between this and the standard WAE is the use of $W_c(P_X, (G \circ E)\#P_X)$ as the reconstruction loss instead of the standard cost which is an upper bound (Lemma 18), and the regularizer is chosen to be $\lambda \cdot \Omega = D_{\lambda f} = \lambda D_f$. We now present the main theorem that captures the relationship between f-GAN and WAE.

Theorem 8 (f-GAN and WAE equivalence) *Suppose $(\mathcal{X}, c)$ is a metric space and let $\mathcal{H}_c$ denote the set of all functions from $\mathcal{X} \to \mathbb{R}$ that are 1-Lipschitz (with respect to c). Let $f : \mathbb{R} \to (-\infty, \infty]$ be a convex function with $f(1) = 0$. Then for all $\lambda > 0$,*

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \leq \mathrm{WAE}_{c,\lambda\cdot D_f}(P_X, G), \tag{5}$$

with equality if G is invertible.

Proof (This is a sketch, see Section A.1 for full proof). The proof begins by proving certain properties of $\mathcal{H}_c$ (Lemma 16), allowing us to use the dual form of restricted GANs (Theorem 15),

$$\begin{aligned}\mathrm{GAN}_f(P_X, G; \mathcal{H}_c) &= \inf_{P' \in \mathscr{P}(\mathcal{X})} \left\{ D_f(P', P_G) + \sup_{h \in \mathcal{H}_c} \{\mathbb{E}_{P_X}[h] - \mathbb{E}_{P'}[h]\} \right\} \\ &= \inf_{P' \in \mathscr{P}(\mathcal{X})} \{D_f(P', P_G) + W_c(P', P_X)\}. \end{aligned} \tag{6}$$

The key is to reparametrize (6) by optimizing over couplings. By rewriting $P' = (G \circ E)\#P_X$ for some $E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))$ and rewriting (6) as an optimization over E (Lemma 20), we obtain

$$\begin{aligned} &\inf_{P' \in \mathscr{P}(\mathcal{X})} \{D_f(P', P_G) + W_c(P', P_X)\} \\ &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{D_f((G \circ E)\#P_X, P_G) + W_c((G \circ E)\#P_X, P_X)\} \end{aligned} \tag{7}$$

We then have

$$D_f((G \circ E)\#P_X, P_G) = D_f(G\#(E\#P_X), G\#P_Z) \overset{(*)}{\leq} D_f(E\#P_X, P_Z),$$

with equality in $(*)$ if G is invertible (Lemma 17). A weaker condition is required if f is differentiable, namely if G is invertible with respect to $f' \circ d(E\#P_X)/dP_Z$ in the sense that

$$G(z) = G(z') \implies f' \circ (d(E\#P_X)/dP_Z)(z) = f' \circ (d(E\#P_X)/dP_Z)(z'), \tag{8}$$

noting that an invertible G trivially satisfies this requirement. Letting $f \leftarrow \lambda f$, we have $D_f(\cdot,\cdot) \leftarrow \lambda D_f(\cdot,\cdot)$, and so from Equation 7, we have

$$\begin{aligned} \mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) &\overset{(*)}{\leq} \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{\lambda D_f(E\#P_X, P_Z) + W_c((G \circ E)\#P_X, P_X)\} \\ &= \overline{W}_{c,\lambda \cdot f}(P_X, G) \\ &\leq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \left\{\lambda D_f(E\#P_X, P_Z) + \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x)\right\} \\ &= \mathrm{WAE}_{c, \lambda \cdot D_f}(P_X, G), \end{aligned}$$

where the final inequality follows from the fact that $W_c(P,Q) \leq \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x)$ (Lemma 18). Using the fact that $\overline{W} \geq \mathrm{WAE}$ (Lemma 19) completes the proof. ■

When G is invertible, we remark that P_G can still be expressive and capable of modelling complex distributions in WAE and GAN models. For example, if G is implemented with feedforward neural networks, and G is invertible then P_G can model *deformed* exponential families [24], which encompasses a large class appearing in statistical physics and information geometry [25, 26]. There exists many invertible activation functions under which G will be invertible. Furthermore, in the proof of the Theorem it is clear that $\overline{W}$ and WAE are the same objective (from Lemma 18 and Lemma 19). When using $f = \mathbf{1}_{\{1\}}$ ($f(x) = 0$ if $x = 1$ and $f(x) = \infty$ otherwise), and noting that $f^\star(x) = x$, meaning that Theorem 8 (with $\lambda = 1$) reduces to

$$\begin{aligned} \sup_{h \in \mathcal{H}_c} \{\mathbb{E}_{x \sim P_X}[h(x)] - \mathbb{E}_{x \sim P_G}[h(x)]\} &= \mathrm{GAN}_f(P_X, G; \mathcal{H}_c) \\ &\leq \overline{W}_{c,f}(P_X, P_G) \\ &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})) : E\#P_X = P_Z} \{W_c(P_X, (G \circ E)\#P_X)\} \\ &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})) : E\#P_X = P_Z} \{W_c(P_X, G\#P_Z\} \\ &= W_c(P_X, P_G), \end{aligned}$$

which is the standard primal-dual relation between Wasserstein distances as in Lemma 4. Hence, Theorem 8 can be viewed as a generalization of this primal-dual relationship, where Autoencoder and Adversarial objectives represent primal and dual forms respectively.

We note that the left hand of Equation (5) does not explicitly engage the prior space Z as much as the right hand side in the sense that one can set $Z = \mathcal{X}$, $G = \mathrm{Id}$ (which is invertible) and $P_Z = P_G$ and indeed results in the exact same f-GAN objective since $G\#P_Z = \mathrm{Id}\#P_G = P_G$, yet the equivalent f-WAE objective (from Theorem 8) will be different. This makes the Theorem versatile in reparametrizations, which we exploit in the proof of Theorem 10. We now consider weighting the reconstruction along with the regularization term in $\overline{W}$ (which is equivalent to weighting WAE), which simply amounts to re-weighting the cost since for any $\gamma > 0$,

$$\overline{W}_{\gamma \cdot c, \lambda \cdot f}(P_X, G) = \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{\gamma W_c((G \circ E)\#P_X, P_X) + \lambda D_f(E\#P_X, P_Z)\}.$$

The idea of weighting the regularization term by λ was introduced by [27] and furthermore studied empirically, showing that the choice of λ influences learning disentanglement in the latent space. [28]. We show that if $\lambda = 1$ and γ is larger than some γ^* then $\overline{W}$ will become an f-divergence (Theorem 9). On the other hand if we fix $\gamma = 1$ and take λ is larger than some λ^*, then $\overline{W}$ becomes the Wasserstein distance and in particular, equality holds in (5) (Theorem 10). We show explicitly how high γ and λ need to be for such equalities to occur. This is surprising since f-divergence and Wasserstein distance are quite different distortions.

We begin with the f-divergence case. Consider $f : \mathbb{R} \to (-\infty, \infty]$ convex, differentiable and $f(1) = 0$ and assume that P_X is absolutely continuous with respect to P_G, so that $D_f(P_X, P_G) < \infty$.

Theorem 9 *Set $c(x, y) = \delta_{x-y}$ and let $f : \mathbb{R} \to (-\infty, \infty]$ be a convex function (with $f(1) = 0$) and differentiable. Let $\gamma^* = \sup_{x,x' \in \mathcal{X}} \left| f'\left(\frac{dP_X}{dP_G}\right) - f'(\frac{dP_X}{dP_G})(x') \right|$ and suppose P_G is absolutely continuous with respect to P_X and that G is invertible, then we have for all $\gamma \geq \gamma^*$*

$$\overline{W}_{\gamma \cdot c, f}(P_X, G) = D_f(P_X, P_G).$$

(Proof in Appendix, Section A.3). It is important to note that $W_c(P_X, P_G) = \mathrm{TV}(P_X, P_G)$ when $c(x, y) = \delta_{x-y}$ and so Theorem 9 tells us that the objective with a weighted total variation reconstruction loss with a f-divergence prior regularization amounts to the f-divergence. It was shown that in [24] that when G is an invertible feedforward neural network then $D_f(P_X, P_G)$ is a *Bregman* divergence (a well regarded quantity in information geometry) between the parametrizations of the network for a fixed choice of activation function of G, which depends on f. Hence, a practioner should design G with such activation function when using f-WAE under the above setting ($c(x, y) = \delta_{x-y}$ and $\gamma = \gamma^*$) with G being invertible, so that the information theoretic divergence (D_f) between the distributions becomes an information geometric divergence involving the network parameters.

We now show that if λ is selected higher than $\lambda^* := \sup_{P' \in \mathscr{P}(\mathcal{X})} \left(W_c(P', P_G)/D_f(P', P_G) \right)$, then $\overline{W}$ becomes W_c and furthermore we have equality between f-GAN, f-WAE and WAE.

Theorem 10 *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a metric. For any $f : \mathbb{R} \to (-\infty, \infty]$ convex function (with $f(1) = 0$), we have for all $\lambda \geq \lambda^*$*

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) = \overline{W}_{c, \lambda \cdot f}(P_X, G) = \mathrm{WAE}_{c, \lambda \cdot D_f}(P_X, G) = W_c(P_X, P_G).$$

(Proof in Appendix, Section A.4). Note that Theorem 10 holds for any f (satisfying properties of the Theorem) and so one can estimate the Wasserstein distance using any f as long as λ is scaled to λ^*. In order to understand how high λ^* can be,there are two extremes in which the supremum may be unbounded. The first case is when P' is taken far from P_G so that $W_c(P', P_G)$ increases, however one should note that in the case when $\Delta = \max_{x,x' \in \mathcal{X}} c(x, x') < \infty$ then $W_c \in [0, \Delta]$ and so W_c will be finite whereas $D_f(P', P_G)$ can possibly diverge to ∞, making $\lambda^* \to 0$. The other case is when P' is made close to P_G, in which case $\frac{1}{D_f(P', P_G)} \to \infty$ however $W_c(P', P_G) \to 0$ so the quantity λ^* can still be small in this case, depending on the rate of decrease between W_c and D_f. Now suppose that $f(x) = |x - 1|$ and $c(x, y) = \delta_{x-y}$, in which case $D_f = W_c$ and thus $\lambda^* = 1$. In this case, Theorem 10 reduces to the standard result [29] regarding the equivalence between Wasserstein distance and f-divergence intersecting at the variational divergence under these conditions.

5 Generalization bounds

We present generalization bounds using machinery developed in [30] with the following definitions.

Definition 11 (Covering Numbers) *For a set $S \subseteq \mathcal{X}$, we denote $N_\eta(S)$ to be the η-*covering number *of S, which is the smallest $m \in \mathbb{N}_*$ such that there exists closed balls $B_1, \ldots, B_m$ of radius η with $S \subseteq \bigcup_{i=1}^m B_i$. For any $P \in \mathscr{P}(\mathcal{X})$, the (η, τ)-*dimension *is $d_\eta(P, \tau) := \frac{\log N_\eta(P, \tau)}{-\log \eta}$, where $N_\eta(P, \tau) := \inf \{N_\eta(S) : P(S) \geq 1 - \tau\}$.*

Definition 12 (1-Upper Wasserstein Dimension) *The* 1-Upper Wasserstein dimension *of any $P \in \mathscr{P}(\mathcal{X})$ is $d^*(P) := \inf\left\{s \in (2, \infty) : \limsup_{\eta \to 0} d_\eta(P, \eta^{\frac{s}{s-2}}) \leq s\right\}$.*

We make an assumption of P_X and P_G having bounded support to achieve the following bounds. For any $P \in \mathscr{P}(\mathcal{X})$ in a metric space $(\mathcal{X}, c)$, we use define $\Delta_{P,c} = \operatorname{diam}_c(\operatorname{supp}(P))$.

Theorem 13 *Let $(\mathcal{X}, c)$ be a metric space and suppose $\Delta := \max\left\{\Delta_{c,P_X}, \Delta_{c,P_G}\right\} < \infty$. For any $n \in \mathbb{N}_*$, let $\hat{P}_X$ and $\hat{P}_G$ denote the empirical distribution with n samples drawn i.i.d from P_X and P_G respectively. Let $s_X > d^*(P_X)$ and $s_G > d^*(P_G)$. For all $f : \mathbb{R} \to (-\infty, \infty]$ convex functions, $f(1) = 0$, $\lambda > 0$ and $\delta \in (0, 1)$, then with probability at least $1 - \delta$, we have*

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \leq \overline{W}_{c,\lambda \cdot f}(\hat{P}_X, P_G) + O\left(n^{-1/s_X} + \Delta\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right), \tag{9}$$

and if $f(x) = |x - 1|$ is chosen then

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \leq \overline{W}_{c,\lambda \cdot f}(\hat{P}_X, \hat{P}_G) + O\left(n^{-1/s_X} + n^{-1/s_G} + \Delta\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right). \tag{10}$$

(Proof in Appendix, Section A.2). First note that there is no requirement on G to be invertible and no restriction on λ. Second, there are the quantities s_X,s_G and Δ that are influenced by the distributions P_X and P_G. It is interesting to note that d^* is related to fractal dimensions [31] and thus relates the convergence of GANs to statistical geometry. If G is invertible in the above then the left hand side of both bounds becomes $\overline{W}_{c,\lambda \cdot f}(P_X, G)$ by Theorem 8. In general, $\hat{P}_X$ and $\hat{P}_G$ will not share the same support, in which case $D_f(\hat{P}_X, \hat{P}_G) = \infty$ – This would lead one to suspect the same from $\overline{W}_{c,\lambda \cdot f}(\hat{P}_X, \hat{P}_G)$, however this is not the case since

$$\overline{W}_{c,\lambda \cdot f}(\hat{P}_X, \hat{P}_G) \leq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{X}))} \left\{W_c((G \circ E)\#P_X, P_X) + \lambda D_f(E\#\hat{P}_X, \hat{P}_Z)\right\},$$

and so $E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))$ would be selected such that $E\#\hat{P}_X$ shares the support of $\hat{P}_Z$, resulting in a bounded value. We now show the relationship between $\overline{W}$ and W_c.

Theorem 14 *For any $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\lambda > 0$ and $f : \mathbb{R} \to (-\infty, \infty]$ convex function (with $f(1) = 0$) we have $\overline{W}_{c,\lambda \cdot f}(P_X, G) \leq W_c(P_X, P_G)$*

(Proof in Appendix, Section A.5). This suggests that in order to minimize $\overline{W}$, one can minimize W_c. Indeed, majority of the solvers are concerned with discrete distributions, which is exactly what is present on the right hand side of the generalization bounds: $\overline{W}_{c,\lambda \cdot f}(\hat{P}_X, \hat{P}_G)$

6 Discussion and Conclusion

This work is the first to prove a generalized primal-dual betweenship between GANs and Autoencoders. Our result elucidated the close performance between WAE and f-GANs. Furthermore, we explored the effect of weighting the reconstruction and regularization on the WAE objective, showing relationships to both f-divergences and Wasserstein metrics along with the impact on the duality relationship. This equivalence allowed us to prove generalization results, which to the best of our knowledge, are the first bounds given for Autoencoder models. The results imply that we can employ efficient (regularized) OT solvers to approximate upper bounds on the generalization bounds, which involve discrete distributions and thus are natural for such solvers.

The consequences of unifying two paradigms are plentiful, generalization bounds being an example. One line of extending and continuing the presented work is to explore the use of a general cost c (as opposed to a metric), invoking the generalized Wasserstein dual in the goal of forming a generalized GAN. Our paper provides a basis to unify Adversarial Networks and Autoencoders through a primal-dual relationship, and opens doors for the further unification of related models.

Acknowledgments

We would like to acknowledge anonymous reviewers and the Australian Research Council of Data61.

References

[1] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[5] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[6] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein autoencoders. *arXiv preprint arXiv:1711.01558*, 2017.

[7] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

[8] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[9] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[10] Aibek Alanov, Max Kochurov, Daniil Yashkov, and Dmitry Vetrov. Pairwise augmented gans with adversarial reconstruction loss. *arXiv preprint arXiv:1810.04920*, 2018.

[11] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

[12] Ke Li and Jitendra Malik. On the implicit assumptions of gans. *arXiv preprint arXiv:1811.12402*, 2018.

[13] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[14] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[15] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[16] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

[17] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

[18] Giorgio Patrini, Marcello Carioni, Patrick Forre, Samarth Bhargav, Max Welling, Rianne van den Berg, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. *arXiv preprint arXiv:1810.01118*, 2018.

[19] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017.

[20] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5545–5553, 2017.

[21] Shuang Liu and Kamalika Chaudhuri. The inductive bias of restricted f-gans. *arXiv preprint arXiv:1809.04542*, 2018.

[22] Farzan Farnia and David Tse. A convex duality framework for gans. In *Advances in Neural Information Processing Systems*, pages 5254–5263, 2018.

[23] Zhiming Zhou, Yuxuan Song, Lantao Yu, Hongwei Wang, Weinan Zhang, Zhihua Zhang, and Yong Yu. Understanding the effectiveness of lipschitz-continuity in generative adversarial nets. 2018.

[24] Richard Nock, Zac Cranko, Aditya K Menon, Lizhen Qu, and Robert C Williamson. f-gans in an information geometric nutshell. In *Advances in Neural Information Processing Systems*, pages 456–464, 2017.

[25] Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

[26] Lisa Borland. Ito-langevin equations within generalized thermostatistics. *Physics Letters A*, 245(1-2):67–72, 1998.

[27] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[28] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.

[29] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

[30] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

[31] Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

[32] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

A Appendix

A.1 Proof of Theorem 8

In order to prove the theorem, we make use of the dual form of the restricted variational form of an f-divergence:

Theorem 15 ([21], Theorem 3) *Let $f : \mathbb{R} \to (-\infty, \infty]$ denote a convex function with property $f(1) = 0$ and suppose H is a convex subset of $\mathscr{F}(\mathcal{X}, \mathbb{R})$ with the property that for any $h \in H$ and $b \in \mathbb{R}$, we have $h + b \in H$. Then for any $P, Q \in \mathscr{P}(\mathcal{X})$ we have*

$$\sup_{h \in H} \{\mathbb{E}_{x \sim P}[h(x)] - \mathbb{E}_{x \sim Q}[f^*(h(x))]\} = \inf_{P' \in \mathscr{P}(\mathcal{X})} \left\{ D_f(P', Q) + \sup_{h \in H} \{\mathbb{E}_P[h(x)] - E_{P'}[h(x)]\} \right\}$$

The goal is now to set $H = \mathcal{H}_c$ however there are some conditions of the above that we require

Lemma 16 *If c is a metric then $\mathcal{H}_c$ is convex and closed under addition.*

Proof Let $f \in \mathcal{H}_c$ and consider define $h = f + b$ for some $b \in \mathbb{R}$, we then have

$$\begin{aligned} |h(x) - h(y)| &= |f(x) + b - f(y) - b| \\ &= |f(x) - f(y)| \\ &\leq c(x, y) \end{aligned}$$

Consider some $\lambda \in [0, 1]$ and set $h(x) = \lambda \cdot f(x) + (1 - \lambda) \cdot g(x)$ for some $f, g \in \mathcal{H}_c$. We then have

$$\begin{aligned} |h(x) - h(y)| &= |\lambda \cdot f(x) + (1 - \lambda) \cdot g(x) - \lambda \cdot f(y) - (1 - \lambda) \cdot g(y)| \\ &= |\lambda \cdot (f(x) - f(y)) + (1 - \lambda) \cdot (g(x) - g(y))| \\ &\leq \lambda \cdot |f(x) - f(y)| + (1 - \lambda) \cdot |g(x) - g(y)| \\ &\leq \lambda \cdot c(x, y) + (1 - \lambda) \cdot c(x, y) \\ &= c(x, y) \end{aligned}$$

for all $x, y \in \mathcal{X}$. ■

We require a lemma regarding the decomposibility of G for f-divergences.

Lemma 17 *Let $G : \mathcal{Z} \to \mathcal{X}$ and let P, Q be two distributions over $\mathcal{Z}$. We have that*

$$D_f(G\#P, G\#Q) \leq D_f(P, Q),$$

with equality if G is invertible. Furthermore, if f is differentiable then we have equality for a weaker condition: for any $z, z' \in \mathcal{Z}, G(z) = G(z') \implies f'(\frac{dP}{dQ}(z)) = f'(\frac{dP}{dQ}(z'))$.

Proof By writing the variational form from [15] (Lemma 1), we have

$$\begin{aligned} D_f(G\#P, G\#Q) &= \sup_{h \in \mathscr{F}(\mathcal{X}, \mathbb{R})} \{\mathbb{E}_{x \sim G\#P}[h(x)] - \mathbb{E}_{x \sim G\#Q}[f^*(h(x))]\} \\ &= \sup_{h \in \mathscr{F}(\mathcal{X}, \mathbb{R})} \{\mathbb{E}_{z \sim P}[h(G(z))] - \mathbb{E}_{z \sim Q}[f^*(h(G(z)))]\} \\ &= \sup_{h \in \mathscr{F}(\mathcal{X}, \mathbb{R}) \circ G} \{\mathbb{E}_{z \sim P}[h(z)] - \mathbb{E}_{z \sim Q}[f^*(h(z))]\} \\ &\leq \sup_{h \in \mathscr{F}(\mathcal{Z}, \mathbb{R})} \{\mathbb{E}_{z \sim P}[h(z)] - \mathbb{E}_{z \sim Q}[f^*(h(z))]\} \\ &= D_f(P, Q), \end{aligned}$$

where we used the fact that $\mathscr{F}(\mathcal{X}, \mathbb{R}) \circ G \subseteq \mathscr{F}(\mathcal{Z}, \mathbb{R})$. If G is invertible then we applying the above with $G \leftarrow G^{-1}$, $P \leftarrow G\#P$ and $Q \leftarrow G\#Q$, we have

$$D_f(G^{-1}\#(G\#P), G^{-1}\#(G\#Q)) \leq D_f(G\#P, G\#Q),$$

which is just the reverse direction $D_f(P,Q) \leq D_f(G\#P, G\#Q)$, and so equality holds. Suppose now that f is differentiable then note that inequality holds when $f'(dP/dQ) \in \mathscr{F}(\mathcal{X},\mathbb{R}) \circ G$ (See proof of Lemma 1 in [15]), which is equivalent to asking if there exists a function $\varphi_f \in \mathscr{F}(\mathcal{X},\mathbb{R})$ such that

$$\varphi_f \circ G = f'\left(\frac{dP}{dQ}\right).$$

For any $z \in \mathcal{Z}$, we can construct φ_f to map $G(z)$ to $f'\left(\frac{dP}{dQ}\right)(z)$ and due to the condition in the lemma, we can guarantee φ_f will indeed be a function and thus exists. ■

We need a Lemma that will allow us to upper bound the Wasserstein distance.

Lemma 18 *For any $E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))$, $G \in \mathscr{F}(\mathcal{Z}, \mathcal{X})$ and $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we have*

$$W_c((G \circ E)\#P_X, P_X) \leq \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x).$$

Proof We quote a reparametrization result from [6] Theorem 1 that if G is deterministic then the Wasserstein distance can be reparametrized as

$$\begin{aligned} W_c(G\#(E\#P_X), P_X) &= \inf_{Q \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})) : Q\#P_X = E\#P_X} \int_{\mathcal{X}} \mathbb{E}_{z \sim Q(x)}[c(x, G(z))] dP_X(x) \qquad (11) \\ &\leq \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x). \end{aligned}$$

■

We also need a Lemma regarding the relationship between $\overline{W}$ and WAE.

Lemma 19 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a convex function with $f(1) = 0$, then we have*

$$\overline{W}_{c,\lambda \cdot f}(P_X, G) \leq \mathrm{WAE}_{c,\lambda \cdot D_f}(P_X, G).$$

Proof Consider the optimal encoder E^* from the f-WAE objective. Let $Q^* = E^*\#P_X$. We then have that

$$\overline{W}_{c,\lambda \cdot f}(P_X, G) = W_c(P_X, G\#Q^*) + \lambda \cdot D_f(Q^*, P_Z).$$

Let $\pi \in \Pi(P_X, E\#Q^*)$ be the optimal coupling under the metric c. By the Gluing lemma [14], one can construct a triple (X, Y, Z) where $(X, Y) \sim \pi$, $Z \sim Q^*$ and $Y = G(Z)$ almost surely. Let π' be the distribution over (Y, Z) and consider the conditional distribution over Z given Y, associated with $E_{\pi'} \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))$. We have $E_{\pi'}\#P_X = Q^*$ and so we have

$$\begin{aligned} \mathrm{WAE}_{c,\lambda \cdot D_f}(P_X, G) &\leq \int_{\mathcal{X}} \mathbb{E}_{z \sim E_{\pi'}(y)}[c(x, G(z))] dP_X + D_f(E_{\pi'}\#P_X, P_Z) \\ &= \int_{\mathcal{X}} \mathbb{E}_{z \sim E_{\pi'}(y)}[c(x, G(z))] dP_X + D_f(Q^*, P_Z) \\ &= \int_{\mathcal{X} \times \mathcal{X}} [c(x, y)] d\pi'(x, y) + D_f(Q^*, P_Z) \\ &= W_c(P_X, G\#Q^*) + \lambda \cdot D_f(Q^*, P_Z). \\ &= \overline{W}_{c,\lambda \cdot f}(P_X, G). \end{aligned}$$

■

Finally, we need a lemma to justify reparametrizations.

Lemma 20 *If $G : \mathcal{Z} \to \mathcal{X}$ is invertible then for any $P' \in \mathscr{P}(\mathcal{X})$ such that $P' \ll P_G$, then there exists an $E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))$ such that $P' = G\#E\#P_X$.*

Proof From the assumption, we have $\text{Supp}(P') \subseteq \text{Supp}(P_G) \subseteq \text{Im}(G)$ and so by invertibility of G, we can set $Q = G^{-1}\#P'$ and construct a conditional distribution E (between marginals Q and P_X) to get $Q = E\#P_X$, hence $P' = G\#E\#P_X$. ■

We are now ready to prove the theorem. Set $H = \mathcal{H}_c$ (the set of 1-Lipschitz functions) and note that λf is a convex function satisfying $\lambda f(1) = 0$ and so substituting $f \leftarrow \lambda f$, we get that $D_{\lambda f}(\cdot,\cdot) = \lambda D_f(\cdot,\cdot)$. Hence, we have

$$\begin{aligned}
\text{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) &= \sup_{h \in H_c} \{\mathbb{E}_{x\sim P_X}[h(x)] - \mathbb{E}_{x\sim P_G}[(\lambda f)^\star(h(x))]\} \\
&= \inf_{P' \in \mathscr{P}(\mathcal{X})} \{\lambda D_f(P', P_G) + W_c(P', P_X)\} \\
&= \inf_{P' \in \mathscr{P}(\mathcal{X}): P' << P_g} \{\lambda D_f(P', P_G) + W_c(P', P_X)\} \\
&= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{\lambda D_f((G\circ E)\#P_X, G\#P_Z) + W_c((G\circ E)\#P_X, P_X)\} \qquad (12) \\
&\overset{(*)}{\leq} \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{\lambda D_f(E\#P_X, P_Z) + W_c((G\circ E)\#P_X, P_X)\} \\
&= \overline{W}_{c,\lambda\cdot f}(P_X, G) \\
&\leq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \left\{\int_{\mathcal{X}} \mathbb{E}_{z\sim E(x)}[c(x, G(z))] dP_X(x) + \lambda D_f(E\#P_X, P_Z)\right\} \\
&= \text{WAE}_{c,\lambda\cdot D_f}(P_X, G), \qquad (13)
\end{aligned}$$

where (12) is an equality when G is invertible from Lemma 20 and $(*)$ is $=$ if G satisfies the requirement of Lemma 17. To prove the final inequality, note that if E^* satisfies the condition of the Theorem then

$$\begin{aligned}
\overline{W}_{c,\lambda\cdot f}(P_X, G) &= W_c((G\circ E^*)\#P_X, P_X) + \lambda D_f(E^*\#P_X, P_Z) \\
&= W_c(G\#(E^*\#P_X), P_X) \\
&= W_c(P_G, P_X). \qquad (14)
\end{aligned}$$

Next, notice that

$$\begin{aligned}
&\text{WAE}_{c,\lambda\cdot D_f}(P_X, G) \\
&= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \left\{\int_{\mathcal{X}} \mathbb{E}_{z\sim E(x)}[c(x, G(z))] dP_X(x) + \lambda D_f(E\#P_X, P_Z)\right\} \\
&\leq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})): E\#P_X = P_Z} \left\{\int_{\mathcal{X}} \mathbb{E}_{z\sim E(x)}[c(x, G(z))] dP_X(x) + \lambda D_f(E\#P_X, P_Z)\right\} \\
&\leq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})): E\#P_X = P_Z} \left\{\int_{\mathcal{X}} \mathbb{E}_{z\sim E(x)}[c(x, G(z))] dP_X(x)\right\} \\
&= W_c(P_X, P_G) \qquad (15) \\
&= \overline{W}_{c,\lambda\cdot f}(P_X, G), \qquad (16)
\end{aligned}$$

where (15) follows from the reparametrized Wasserstein distance from [6] (Theorem 1), which we used in (11) and the final step follows from (14). Combining $\text{WAE}_{c,\lambda\cdot D_f}(P_X, G) \leq \overline{W}_{c,\lambda\cdot f}(P_X, G)$ with $\text{WAE}_{c,\lambda\cdot D_f}(P_X, G) \geq \overline{W}_{c,\lambda\cdot f}(P_X, G)$ (from 13) yields equality and concludes the proof.

A.2 Proof of Theorem 13

We first prove a lemma that will apply to both cases. Recalling that for any metric space $(\mathcal{X}, c)$ and $P \in \mathscr{P}(\mathcal{X})$ we define $\Delta_{P,c} = \text{diam}_c(\text{supp}(P))$.

Lemma 21 *Let $(\mathcal{X}, c)$ be a metric space. For any $P \in \mathscr{P}(\mathcal{X})$, suppose $\Delta_{P,c} < \infty$ and let $\hat{P}$ denote the empirical distribution after drawing n i.i.d samples for some $n \in \mathbb{N}_*$. If $s > d^*(P)$, then we have*

$$\text{IPM}_{\mathcal{H}_c}(P, \hat{P}) \leq O(n^{-1/s}) + \frac{\Delta_{P,c}}{2}\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)}$$

Proof We appeal to McDiarmind's Inequality and use a standard method, as shown in [32], to bound the quantity.

Theorem 22 (McDiarmind's Inequality) *Let $X_1, \dots, X_n$ be n independent random variables and consider a function $\Phi : \mathcal{X}^n \to \mathbb{R}$ such that there exists constants $c_i > 0$ (for $i = 1, \dots, n$) with*

$$\sup_{x_1,\dots,x_n,x_i'} |\Phi(x_1, \dots, x_n) - \Phi(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i.$$

Then for any $t > 0$, we have

$$\Pr\left[\Phi(X_1, \dots, X_n) - \mathbb{E}\left[\Phi(X_1, \dots, X_n)\right] \geq t\right] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right)$$

Let $\mathcal{F} = \mathcal{H}_c$ then let

$$\Phi(S) = \text{IPM}_{\mathcal{H}_c}(P, \hat{P}).$$

Noting that

$$\begin{aligned}
|\Phi(x_1, \dots, x_n) - \Phi(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| &\leq \frac{1}{n}\left|f(x_i) - f(x_i')\right| \\
&\leq \frac{1}{n} \cdot c(x_i, x_i') \\
&\leq \frac{\Delta_{P,c}}{n},
\end{aligned}$$

where the first inequality follows as each f is 1-Lipschitz and the second follows from the fact that each $x, x' \in \text{supp}(P)$. This allows us to set $c_i = \Delta/n$ for all $i = 1, \dots, n$. Now applying McDiarmind's inequality with $t = \Delta_{P,c}/2\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)}$ yields (for a sample $S \sim P^n$)

$$\begin{aligned}
\Pr\left[\Phi(S) - \mathbb{E}\Phi(S) \geq \frac{\Delta_{P,c}}{2}\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)}\right] &\leq \delta \\
\Pr\left[\Phi(S) - \mathbb{E}\Phi(S) \leq \frac{\Delta_{P,c}}{2}\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)}\right] &\geq 1 - \delta,
\end{aligned}$$

and thus

$$\Phi(S) \leq \mathbb{E}\Phi(S) + \frac{\Delta_{P,c}}{2}\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)}.$$

Noting that $\mathbb{E}\Phi(S) = \mathbb{E}[W_c(P, \hat{P})]$ (from Lemma 4), we appeal to a case of Theorem 1 in [30] where $p = 1$, which tells us that if $s > d^*(P)$ then $\mathbb{E}[W_c(P, \hat{P})] = O(n^{-1/s})$. Since this is the requirement in the lemma, the proof concludes. ■

We will make use of this lemma for both P_X and P_G and use Δ for both cases since $\Delta \geq \Delta_{P_X,c}$ and $\Delta \geq \Delta_{P_G,c}$. For the general case of any f, let (abusing notation) $G = \text{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c)$ and $\hat{G}$ denote the empirical counterpart with n samples, and let $h^1, h^2 \in \mathcal{H}_c$ denote their witness functions. We then have

$$\begin{aligned}
&G - \hat{G} \\
&= \sup_{h \in \mathcal{H}_c} \left\{\mathbb{E}_{x \sim P_X}[h(x)] - \mathbb{E}_{x \sim P_G}[(\lambda f)^\star(h(x))]\right\} - \sup_{h \in \mathcal{H}_c} \left\{\mathbb{E}_{x \sim \hat{P}_X}[h(x)] - \mathbb{E}_{x \sim P_G}[(\lambda f)^\star(h(x))]\right\} \\
&= \mathbb{E}_{x \sim P_X}[h^1(x)] - \mathbb{E}_{x \sim P_G}[(\lambda f)^\star(h^1(x))] - \mathbb{E}_{x \sim \hat{P}_X}[h^2(x)] + \mathbb{E}_{x \sim P_G}[(\lambda f)^\star(h^2(x))] \\
&\leq \mathbb{E}_{x \sim P_X}[h^1(x)] - \mathbb{E}_{x \sim \hat{P}_X}[h^1(x)] + \mathbb{E}_{x \sim P_G}[(\lambda f)^\star(h^1(x))] - \mathbb{E}_{x \sim P_G}[(\lambda f)^\star(h^1(x))] \\
&= \mathbb{E}_{x \sim P_X}[h^1(x)] - \mathbb{E}_{x \sim \hat{P}_X}[h^1(x)] \\
&\leq \sup_{h \in \mathcal{H}_c} \left\{\mathbb{E}_{x \sim P_X}[h(x)] - \mathbb{E}_{x \sim \hat{P}_X}[h(x)]\right\} \\
&= \text{IPM}_{\mathcal{H}_c}(P_X, \hat{P}_X) \\
&\leq O(n^{-1/s_X}) + \frac{\Delta}{2}\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)},
\end{aligned}$$

where the last step is an application of Lemma 21. Applying Theorem 8, we get $\hat{G} \leq \overline{W}_{c,\lambda\cdot f}$ and rearrangement of the above shows the first bound. For the case of $f(x) = |x-1|$, note that if $\mathcal{F} \subseteq \mathscr{F}(\mathcal{X}, \mathbb{R})$ is such that $-\mathcal{F} = \mathcal{F}$, then $\text{IPM}_{\mathcal{F}}$ is a pseudo-metric and satisfies the triangle inequality, which allows us to have

$$\begin{aligned}\text{IPM}_{\mathcal{F}}(P_X, P_G) &\leq \text{IPM}_{\mathcal{F}}(P_X, \hat{P}_X) + \text{IPM}_{\mathcal{F}}(\hat{P}_X, P_G) \\ &\leq \text{IPM}_{\mathcal{F}}(P_X, \hat{P}_X) + \text{IPM}_{\mathcal{F}}(P_G, \hat{P}_G) + \text{IPM}_{\mathcal{F}}(\hat{P}_X, \hat{P}_G). \qquad (17)\end{aligned}$$

Next, we set $\mathcal{F} = \mathcal{F}_{c,\lambda}$, and noting that $\mathcal{F}_{c,\lambda} \subseteq \mathcal{H}_c$, we have

$$\begin{aligned}\text{IPM}_{\mathcal{F}_{c,\lambda}}(P_X, P_G) &\leq \text{IPM}_{\mathcal{F}_{c,\lambda}}(P_X, \hat{P}_X) + \text{IPM}_{\mathcal{F}_{c,\lambda}}(P_G, \hat{P}_G) + \text{IPM}_{\mathcal{F}_{c,\lambda}}(\hat{P}_X, \hat{P}_G) \\ &\leq \text{IPM}_{\mathcal{H}_c}(P_X, \hat{P}_X) + \text{IPM}_{\mathcal{H}_c}(P_G, \hat{P}_G) + \text{IPM}_{\mathcal{H}_c}(\hat{P}_X, \hat{P}_G) \\ &\leq \text{IPM}_{\mathcal{H}_c}(\hat{P}_X, \hat{P}_G) + O(n^{-1/s_X} + n^{-1/s_G}) + \Delta\sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}, \qquad (18)\end{aligned}$$

where the final inequality is an application of Lemma 21 like before. However since we use McDiarmind's inequality twice, we set $\delta \leftarrow \delta/2$ and use union bound to have the above inequality with probability $1-\delta$. The final step is to note that when $f(x) = |x-1|$ then for any $\lambda > 0$,

$$(\lambda f)^\star(x) = \begin{cases} x & x \leq \lambda \\ \infty & x > \lambda \end{cases}$$

and so we have

$$\begin{aligned}\text{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) &= \sup_{h \in \mathcal{H}_c} \left\{ \mathbb{E}_{x\sim P_X}[h(x)] - \mathbb{E}_{x\sim P_G}[(\lambda f)^\star(h(x))] \right\} \\ &= \sup_{h \in \mathcal{H}_c : |h| \leq \lambda} \left\{ \mathbb{E}_{x\sim P_X}[h(x)] - \mathbb{E}_{x\sim P_G}[h(x)] \right\} \\ &= \sup_{h \in \mathcal{F}_{c,\lambda}} \left\{ \mathbb{E}_{x\sim P_X}[h(x)] - \mathbb{E}_{x\sim P_G}[h(x)] \right\} \\ &= \text{IPM}_{\mathcal{F}_{c,\lambda}}(P_X, P_G).\end{aligned}$$

By Theorem 8, we have $\text{IPM}_{\mathcal{F}_{c,\lambda}}(\hat{P}_X, \hat{P}_G) = \text{GAN}_{\lambda f}(\hat{P}_X, G; \mathcal{H}_c) \leq \overline{W}_{c,\lambda\cdot f}(\hat{P}_X, G)$ where $\text{GAN}_{\lambda f}(\hat{P}_X, G; \mathcal{H}_c)$ is the objective with $\hat{P}_X$ and $\hat{P}_G$. Putting this together with (18), we get

$$\begin{aligned}\text{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) &= \text{IPM}_{\mathcal{F}_{c,\lambda}}(P_X, P_G) \\ &\leq \text{IPM}_{\mathcal{H}_c}(\hat{P}_X, \hat{P}_G) + O(n^{-1/s}) + \Delta\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)} \\ &= \text{GAN}_{\lambda f}(\hat{P}_X, G; \mathcal{H}_c) + O(n^{-1/s}) + \Delta\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)} \\ &\leq \overline{W}_{c,\lambda\cdot f}(\hat{P}_X, G) + O(n^{-1/s_X} + n^{-1/s_G}) + \Delta\sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}.\end{aligned}$$

A.3 Proof of Theorem 9

First, using Theorem 8 and the fact that the f-GAN objective is a lower bound to D_f, we have that

$$\begin{aligned}\overline{W}_{\gamma\cdot c,f}(P_X, G) &= \text{GAN}_f(P_X, G, \mathcal{H}_{\gamma c}) \\ &\leq D_f.\end{aligned}$$

It is known that $f'(dP_X/dP_G)$ is the maximizer of $L(h) = \mathbb{E}_{x\sim P_X}[h(x)] - \mathbb{E}_{x\sim P_G}[f^\star(h(x))]$ [15], and so the proof concludes by showing that $f'(dP_X/dP_G) \in \mathcal{H}_{\gamma^*\cdot c}$. Note that $h \in \mathcal{H}_{\gamma\cdot c}$ if and only if for all $x, x' \in \mathcal{X}, x \neq x'$

$$\begin{aligned}|h(x) - h(x')| &\leq \gamma \cdot \delta_{x-x'}(0) \\ &= \gamma\end{aligned}$$

and so the 1-Lipschitz functions are those that are bounded by their maximum and minimum value by γ. For any $x, x' \in \mathcal{X}, x \neq x'$ we have

$$\begin{aligned}\left|f'\left(\frac{dP_X}{dP_G}\right)(x) - f'\left(\frac{dP_X}{dP_G}\right)(x')\right| &= \gamma^* \left|f'\left(\frac{dP_X}{dP_G}\right)(x) - f'\left(0\right)\right| \\ &\leq \gamma,\end{aligned}$$

and thus $f'(dP_X/dP_G) \in \mathcal{H}_{\gamma \cdot c}$.

A.4 Proof of Theorem 10

First note that

$$\begin{aligned}\text{WAE}_{c,\lambda\cdot f}(P_X, P_G) &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z\sim E(x)}[c(x, G(z))] dP_X(x) + \lambda \cdot D_f(E\#P_X, P_Z) \right\} \\ &\leq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})) : E\#P_X = P_Z} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z\sim E(x)}[c(x, G(z))] dP_X(x) \right\} \\ &= W_c(P_X, P_G),\end{aligned}$$

where the last equality holds from [6] Theorem 1. Thus we have the chain of inequalities for all λ and $f : \mathbb{R} \to (-\infty, \infty]$ (convex with $f(1) = 0$)

$$\text{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \leq \overline{W}_{c,\lambda\cdot}(P_X, P_G) \leq \text{WAE}_{c,\lambda\cdot f}(P_X, P_G) \leq W_c(P_X, P_G).$$

We now show the opposite direction, which will conclude the proof.

Lemma 23 *For any metric c and $f : \mathbb{R} \to (-\infty, \infty]$ convex function with $f(1) = 0$, if*

$$\lambda \geq \lambda^* = \sup_{P' \in \mathscr{P}(\mathcal{X})} \left(W_c(P', P_G)/D_f(P', P_G)\right),$$

then we have

$$\text{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \geq W_c(P_X, P_G)$$

Proof First noting that $\lambda \geq \sup_{P' \in \mathscr{P}(\mathcal{X})} (W_c(P', P_G)/D_f(P', P_G))$, for all $P' \in \mathscr{P}(\mathcal{X})$, we have

$$\lambda D_f(P', P_G) - W_c(P', P_G) \geq 0.$$

Let $\tilde{\mathcal{Z}} = \mathcal{X}$,$\tilde{G} = \text{Id}$, $P_{\tilde{\mathcal{Z}}} = P_G$ and noting that $\tilde{G}$ is invertible, we can apply Theorem 8 to get

$$\begin{aligned}\text{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) &= \overline{W}_{c,\lambda\cdot f}(P_X, \tilde{G}\#P_{\tilde{\mathcal{Z}}}) \\ &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathcal{P}(\mathcal{X}))} \{W_c(E\#P_X, P_X) + \lambda D_f(E\#P_X, P_G)\} \\ &\geq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathcal{P}(\mathcal{X}))} \{W_c(P_X, P_G) - W_c(E\#P_X, P_G) + \lambda D_f(E\#P_X, P_G)\} \\ &\geq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathcal{P}(\mathcal{X}))} \{W_c(P_X, P_G)\} \\ &= W_c(P_X, P_G).\end{aligned}$$

■

A.5 Proof of Theorem 14

We have

$$\begin{aligned}\overline{W}_{c,\lambda\cdot f}(P_X, G) &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{W_c(P_X, (G \circ E)\#P_X) + \lambda D_f(E\#P_X, P_Z)\} \\ &\leq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})) : E\#P_X = P_Z} \{W_c(P_X, (G \circ E)\#P_X) + \lambda D_f(E\#P_X, P_Z)\} \\ &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})) : E\#P_X = P_Z} \{W_c(P_X, (G \circ E)\#P_X)\} \\ &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})) : E\#P_X = P_Z} \{W_c(P_X, P_G)\} \\ &= W_c(P_X, P_G).\end{aligned}$$

References

[1] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[5] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[6] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein autoencoders. *arXiv preprint arXiv:1711.01558*, 2017.

[7] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

[8] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[9] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[10] Aibek Alanov, Max Kochurov, Daniil Yashkov, and Dmitry Vetrov. Pairwise augmented gans with adversarial reconstruction loss. *arXiv preprint arXiv:1810.04920*, 2018.

[11] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

[12] Ke Li and Jitendra Malik. On the implicit assumptions of gans. *arXiv preprint arXiv:1811.12402*, 2018.

[13] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[14] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[15] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[16] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

[17] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

[18] Giorgio Patrini, Marcello Carioni, Patrick Forre, Samarth Bhargav, Max Welling, Rianne van den Berg, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. *arXiv preprint arXiv:1810.01118*, 2018.

[19] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017.

[20] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5545–5553, 2017.

[21] Shuang Liu and Kamalika Chaudhuri. The inductive bias of restricted f-gans. *arXiv preprint arXiv:1809.04542*, 2018.

[22] Farzan Farnia and David Tse. A convex duality framework for gans. In *Advances in Neural Information Processing Systems*, pages 5254–5263, 2018.

[23] Zhiming Zhou, Yuxuan Song, Lantao Yu, Hongwei Wang, Weinan Zhang, Zhihua Zhang, and Yong Yu. Understanding the effectiveness of lipschitz-continuity in generative adversarial nets. 2018.

[24] Richard Nock, Zac Cranko, Aditya K Menon, Lizhen Qu, and Robert C Williamson. f-gans in an information geometric nutshell. In *Advances in Neural Information Processing Systems*, pages 456–464, 2017.

[25] Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

[26] Lisa Borland. Ito-langevin equations within generalized thermostatistics. *Physics Letters A*, 245(1-2):67–72, 1998.

[27] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[28] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.

[29] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

[30] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

[31] Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

[32] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Chapter 4

Local Differential Privacy for Sampling

One of the significant drawbacks of Generative Adversarial Networks (GANs) is the inability to derive a closed-form density of the final generative distribution learned. Not only does it withhold our ability to understand the capacity of GANs, but it restricts us from having guarantees from the perspective of privacy, which is a topical application of data generation that requires control on the density. Inspired by the discriminator in GANs, this chapter will build a boosting-based approach for learning densities whose samples comply with local differential privacy. Such a method will serve helpful to privatize any mechanism that employs the sampler by way of post-processing. Theoretical guarantees such as convergence and mode-coverage will also be proven along with experimental results against state-of-the-art privacy generation methods.

4.0.1 Errata

We remark a minor mistake in the presentation of Theorem 6 which should be stated as

Theorem 2 *We have $\Delta(Q) \leq \varepsilon/2, \forall Q \in \mathcal{M}_\varepsilon$, and if* IPB-DE *is in the high boosting regime, then*

$$\Delta(Q_T) \quad \geq \quad \frac{\varepsilon}{2} \cdot \left\{ \frac{\gamma_P + \overline{\Gamma}(\gamma_Q)}{2} \cdot (1 - \theta_T(\varepsilon)) \right\}, \tag{4.1}$$

where $\overline{\Gamma}(z) = \Gamma(z)/\log 2$.

This does not change any other aspect of the publication since indeed as $\gamma_P, \gamma_Q \to 1$, we still have $\Delta(Q_T) \geq (\varepsilon/2) \cdot (1 - \theta_T(\varepsilon))$.

Local Differential Privacy for Sampling

Hisham Husain[○,‡] **Borja Balle**[†] **Zac Cranko**[○,‡] **Richard Nock**[‡,○]

Abstract

Differential privacy (DP) is a leading privacy protection focused by design on individual privacy. In the local model of DP, strong privacy is achieved by privatizing each user's individual data before sending it to an untrusted aggregator for analysis. While in recent years local DP has been adopted for practical deployments, most research in this area focuses on problems where each individual holds a single data record. In many problems of practical interest this assumption is unrealistic since nowadays most user-owned devices collect large quantities of data (e.g. pictures, text messages, time series). We propose to model this scenario by assuming each individual holds a distribution over the space of data records, and develop novel local DP methods to sample privately from these distributions. Our main contribution is a boosting-based density estimation algorithm for learning samplers that generate synthetic data while protecting the underlying distribution of each user with local DP. We give approximation guarantees quantifying how well these samplers approximate the true distribution. Experimental results against DP kernel density estimation and DP GANs displays the quality of our results.

1 Introduction

Over the past decade, differential privacy (DP) has evolved as the leading statistical protection model for individuals' data (Dwork and Roth, 2014). The basis of DP is that a mechanism is private whenever its output provides insufficient information to distinguish between two potential input datasets that differ on a single individual. In doing so, it guarantees plausible deniability regarding the presence of an individual in the input of the mechanism. Despite the popularity of DP, one shortcoming of the standard definition is the assumption of a *trusted* curator who has access to the full dataset of individuals. One way to get around this is to have individuals run their data through a DP mechanism at the local level before sending it for processing, ensuring that the curator only gets access to privatized data. This approach is called the *local model* of differential privacy (Raskhodnikova et al., 2008). It requires considerably weaker trust assumptions than the curator model, and was in fact the basis of the first large-scale deployments of DP by Apple (Differential privacy team, Apple, 2017) and Google (Erlingsson et al., 2014).

The interest in the local model has spurred research into local DP protocols for a number of practical tasks (see (Cormode et al., 2018) and references therein), as well as the search for intermediate privacy models achieving a compromise between the local and curator DP Bittau et al. (2017). However, while most of this research focuses, often implicitly, on the setting where each individual owns a *single* data record, a growing number of applications involve one individual contributing *multiple* data records. Examples include problems where the data evolves over time, as well as settings where locally each individual owns a whole dataset containing, e.g., pictures, text messages or historical device usage information.

In this paper we investigate a method to leverage sensitive user-level datasets in local DP protocols by constructing *locally private samplers* which can release synthetic data points from the distribution of the underlying dataset. Our framework accommodates local datasets of arbitrary sizes by modelling an individual's private data as a probability distribution – this is also applicable in situations where the dataset does not exist *per se* but an algorithm can sample from it by, e.g., interacting with the user. We formalize the problem by (1) introducing the notion of mollifier – a collection of valid distributions from which one can obtain samples with a desired privacy level; and, (2) cast the goal of learning a private sampler as the problem of computing the information-geometric projection of a private distribution onto a given mollifier – a process we call mollification. Our main contribution is an efficient approximate mollification algorithm based on recent advances in boosted density estimation (Cranko and Nock, 2019). In contrast with

[○] The Australian National University. [‡] Data61. [†] Currently at DeepMind.

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

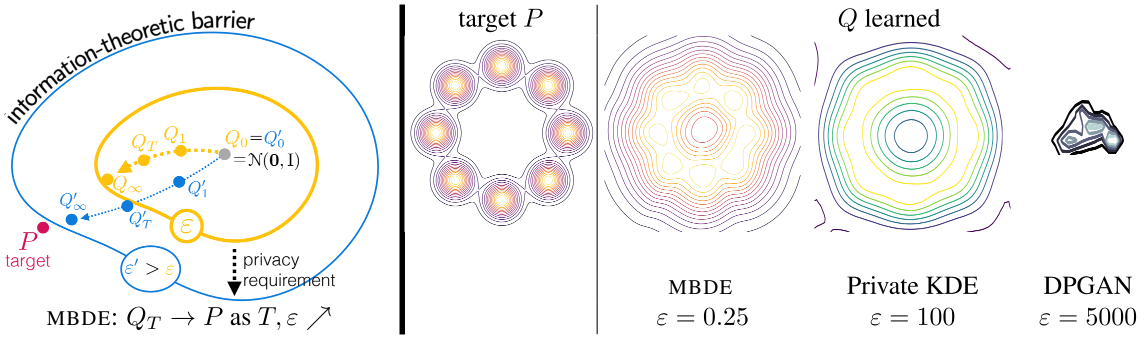


Figure 1: *Left*: Our method is guaranteed to get a Q_T that converges to P as the privacy constraint is relaxed and the number of boosting iterations increases (under a weak learning assumption). *Right*: Our method vs private KDE (Aldà and Rubinstein, 2017) and DPGAN (Xie et al., 2018) on a ring Gaussian mixture (see Section 5, $m = k = 10000$). Remark that the GAN is subject to mode collapse.

naive solutions we discuss below, our algorithm works on arbitrary data, including continuous unbounded domains. This algorithm comes with convergence rate guarantees in the classical boosting model, that is, under lightweight assumptions on the distribution iterates used in the mollification process. Under slightly stronger assumptions, we are able to show guaranteed approximation with respect to the *optimal* distribution in the mollifier. As the privacy constraint is relaxed, we get better approximation guarantees with respect to the target distribution itself. This is illustrated in Figure 1 (left). Last but not least, we provide guarantees in terms of capturing the modes of the target distribution, which is a prominent problem in generative approaches (Figure 1, right).

The rest of this paper is organized as follows. Section 2 introduces locally private sampling, mollifiers and their relationships. Section 3 introduces our algorithm that learns a density in a mollifier and shows several approximation properties in the boosting model. Section 4 summarizes related work, Section 5 presents and discusses experiments.

2 Private sampling and mollifiers

We now proceed to formalize the task of sampling from a private distribution in the local DP model. Then introduce the concept of mollification which solves this problem by first projecting the distribution into a carefully constructed set and releases a sample from the resulting projection.

Locally private sampling Suppose a user holds a private probability distribution $P \in \mathcal{D}(\mathcal{X})$ over some domain $\mathcal{X}$ and wants to release a sample from P while preserving the their privacy. We introduce a user-defined parameter, $\varepsilon > 0$, which represents a privacy budget – smaller ε correspond to a stronger privacy demand. An *ε-private sampler* is a randomized mapping $A : \mathcal{D}(\mathcal{X}) \to \mathcal{X}$ such that for any $x \in \mathcal{X}$ and any two distributions $P, P' \in \mathcal{D}(\mathcal{X})$ we have

$$\frac{\Pr[A(P) = x]}{\Pr[A(P') = x]} \leq \exp(\varepsilon) \ . \quad (1)$$

This is the same as saying that A is an *ε-locally differentially private* (LDP) mechanism[1] with inputs in $\mathcal{D}(\mathcal{X})$ and outputs in $\mathcal{X}$, which allows a user to release a privatized sample from their distribution P. Note that when the user has a dataset with records from $\mathcal{X}$ we can take P to be the empirical distribution over the sample.

A simple way to construct ε-private samplers given an ε-LDP mechanism $R : \mathcal{X} \to \mathcal{X}$ is a follows: take a sample $x_0 \sim P$ and then release the output of $R(x_0)$. This construction, which we denote by A_R, is appealing because it enables us to leverage any of the many local randomizers R that have been proposed in the literature, including, e.g., randomized response for discrete input spaces, and the Laplace mechanism with inputs on a bounded real interval. On the other hand, this generic construction is limited by the fact that A_R only accesses the private distribution P through a *single* sampling operation and has no information about the global shape of P.

Mollifiers To address this issue we propose to build private samplers by first projecting the distribution P onto a given mollifier and then releasing one sample from the projected distribution.

Definition 1 *Let $\mathcal{M} \subset \mathcal{D}(\mathcal{X})$ be a set of distributions*[2] *and $\varepsilon > 0$. We say $\mathcal{M}$ is an ε-**mollifier** iff*

$$Q(x) \quad \leq \quad \exp(\varepsilon) \cdot Q'(x), \forall Q, Q' \in \mathcal{M}, \forall x \in \mathcal{X}. \quad (2)$$

[1] A randomized mechanism $R : \mathcal{Y} \to \mathcal{Z}$ is ε-LDP if $\Pr[R(y) = z] \leq e^\varepsilon \Pr[R(y') = z]$ for all y, y', z.

[2] For the sake of simplicity (and at the expense of slight abuses of language) we use the same notation for distributions and their densities with respect to some base measure throughout the paper.

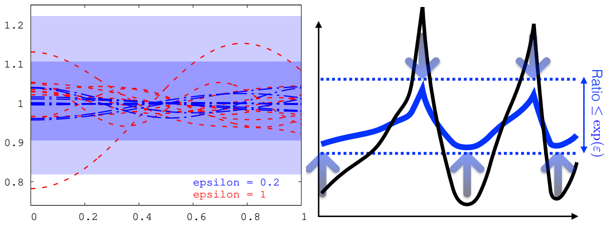


Figure 2: Left: example of mollifiers for two values of ε, $\varepsilon = 1$ (red curves) or $\varepsilon = 0.2$ (blue curves), with $\mathcal{X} = [0, 1]$. For that latter case, we also indicate in light blue the *necessary* range of values to satisfy (2), and in dark blue a *sufficient* range that allows to satisfy (2). Right: schematic depiction of how one can transform any set of finite densities in an ε-mollifier without losing the modes and keeping derivatives up to a positive constant scaling.

For example, a singleton $\mathcal{M} = \{Q\}$ is a 0-mollifier. Intuitively, these mollifiers consist of distributions which are all close to each other with respect to the divergence used to define local DP. Figure 2 (left) features examples of mollifiers with densities supported in $\mathcal{X} = [0, 1]$. Two ranges indicated in blue depict necessary or sufficient conditions on the overall range of a set of densities to be a mollifier. For the necessary part, we note that any continuous density must have 1 in its range of values (otherwise its total mass cannot be unit), so if it belongs to an ε-mollifier, its maximal value cannot be $\geq e^{\varepsilon}$ and its minimal value cannot be $\leq e^{-\varepsilon}$. We end up with the range in light blue, in which any ε-mollifier has to fit. For the sufficiency part, we indicate in dark blue a possible range of values, $[e^{-\varepsilon/2}, e^{\varepsilon/2}]$, which gives a sufficient condition for the range of all elements in a set $\mathcal{M}$ for this set to be an ε-mollifier.

Mollifiers play a central role in the theory developed in this paper, and they might also be of independent interest in the field of differential privacy. Before we show how they relate to private samplers, we first discuss some properties.

Constructing mollifiers Taking the convex hull of a mollifier produces a new mollifier. That is, given an ε-mollifier[3] $\mathcal{M} = \{Q_1, \ldots, Q_m\}$, the convex hull

$$\mathsf{cvx}(\mathcal{M}) = \left\{ \sum \alpha_i Q_i \ : \ \alpha_i \geq 0, \ \sum \alpha_i = 1 \right\} \qquad (3)$$

is again an ε-mollifier. We call $\mathsf{cvx}(\mathcal{M})$ the mollifier *generated* by $\mathcal{M}$. A mollifier is *convex* if $\mathsf{cvx}(\mathcal{M}) = \mathcal{M}$. Of particular interest are the convex ε-mollifiers generated by a ε-LDP mechanism R on some finite set $\mathcal{X}$, obtained as $\mathcal{M}_R := \mathsf{cvx}(\{R(x) : x \in \mathcal{X}\})$. This mollifier is in fact equivalent to the range of distributions of the naive sampler A_R, in the sense that

$$\mathcal{M}_R = \{\mathsf{Law}(A_R(P)) \ : \ P \in \mathcal{D}(\mathcal{X})\} \ , \qquad (4)$$

where $\mathsf{Law}(A_R(P))$ denotes the distribution of the output of $A_R(P)$ which can be written as the mixture $\mathsf{Law}(A_R(P)) = \sum_{x \in \mathcal{X}} P(x) \cdot \mathsf{Law}(R(x))$. This construction can be directly extended to bounded $\mathcal{X} \subset \mathbb{R}^d$, but for unbounded domains it is unclear how to proceed as most known LDP mechanisms R require bounded sensitivity.

Another way to obtain mollifiers starting from a reference distribution Q_0 is to consider the set of all distributions which are close to Q_0. In particular, we define the ε-mollifier *relative* to Q_0, denoted $\mathcal{M}_{\varepsilon, Q_0}$, to be the set of all distributions Q such that

$$\sup_x \max \left\{ \frac{Q_0(x)}{Q(x)}, \frac{Q(x)}{Q_0(x)} \right\} \leq \exp(\varepsilon/2) \ . \qquad (5)$$

To verify that this is indeed an ε-mollifier just note that for any $Q, Q' \in \mathcal{M}_{\varepsilon, Q_0}$ we have

$$\frac{Q(x)}{Q'(x)} = \frac{Q(x)}{Q_0(x)} \frac{Q_0(x)}{Q'(x)} \leq \exp(\varepsilon) \ . \qquad (6)$$

Whenever Q_0 is clear from the context we shall omit if from our notation.

Unlike with finitely generated mollifiers, relative mollifiers are not easy to parametrize in closed form. This is due to the “non-parametric” nature of the definition of $\mathcal{M}_{\varepsilon, Q_0}$, as opposed to the parametric definition of $\mathsf{cvx}(\{Q_1, \ldots, Q_m\})$. However, from the point of view of the problem we consider in the sequel – namely, finding the closest projection of a distribution onto a given mollifier – we shall see that relative mollifiers are also computationally tractable. In particular, we show that finding such projections when $\mathcal{X}$ is finite can be done in closed-form, and that when $\mathcal{X}$ is infinite one can use boosting-based techniques to efficiently approximate the corresponding projection.

Private sampling via mollification We call *mollification* the process of taking a distribution P and finding a distribution $\hat{P}$ inside a given mollifier $\mathcal{M}$ that minimizes the KL divergence:

$$\hat{P} \in \operatorname*{argmin}_{Q \in \mathcal{M}} \mathrm{KL}(P, Q) \ . \qquad (7)$$

We pick the KL divergence for its popularity and the fact that it is the canonical divergence for broad sets of distributions (Amari and Nagaoka, 2000). The appeal of this construction stems from the following result, which says that a mechanism that releases samples from some distribution in a mollifier provides privacy.

Lemma 2 *Let $A : \mathcal{D}(\mathcal{X}) \to \mathcal{X}$ by a randomized mechanism such that, for any P, $A(P)$ releases a sample from some $Q \in \mathcal{M}$. If $\mathcal{M}$ is an ε-mollifier, then A is an ε-private sampler.*

[3] Assumed finite for simplicity of exposition.

Thus, the *mollification mechanism* $A_{\mathcal{M}}$ that on input P releases a sample from the mollification $\hat{P}$ is a private sampler which tries to maximize utility by finding the closest distribution to P in a given mollifier. In order to implement the mechanism $A_{\mathcal{M}}$ it is necessary to solve the optimization problem (7). Furthermore, one also requires that the resulting distribution admits an efficient sampling procedure. With respect to the first requirement, we note that the problem in (7) is convex whenever the mollifier $\mathcal{M}$ is convex. Thus, the mollification problem could be solved efficiently using (stochastic[4]) convex optimization methods as long as $\mathcal{M}$ has a tractable representation. However, here we take a different approach.

For the case where the domain $\mathcal{X}$ is finite, the optimum of (7) admits a simple closed-form whenever $\mathcal{M}$ is a relative mollifier. In particular, for $\mathcal{M}_{\varepsilon,Q_0}$ it is easy to solve the Karush-Kuhn-Tucker (KKT) optimality conditions for (7) to show that the optimum is given by

$$\hat{P}(x) = \min\left\{\max\left\{\frac{Q_0(x)}{e^{\varepsilon/2}}, \frac{P(x)}{C}\right\}, e^{\varepsilon/2}Q_0(x)\right\}, \tag{8}$$

where C is a constant such that $\hat{P}$ sums to one. If P is only accessible through sampling, one can plug estimators for the probability of each element in $\mathcal{X}$ into the closed-form solution to obtain approximations to $\hat{P}$. An important observation is that no matter how bad this approximation is, the overall mechanism $A_{\mathcal{M}}$ remains private because the form of these closed-form solutions ensures the approximation is always inside the mollifier $\mathcal{M}_{\varepsilon,Q_0}$; this is a property that any private sampler using approximate mollification should satisfy.

When $\mathcal{X}$ is infinite this strategy is not immediately tractable, although one could try to obtain a non-parametric approximation to P and use it as a plug-in estimator in (8). Known properties of such estimators could be used to analyze the convergence of these non-parametric approximations, but the alternative approach we consider in this paper is more in line with modern methods in generative modelling. In particular, in Section 3 we provide a method for approximate mollification with relative mollifiers based on boosted density estimation. The boosting-based approach allows us to encode prior knowledge about the distributions P that we expect to encounter in practice in the choice of Q_0 and the architecture of the weak classifier trained at each iteration. This opens the door to using mollifiers learned from (non-private) data to improve the sample efficiency of private samplers; we leave this question for future research.

Algorithm 1 MBDE(WL, T, ε, Q_0)

```
input: Weak learner WL, # iterations $T$, privacy parameter $\varepsilon$, initial distribution $Q_0$, private target $P$;
for $t = 1, \ldots, T$ do
    $\theta_t(\varepsilon) \leftarrow \left(\frac{\varepsilon}{\varepsilon + 4\log(2)}\right)^t$
    $c_t \leftarrow \mathrm{WL}(P, Q_t)$
    $Q_t \propto Q_{t-1} \cdot \exp(\theta_t(\varepsilon) \cdot c_t)$
end for
return: $Q_T$
```

3 Mollification with approximation guarantees

The cornerstone of our approach to locally private sampling is an algorithm that (i) learns an explicit density in an ε-mollifier and (ii) with approximation guarantees with respect to the target P. We refer to the algorithm as MBDE, for Mollified Boosted Density Estimation; its pseudo-code is given in Algorithm 1.

To show convergence result on MBDE, we borrow the standard machinery from boosting, which includes classifiers $c : \mathcal{X} \to \mathbb{R}$ where $\mathrm{sign}(c(x)) \in \{-1, 1\}$ denotes classes. For technical convenience we assume $c(x) \in [-\log 2, \log 2]$ and so the output of c is bounded. This is a common assumption in the boosting literature (Schapire and Singer, 1999). We also require a pivotal condition from boosting: the *weak learning* assumption.

Definition 3 (WLA) *Fix $\gamma_P, \gamma_Q \in (0, 1]$ two constants. We say that WeakLearner$(.,.)$ satisfies the* ***weak learning assumption*** *(WLA) for γ_P, γ_Q iff for any P, Q, WeakLearner(P, Q) returns a classifier c satisfying $\mathbb{E}_P[c] > c^* \cdot \gamma_P$ and $\mathbb{E}_Q[-c] > c^* \cdot \gamma_Q$, where $c^* = \sup_{x \in \mathcal{X}} |c(x)|$.*

Briefly stated, a weak learner can be thought of as an oracle taking as inputs two distributions P and Q and is required to always return a classifier c that weakly guesses the sampling from P vs Q. Remark that as the two inputs P and Q become "closer" in some sense to one another, it is harder to satisfy the WLA. However, this is not a problem as whenever this happens, we shall have successfully learned P through Q. The classical theory of boosting would just assume one constraint over a distribution M whose marginals over classes would be P and Q (Kearns, 1988), but our definition can in fact easily be shown to coincide with that of boosting (Cranko and Nock, 2019).

MBDE is a private refinement of the DISCRIM algorithm of (Cranko and Nock, 2019, Section 3). It uses a weak learner whose objective is to distinguish between the target P and the current guessed density Q_t — the index indicates the iterative nature of the algorithm. Q_t is progressively

[4] Depending on whether we have access to P through a probability oracle for evaluating $P(x)$ or just through sampling.

refined using the weak learner's output classifier c_t, for a total number of user-fixed iterations T. We start boosting by setting Q_0 as the starting distribution, typically a simple non-informed (to be private) distribution such as a standard Gaussian (see also Figure 1, center). The classifier is then aggregated into Q_{t-1} as:

$$\begin{aligned} Q_t &= \frac{\exp(\theta_t(\varepsilon)c_t)Q_{t-1}}{\int \exp(\theta_t(\varepsilon)c_t)Q_{t-1}dx} \\ &= \exp\left(\langle\theta(\varepsilon), c\rangle - \varphi(\theta(\varepsilon))\right) Q_0, \end{aligned} \quad (9)$$

where $\theta(\varepsilon) = (\theta_1(\varepsilon), \ldots, \theta_t(\varepsilon))$, $c = (c_1, \ldots, c_t)$ (from now on, c denotes the vector of all classifiers) and $\varphi(\theta(\varepsilon))$ is the log-normalizer given by

$$\varphi(\theta(\varepsilon)) = \log \int_{\mathcal{X}} \exp\left(\langle\theta(\varepsilon), c\rangle\right) dQ_0. \quad (10)$$

This process repeats until $t = T$ and the proposed distribution is $Q_\varepsilon(x; P) \doteq Q_T$. We now show three formal results on MBDE.

MBDE is a private sampler Recall $\mathcal{M}_\varepsilon := \mathcal{M}_{\varepsilon,Q_0}$ is the set of densities whose range is in $\exp[-\varepsilon/2, \varepsilon/2]$ with respect to Q_0. Due to Lemma 2, it suffices to show that the output density Q_T of MBDE is in $\mathcal{M}_\varepsilon$.

Theorem 4 $Q_T \in \mathcal{M}_\varepsilon$.

We observe that privacy comes with a price, as for example $\lim_{\varepsilon \to 0} \theta_t(\varepsilon) = 0$, so as we become more private, the updates on $Q_\bullet$ become less and less significant and we somehow flatten the learned density — such a phenomenon is not a particularity of our method as it would also be observed for standard DP mechanisms (Dwork and Roth, 2014).

Convergence guarantees for MBDE As explained in Section 2, it is not hard to fit a density in $\mathcal{M}_\varepsilon$ to make its sampling private. An important question is however what guarantees of approximation can we still have with respect to P, given that P may not be in $\mathcal{M}_\varepsilon$. We now give such guarantees to MBDE in the boosting framework, and we also show that the approximation is within close order to the best possible given the constraint to fit $Q_\bullet$ in $\mathcal{M}_\varepsilon$. We start with the former result, and for this objective include the iteration index t in the notations from Definition 3 since the actual weak learning guarantees may differ across iterations, even when they are still within the prescribed bounds.

Theorem 5 *For any $t \geq 1$, suppose* WL *satisfies at iteration t the WLA for γ_P^t, γ_Q^t. Then we have:*

$$KL(P, Q_t) \leq KL(P, Q_{t-1}) - \theta_t(\varepsilon) \cdot \Lambda_t, \quad (11)$$

where (letting $\Gamma(z) \doteq \log(4/(5-3z))$):

$$\Lambda_t = \begin{cases} c_t^* \gamma_P^t + \Gamma(\gamma_Q^t) & \textit{if } \gamma_Q^t \in [1/3, 1] \textit{ ("HBS")} \\ \gamma_P^t + \gamma_Q^t - \frac{c_t^* \cdot \theta_t(\varepsilon)}{2} & \textit{if } \gamma_Q^t \in (0, 1/3) \textit{ ("LBS")} \end{cases}. \quad (12)$$

Here, HBS means high boosting regime and LBS means low boosting regime.

Remark that in the *high* boosting regime, we are guaranteed that $\Lambda_t \geq 0$ so the bound on the KL divergence is guaranteed to decrease. This is a regime we are more likely to encounter during the first boosting iterations since Q_{t-1} and P are then easier to tell apart — we can thus expect a larger γ_Q^t. In the low boosting regime, the picture can be different since we need $\gamma_P^t + \gamma_Q^t \geq c_t^* \cdot \theta_t(\varepsilon)/2$ to make the bound not vacuous. Since $\theta_t(\varepsilon) \to_t 0$ exponentially fast and $c_t^* \leq \log 2$, a constant, the constraint for (12) to be non-vacuous vanishes and we can also expect the bound on the KL divergence to also decrease in the *low* boosting regime. We now check that the guarantees we get are close to the best possible in an information-theoretic sense. Let us define $\Delta(Q) \doteq \mathrm{KL}(P, Q_0) - \mathrm{KL}(P, Q)$. Intuitively, the farther P is from Q_0, the farther we should be able to get from Q_0 to approximate P, and so the larger should be $\Delta(Q)$. Notice that this would typically imply to be in the high boosting regime for MBDE. For the sake of simplicity, we consider γ_P, γ_Q to be the same throughout all iterations.

Theorem 6 *We have $\Delta(Q) \leq \varepsilon/2$, $\forall Q \in \mathcal{M}_\varepsilon$, and if* MBDE *is in the high boosting regime, then*

$$\Delta(Q_T) \geq \frac{\varepsilon}{2} \cdot \left\{ \frac{\gamma_P + \gamma_Q}{2} \cdot (1 - \theta_T(\varepsilon)) \right\}. \quad (13)$$

Hence, as $\gamma_P \to 1$ and $\gamma_Q \to 1$, we have $\Delta(Q_T) \geq (\varepsilon/2) \cdot (1 - \theta_T(\varepsilon))$ and since $\theta_T(\varepsilon) \to 0$ as $T \to \infty$, MBDE indeed reaches (in the high boosting regime) the information-theoretic limit, which is the mollification of P. As ε increases (the privacy constraint is reduced), Theorem 6 shows that we are guaranteed to progressively come closer to P, and if we make the additional assumption that there exists $\varepsilon_P \ll \infty$ such that $P \in \mathcal{M}_{\varepsilon_P}$ – which appears to be quite reasonable given the definition in (5) –, then Theorem 6 delivers a direct approximability result for MBDE with respect to P for all privacy levels $\varepsilon \geq \varepsilon_P$. This is a new result compared to the privacy-free approximation bounds of P in (Cranko and Nock, 2019), but it requires to be in the high boosting regime.

MBDE captures the modes of P Mode capture is a prominent problem in the area of generative models (Tolstikhin et al., 2017). We have already seen that enforcing mollification can be done while keeping modes, but we would like to show that MBDE is indeed efficient at building some Q_T with guarantees on mode capture. For this objective, we define for any set $B \subseteq \mathcal{X}$ and distribution Q,

$$\mathrm{M}_B(Q) \doteq \int_B dQ\ ,\ \mathrm{KL}_B(P, Q) \doteq \int_B \log\left(\frac{P}{Q}\right) dP,$$

respectively the total mass of B on Q and the KL divergence between P and Q restricted to B.

Theorem 7 *Suppose* MBDE *stays in the high boosting. Then* $\forall \alpha \in [0, 1]$, $\forall B \subseteq \mathcal{X}$, *if*

$$\mathrm{M}_B(P) \quad \geq \quad \varepsilon \cdot \frac{h((2 - \gamma_P - \gamma_Q) \cdot T)}{h(\alpha) \cdot h(T)}, \qquad (14)$$

then $\mathrm{M}_B(Q_T) \geq (1 - \alpha)\mathrm{M}_B(P) - \mathrm{KL}_B(P, Q_0)$, *where* $h(x) \doteq \varepsilon + 2x$.

There is not much we can do to control $\mathrm{KL}_B(P, Q_0)$ as this term quantifies our luck in picking Q_0 to approximate P in B but if this restricted KL divergence is small compared to the mass of B, then we are guaranteed to capture a substantial part of it through Q_T. As a mode, in particular “fat”, would tend to have large mass over its region B, Theorem 7 says that we can indeed hope to capture a significant part of it as long as we stay in the high boosting regime. As $\gamma_P \to 1$ and $\gamma_Q \to 1$, the condition on $\mathrm{M}_B(P)$ in (14) vanishes with T and we end up capturing any fat region B (and therefore, modes, assuming they represent ”fatter” regions) whose mass is sufficiently large with respect to $\mathrm{KL}_B(P, Q_0)$.

With regards to a practical application, consider the problem of generating synthetic text data from conversational English. Each individual user holds their own distribution (own speech patterns and vocabulary) and the goal is to be able to model these distributions with privacy and approximation guarantees. We point out two implicit advantages of our method over standard local DP and federated learning methods: (i) Our method relies on the reference distribution Q_0, which in this application, one may use public conversational data to learn Q_0 using a strong non-private algorithm. In this case, the ε-mollifier centered at Q_0 will contain admissible conversations with relatively high utility, meaning that mollifications will still be reasonable. (ii) Our method is non-interactive: each user generates a privatized sample which is submitted to the server for post-processing.

To finish up this Section, recall that $\mathcal{M}_\varepsilon$ is also defined (in disguise) and analyzed in (Wang et al., 2015, Theorem 1) for posterior sampling. However, the convergence in (Wang et al., 2015, Section 3) does not dig into specific forms for the likelihood of densities chosen — as a result, it remains essentially in weak asymptotic form, and furthermore it is only applied to DP in the curator model. We exhibit particular choices for these mollifier densities, along with a specific training algorithm to learn them, that allow for significantly better approximation, quantitatively and qualitatively (mode capture) in the local DP setting.

4 Related work

A broad literature has been developed early for discrete distributions (Machanavajjhala et al., 2008) (and references therein). For a general Q not necessarily discrete, more sophisticated approaches have been tried, most of which exploit randomisation and the basic toolbox of differential privacy (Dwork and Roth, 2014, Section 3): given non-private $\tilde{Q}$, one compute the *sensitivity* s of the approach, then use a standard mechanism $M(\tilde{Q}, s)$ to compute a private Q. Such general approaches have been used for Q being the popular kernel density estimation (KDE, (Givens and Hoeting, 2013)) with variants (Aldà and Rubinstein, 2017; Hall et al., 2013; Rubinstein and Aldà, 2017).

On the algorithmic side, our work shares some ideas with DP methods based on the *multiplicate weights* technique (Hardt and Rothblum, 2010; Hardt et al., 2012; Ullman, 2015). These papers leverage ideas similar to boosting to solve problems like answering linear queries, solving convex minimization problems, or releasing synthetic data to accurately answer a pre-determined set of queries. None of these works, however, apply directly to the local DP model.

5 Experiments

Architectures We carried out experiments on a simulated setting inspired by (Aldà and Rubinstein, 2017), to compare MBDE (implemented following its description in Section 3) against differentially private KDE (Aldà and Rubinstein, 2017). As a weak learner for MBDE, we fit for each c_t a neural network (NN) classifier:

$$\mathcal{X} \xrightarrow[\text{dense}]{\tanh} \mathbb{R}^{25} \xrightarrow[\text{dense}]{\tanh} \mathbb{R}^{25} \xrightarrow[\text{dense}]{\tanh} \mathbb{R}^{25} \xrightarrow[\text{dense}]{\text{sigmoid}} (0, 1), \quad (15)$$

where $\mathcal{X} \in \{\mathbb{R}, \mathbb{R}^2\}$ depending on the experiment. At each iteration t of boosting, c_t is trained using 10000 samples from P and Q_{t-1} using Nesterov’s accelerated gradient descent with $\eta = 0.01$ based on cross-entropy loss with 750 epochs. Random walk Metropolis-Hastings is used to sample from Q_{t-1} at each iteration. For the number of boosting iterations in MBDE, we pick $T = 3$. This is quite a small value but given the rate of decay of $\theta_t(\varepsilon)$ and the small dimensionality of the domain, we found it a good compromise for complexity vs accuracy. Finally, Q_0 is a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathrm{I}_d)$.

Contenders We know of no local differentially private sampling approach operating under conditions equivalent to ours, so our main contender is going to be a particular

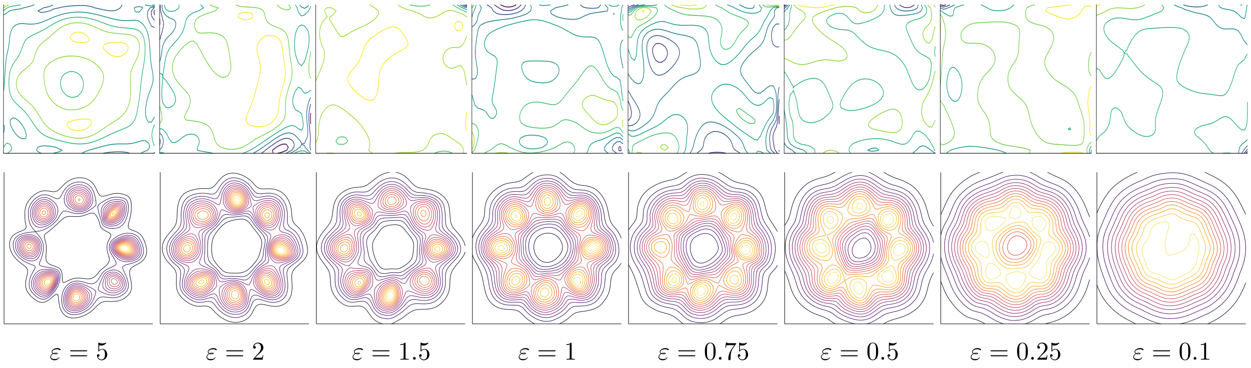


Figure 3: Gaussian ring: densities obtained for DPB (upper row) against MBDE (lower row)

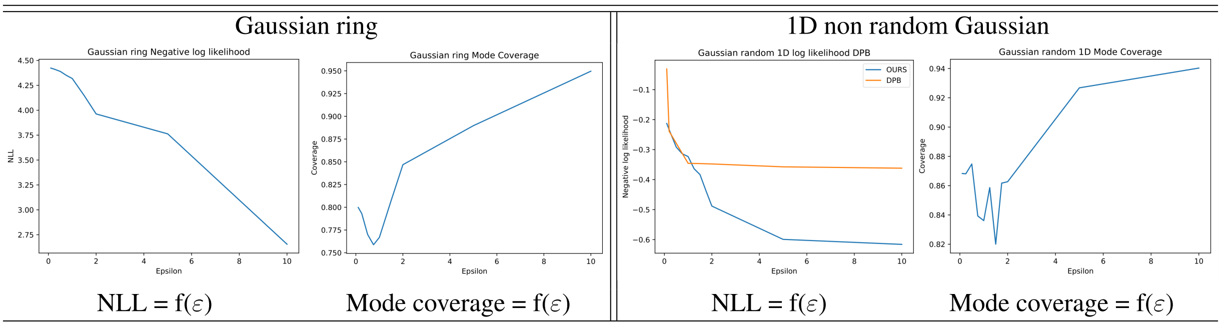


Figure 4: Metrics for MBDE (blue): NLL (lower is better) and mode coverage (higher is better). Orange: DPB (see text).

state of the art ε-differentially private approach which provides a private *density*, DPB (Aldà and Rubinstein, 2017). We choose this approach because digging in its technicalities reveal that *its **local differential privacy budget** would be roughly equivalent to ours, mutatis mutandis*. Here is why: this approach allows to sample a dataset of arbitrary size (say, k) while keeping the same privacy budget, *but* needs to be scaled to accomodate local differential privacy, while in our case, MBDE allows to obtain local differential privacy for one observation ($k = 1$), *but* its privacy budget needs to be scaled to accomodate for larger k. It turns out that in both approaches, the scaling of the privacy parameter to accomodate for arbitrary k and local differential privacy is roughly the same. In our case, the change is obvious: the privacy parameter ε is naturally scaled by k by the composition property of ε-LDP. In the case of (Aldà and Rubinstein, 2017), the requirement of local differential privacy multiplies the sensitivity[5] by k by the group privacy property.

We have also compared with a private GAN approach, which has the benefit to yield a simple sampler but involves a weaker privacy model (Xie et al., 2018) (DPGAN). For DPB, we use a bandwidth kernel and learn the bandwidth parameter via 10-fold cross-validation. For DPGAN, we train the WGAN base model using batch sizes of 128 and 10000 epochs, with $\delta = 10^{-1}$. We found that DPGAN is significantly outperformed by both DPB and MBDE, so to save space we have only included the experiment in Figure 1 (right). We observed that DPB does not always yield a positive measure. To ensure positivity, we shift and scale the output.

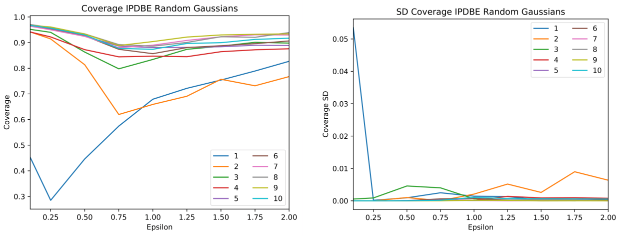
Metrics We consider two metrics, inspired by those we consider for our theoretical analysis and one investigated in (Tolstikhin et al., 2017) for mode capture. We first investigate the ability of our method to learn highly dense regions by computing *mode coverage*, which is defined to be $P(dQ < t)$ for t such that $Q(dQ < t) = 0.95$. Mode coverage essentially attempts to find high density regions of the model Q (based on t) and computes the mass of the target P under this region. Second, we compare the negative log likelihood, $-E_P[\log Q]$ as a general loss measure.

Domains We essentially consider three different problems. The first is the ring Gaussians problem now common to generative approaches (Goodfellow, 2016), in which 8 Gaussians have their modes regularly spaced on a

[5] Cf (Aldà and Rubinstein, 2017, Definition 4) for the sensitivity, (Aldà and Rubinstein, 2017, Section 6) for the key function $F_H(.,.)$ involved.

circle. The target P is shown in Figure 1. Second, we consider a mixture of three 1D gaussians with pdf $P(x) = \frac{1}{3}\left(\mathcal{N}(0.3, 0.01) + \mathcal{N}(0.5, 0.1) + \mathcal{N}(0.7, 0.1)\right)$. For the final experiment, we consider a 1D domain and randomly place m gaussians with means centered in the interval $[0, 1]$ and variances 0.01. We vary $m = 1, \ldots, 10$, $\varepsilon \in (0, 2]$ and repeat the experiment four times to get means and standard deviations. More experiments can be found in the Appendix.

Results Figure 3 displays contour plots of the learned Q against DPB (Aldà and Rubinstein, 2017). Figure 4 provides metrics. We indicate the metric performance for DPB on one plot only since density estimates obtained for some of the other metrics could not allow for an accurate computation of metrics. The experiments bring the following observations: MBDE is significantly better at local differentially private density estimation than DPB if we look at the ring Gaussian problem. MBDE essentially obtains the same results as DPB for values of ε that are 400 times smaller as seen from Figure 1. We also remark that the density modelled are more smooth and regular for MBDE in this case. One might attribute the fact that our performance is much better on the ring Gaussians to the fact that our Q_0 is a standard Gaussian, located at the middle of the ring in this case, but experiments on random 2D Gaussians (see Appendix) display that our performances also remain better in other settings where Q_0 should represent a handicap. All domains, including the 1D random Gaussians experiments in Figure 1 (Appendix), display a consistent decreasing NLL for MBDE as ε increases, with sometimes very sharp decreases for $\varepsilon < 2$ (See also Appendix, Section 2). We attribute it to the fact that it is in this regime of the privacy parameter that MBDE captures all modes of the mixture. For larger values of ε, it justs fits better the modes already discovered. We also remark on the 1D Gaussians that DPB rapidly reaches a plateau of NLL which somehow show that there is little improvement as ε increases, for $\varepsilon \geq 1$. This is not the case for MBDE, which still manages some additional improvements for $\varepsilon > 5$ and significantly beats DPB. We attribute it to the flexibility of the sufficient statistics as (deep) classifiers in MBDE. The 1D random Gaussian problem (Figure 1 in Appendix) displays the same pattern for MBDE. We also observe that the standard deviation of MBDE is often 100 times *smaller* than for DPB, indicating not just better but also much more stable results. In the case of mode coverage, we observe for several experiments (*e.g.* ring Gaussians) that the mode coverage *decreases* until $\varepsilon \approx 1$, and then increases, on all domains, for MBDE. This, we believe is due to our choice of Q_0, which as a Gaussian, already captures with its mode a part of the existing modes. As ε increases however, MBDE performs better and obtains in general a significant improvement over Q_0. We also observe this phenomenon for the random 1D Gaussians (Figure 5) where the very small standard deviations (at least for $\varepsilon > .25$ or $m > 1$) display a significant stability for the solutions of MBDE.



Mean = f(ε) StDev = f(ε)

Figure 5: Mode coverage for MBDE on 1D random Gaussian.

6 Discussion and Conclusion

In this paper, we proposed a new method to learn densities that can be sampled from privately at the local level, paving the way for synthetic data generation. In order to prove privacy guarantees, we introduced the notion of mollifiers, which are of independent interest. Furthermore, we proved convergence guarantees of our method in the context of boosting along with additional formal results regarding capturing of modes and approximation of the target density. The use of the boosting framework allows to dampen the effects of a "curse of complexity" – *e.g.* when the dimension of the support of P increases –, as convergence primarily relies on weak guessing in sampling P vs sampling vs $Q_.$. Additional assumptions, like sparsity in the expected parameters of the target or publicly available information allowing to tune Q_0, could boost further convergence. Finally, we conducted experiments, which advocate for our method, especially on the utility side of things when it comes to capturing statistical features of the true distribution.

Acknowledgments

We are indebted to Benjamin Rubinstein for providing us with the Private KDE code. We would also like to thank Arthur Street, Gianpaolo Gioiosa and anonymous reviewers for significant help in correcting and improving focus, clarity and presentation.

References

Aldà, F. and Rubinstein, B. (2017). The Bernstein mechanism: Function release under differential privacy. In *AAAI'17*.

Amari, S.-I. and Nagaoka, H. (2000). *Methods of Information Geometry*. Oxford University Press.

Bittau, A., Erlingsson, Ú., Maniatis, P., Mironov, I., Raghunathan, A., Lie, D., Rudominer, M., Kode, U., Tinnes, J., and Seefeld, B. (2017). Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 441–459. ACM.

Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., and Wang, T. (2018). Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 1655–1658, New York, NY, USA. ACM.

Cranko, Z. and Nock, R. (2019). Boosted density estimation remastered. In *ICML'19*.

Differential privacy team, Apple (2017). Learning with differential privacy at scale.

Dwork, C. and Roth, A. (2014). The algorithmic foudations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9:211–407.

Erlingsson, U., Pihur, V., and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA. ACM.

Givens, G.-F. and Hoeting, J.-A. (2013). *Computational Statistics*. Wiley.

Goodfellow, I. (2016). Generative adversarial networks. NIPS'16 tutorials.

Hall, R., Rinaldo, A., and Wasserman, L.-A. (2013). Differential privacy for functions and functional data. *JMLR*, 14(1):703–727.

Hardt, M., Ligett, K., and McSherry, F. (2012). A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems*, pages 2339–2347.

Hardt, M. and Rothblum, G. N. (2010). A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 61–70. IEEE.

Kearns, M. (1988). Thoughts on hypothesis boosting. ML class project.

Machanavajjhala, A., Kifer, D., Abowd, J.-M., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *ICDE'08*, pages 277–286.

Raskhodnikova, S., Smith, A., Lee, H. K., Nissim, K., and Kasiviswanathan, S. P. (2008). What can we learn privately. In *Proceedings of the 54th Annual Symposium on Foundations of Computer Science*, pages 531–540.

Rubinstein, B. and Aldà, F. (2017). Pain-free random differential privacy with sensivity sampling. In *34^{th} ICML*.

Schapire, R. E. and Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *MLJ*, 37:297–336.

Tolstikhin, I.-O., Gelly, S., Bousquet, O., Simon-Gabriel, C., and Schölkopf, B. (2017). Adagan: Boosting generative models. In *NIPS*30*, pages 5430–5439.

Ullman, J. (2015). Private multiplicative weights beyond linear queries. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 303–312. ACM.

Wang, Y.-X., Fienberg, S., and Smola, A.-J. (2015). Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *32^{nd} ICML*, pages 2493–2502.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. (2018). Differentially private generative adversarial network. *CoRR*, abs/1802.06739.

Supplementary Material to Paper ”Local Differential Privacy for Sampling”

Abstract

This is the Supplementary Material to Paper ”Local Differential Privacy for Sampling”. Notation “main file” indicates reference to the submitted draft.

Appendix: table of contents

1 Proofs and formal results

1.1 Proof of Lemma 2

For any $x \in \mathcal{X}$ and $P, P' \in \mathcal{D}(\mathcal{X})$, we have Pr$[A(P) = x] \in \mathcal{M}$ and Pr$[A(P') = x] \in \mathcal{M}$ by the fact that $A(P)$ samples from densities that lie in the mollifier $\mathcal{M}$. By definition of ε-mollifiers, the density ratio between any two densities in the ε-mollifiers is bounded by $\exp(\varepsilon)$, meaning we have

$$\frac{\text{Pr}[A(P) = x]}{\text{Pr}[A(P') = x]} \leq \exp(\varepsilon), \tag{1}$$

and thus A is an ε-private sampler.

1.2 Proof of Theorem 4

The proof follows from two Lemma which we state and prove.

Lemma 1 *For any $T \in \mathbb{N}_*$, we have that*

$$\sum_{t=1}^{T} \theta_t(\varepsilon) = \sum_{t=1}^{T} \left(\frac{\varepsilon}{\varepsilon + 4\log(2)}\right)^t < \frac{\varepsilon}{4\log(2)}. \tag{2}$$

Proof Since $(\varepsilon/(\varepsilon + 4\log(2)) < 1$ for any ε and noting that $\theta_t(\varepsilon) = (\varepsilon/(\varepsilon + 4\log(2))\theta_{t-1}(\varepsilon)$, we can conclude that $\theta_t(\varepsilon)$ is a geometric sequence. For any geometric series with ratio r, we have that

$$\sum_{t=1}^{T} r^t = r\left(\frac{1 - r^T}{1 - r}\right) \tag{3}$$

$$= \frac{r}{1-r} - \frac{r^{T+1}}{1-r} \tag{4}$$

$$< \frac{r}{1-r} \tag{5}$$

Indeed, $\frac{r}{1-r}$ is the limit of the geometric series above when $T \to \infty$. In our case, we let $r = (\varepsilon/(\varepsilon + 4\log(2)))$ to show that

$$\frac{r}{1-r} = \frac{\frac{\varepsilon}{\varepsilon+4\log(2)}}{1 - \frac{\varepsilon}{\varepsilon+4\log(2)}} = \frac{\frac{\varepsilon}{\varepsilon+4\log(2)}}{\frac{4\log(2)}{\varepsilon+4\log(2)}} = \frac{\varepsilon}{4\log(2)}, \tag{6}$$

which concludes the proof. ■

Lemma 2 *For any $\varepsilon > 0$ and $T \in \mathbb{N}_*$, let $\theta(\varepsilon) = (\theta_1(\varepsilon), \ldots, \theta_T(\varepsilon))$ denote the parameters and $c = (c_1, \ldots, c_t)$ denote the sufficient statistics returned by Algorithm 1, then we have*

$$-\frac{\varepsilon}{2} \leq \langle \theta(\varepsilon), c \rangle - \varphi(\theta(\varepsilon)) \leq \frac{\varepsilon}{2}. \tag{7}$$

Proof Since the algorithm returns classifiers such that $c_t(x) \in [-\log 2, \log 2]$ for all $1 \leq t \leq T$, we have from Lemma 1,

$$\sum_{t=1}^{T} \theta_t(\varepsilon) c_t \leq \log(2) \sum_{t=1}^{T} \theta_t(\varepsilon) < \log(2) \frac{\varepsilon}{4 \log(2)} = \frac{\varepsilon}{4}, \tag{8}$$

and similarly,

$$\sum_{t=1}^{T} \theta_t(\varepsilon) c_t \geq -\log(2) \sum_{t=1}^{T} \theta_t(\varepsilon) > -\log(2) \frac{\varepsilon}{4 \log(2)} = -\frac{\varepsilon}{4}. \tag{9}$$

Thus we have

$$-\frac{\varepsilon}{4} \leq \langle \theta(\varepsilon), c \rangle \leq \frac{\varepsilon}{4}. \tag{10}$$

By taking exponential, integrand (w.r.t Q_0) and logarithm of 10, we get

$$\log \int_{\mathcal{X}} \exp\left(-\frac{\varepsilon}{4}\right) dQ_0 \leq \log \int_{\mathcal{X}} \exp\left(\langle \theta(\varepsilon), c \rangle\right) dQ_0 \leq \log \int_{\mathcal{X}} \exp\left(\frac{\varepsilon}{4}\right) dQ_0 \tag{11}$$

$$-\frac{\varepsilon}{4} \leq \varphi(\theta(\varepsilon)) \leq \frac{\varepsilon}{4} \tag{12}$$

Since $\langle \theta(\varepsilon), c \rangle \in [-\varepsilon/4, \varepsilon/4]$ and $\varphi(\theta(\varepsilon)) \in [-\varepsilon/4, \varepsilon/4]$, the proof concludes by considering highest and lowest values. ■

The proof of Theorem 4 now follows from taking the $\exp$ of all quantities in (7), which makes appear Q_T in the middle and conditions for membership to $\mathcal{M}_\varepsilon$ in the bounds.

1.3 Proof of Theorem 5

We begin by first deriving the KL drop expression. At each iteration, we learn a classifier c_t, fix some step size $\theta > 0$ and multiply Q_{t-1} by $\exp(\theta \cdot c_t)$ and renormalize to get a new distribution which we will denote by $Q_t(\theta)$ to make the dependence of θ explicit.

Lemma 3 *For any $\theta > 0$, let $\varphi(\theta) = \log \int_{\mathcal{X}} \exp(\theta \cdot c_t) dQ_{t-1}$. The drop in KL is*

$$\mathit{DROP}(\theta) := \mathit{KL}(P, Q_{t-1}) - \mathit{KL}(P, Q_t(\theta)) = \theta \cdot \int_{\mathcal{X}} c_t dP - \varphi(\theta) \tag{13}$$

Proof Note that $Q_t(\theta)$ is indeed a one dimensional exponential family with natural parameter θ, sufficient statistic c_t, log-partition function $\varphi(\theta)$ and base measure Q_{t-1}. We can write out the KL divergence as

$$\mathrm{KL}(P, Q_{t-1}) - \mathrm{KL}(P, Q_t(\theta)) = \int_{\mathcal{X}} \log\left(\frac{P}{Q_{t-1}}\right) dP - \int_{\mathcal{X}} \log\left(\frac{P}{\exp(\theta \cdot c_t - \varphi(\theta))Q_{t-1}}\right) dP \tag{14}$$

$$= \int_{\mathcal{X}} \log\left(\frac{\exp(\theta \cdot c_t - \varphi(\theta))Q_{t-1}}{Q_{t-1}}\right) dP \tag{15}$$

$$= \int_{\mathcal{X}} \theta \cdot c_t - \varphi(\theta) dP \tag{16}$$

$$= \theta \cdot \int_{\mathcal{X}} c_t dP - \varphi(\theta) \tag{17}$$

■

It is not hard to see that the drop is indeed a concave function of θ, suggesting that there exists an optimal step size at each iteration. We split our analysis by considering two cases and begin when $\gamma_Q^t < 1/3$. Since $\theta > 0$, we can lowerbound the first term of the KL drop using WLA. The trickier part however, is bounding $\varphi(\theta)$ which we make use of Hoeffding's lemma.

Lemma 4 (Hoeffding's Lemma) *Let X be a random variable with distribution Q, with $a \leq X \leq b$ such that $\mathbb{E}_Q[X] = 0$, then for all $\lambda > 0$, we have*

$$\mathbb{E}_Q[\exp(\lambda \cdot X)] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \tag{18}$$

Lemma 5 *For any classifier c_t satisfying Assumption 3 (WLA), we have*

$$\mathbb{E}_{Q_{t-1}}[\exp(\theta_t(\varepsilon) \cdot c_t)] \leq \exp\left(\theta_t^2(\varepsilon) \cdot \frac{(c_t^*)^2}{2} - \theta_t(\varepsilon) \cdot \gamma_Q^t \cdot c_t^*\right) \tag{19}$$

Proof Let $X = c_t - \cdot\mathbb{E}_{Q_{t-1}}[c_t]$, $b = c_t^*$, $a = -c_t^*$ and $\lambda = \theta_t(\varepsilon)$ and noticing that

$$\mathbb{E}_{Q_{t-1}}[\lambda \cdot X] = \mathbb{E}_{Q_{t-1}}[c_t - \mathbb{E}_{Q_{t-1}}[c_t]] = \mathbb{E}_{Q_{t-1}}[c_t] - \mathbb{E}_{Q_{t-1}}[c_t] = 0, \tag{20}$$

allows us to apply Lemma 4. By first realizing that

$$\exp(\lambda \cdot X) = \exp(\theta_t(\varepsilon) \cdot c_t) \cdot \exp(\theta_t(\varepsilon) \cdot \mathbb{E}_{Q_{t-1}}[-c_t]), \tag{21}$$

We get that

$$\mathbb{E}_{Q_{t-1}}[\exp(\theta_t(\varepsilon) \cdot c_t)] \cdot \exp\left(\theta_t(\varepsilon) \cdot \mathbb{E}_{Q_{t-1}}[-c_t]\right) \leq \exp\left(\theta_t^2(\varepsilon) \cdot \frac{(c_t^*)^2}{2}\right). \tag{22}$$

Re-arranging and using the WLA inequality yields

$$\mathbb{E}_{Q_{t-1}}[\exp(\theta_t(\varepsilon) \cdot c_t)] \leq \exp\left(\theta_t^2(\varepsilon) \cdot \frac{(c_t^*)^2}{2} - \theta_t(\varepsilon) \cdot \mathbb{E}_{Q_{t-1}}[-c_t]\right) \tag{23}$$

$$\leq \exp\left(\theta_t^2(\varepsilon) \cdot \frac{(c_t^*)^2}{2} - \theta_t(\varepsilon) \cdot \gamma_Q^t \cdot c_t^*\right) \tag{24}$$

■

Applying Lemma 5 and Lemma 3 (writing $Q_t = Q_t(\varepsilon)$) together gives us

$$\begin{aligned}
\mathrm{KL}(P, Q_t) &= \mathrm{KL}(P, Q_{t-1}) - \mathrm{DROP}(\theta_t(\varepsilon)) && (25)\\
&= \mathrm{KL}(P, Q_{t-1}) - \theta_t(\varepsilon) \cdot \int_{\mathcal{X}} c_t dP + \log \mathbb{E}_{Q_{t-1}}[\exp(\theta_t(\varepsilon) \cdot c_t)] && (26)\\
&\leq \mathrm{KL}(P, Q_{t-1}) - c_t^* \cdot \theta_t(\varepsilon) \cdot \left(\frac{1}{c_t^*} \int_{\mathcal{X}} c_t dP \right) + \left(\theta_t^2(\varepsilon) \cdot \frac{(c_t^*)^2}{2} - \theta_t(\varepsilon) \cdot \gamma_Q^t \cdot c_t^* \right) && (27)\\
&\leq \mathrm{KL}(P, Q_{t-1}) - c_t^* \theta_t(\varepsilon) \left(\gamma_P^t + \gamma_Q^t - \frac{c_t^* \cdot \theta_t(\varepsilon)}{2} \right) && (28)
\end{aligned}$$

Now we move to the case of $\gamma_Q^t \geq 1/3$.

Lemma 6 *For any classifier c_t returned by Algorithm 1, we have that*

$$\mathbb{E}_{Q_{t-1}}[\exp(c_t)] \leq \exp\left(-\Gamma(\gamma_Q^t)\right) \qquad (29)$$

where $\Gamma(z) = \log(4/(5-3z))$.

Proof Consider the straight line between $(-\log 2, 1/2)$ and $(\log 2, 2)$ given by $y = 5/4 + (3/(4 \cdot \log 2))x$, which by convexity is greater then $y = \exp(x)$ on the interval $[-\log 2, \log 2]$. To this end, we define the function

$$f(x) = \begin{cases} \frac{5}{4} + \frac{3}{4 \cdot \log 2} \cdot x, & \text{if } x \in [-\log 2, \log 2] \\ 0, & \text{otherwise} \end{cases} \qquad (30)$$

Since $c_t(x) \in [-\log 2, \log 2]$ for all $x \in \mathcal{X}$, we have that $f(c_t(x)) \geq \exp(c_t(x))$ for all $x \in \mathcal{X}$. Taking $\mathbb{E}_{Q_{t-1}}[\cdot]$ over both sides and using linearity of expectation gives

$$\begin{aligned}
\mathbb{E}_{Q_{t-1}}[\exp(c_t(x))] &\leq \mathbb{E}_{Q_{t-1}}[f(c_t(x))] && (31)\\
&= \frac{5}{4} + \frac{3}{4 \log 2} \left(\mathbb{E}_{Q_{t-1}}[c_t(x)] \right) && (32)\\
&= \frac{5}{4} - \frac{3}{4} \left(\frac{1}{\log 2} \mathbb{E}_{Q_{t-1}}[-c_t(x)] \right) && (33)\\
&< \frac{5}{4} - \frac{3}{4} \gamma_Q^t && (34)\\
&= \exp\left(-\log \left(\frac{5 - 3\gamma_Q^t}{4} \right)^{-1} \right) && (35)\\
&= \exp\left(-\log \left(\frac{4}{5 - 3\gamma_Q^t} \right) \right) && (36)\\
&= \exp\left(-\Gamma(\gamma_Q^t) \right), && (37)
\end{aligned}$$

as claimed. ■

Now we use Lemma 3 and Jensen's inequality since $\theta_t(\varepsilon) < 1$ so that

$$\begin{aligned}
\mathrm{KL}(P, Q_t) &= \mathrm{KL}(P, Q_{t-1}) - \mathrm{DROP}(\theta) && (38)\\
&= \mathrm{KL}(P, Q_{t-1}) - \theta_t(\varepsilon) \cdot \int_{\mathcal{X}} c_t dP + \log \mathbb{E}_{Q_{t-1}}[\exp(\theta_t \cdot c_t)] && (39)\\
&\leq \mathrm{KL}(P, Q_{t-1}) - \theta_t(\varepsilon) \cdot \mathbb{E}_P[c_t] + \theta_t \cdot \log \mathbb{E}_{Q_{t-1}}[\exp(c_t)] && (40)\\
&\leq \mathrm{KL}(P, Q_{t-1}) - \theta_t(\varepsilon) \left(\mathbb{E}_P[c_t] - \log \mathbb{E}_{Q_{t-1}}[\exp(c_t)]\right) && (41)\\
&= \mathrm{KL}(P, Q_{t-1}) - \theta_t(\varepsilon) \left(c_t^* \left(\frac{1}{c_t^*}\mathbb{E}_P[c_t]\right) - \log \mathbb{E}_{Q_{t-1}}[\exp(c_t)]\right) && (42)\\
&< \mathrm{KL}(P, Q_{t-1}) - \theta_t(\varepsilon) \left(c_t^* \gamma_P^t - \log\left(\exp\left(-\Gamma(\gamma_Q^t)\right)\right)\right) && (43)\\
&= \mathrm{KL}(P, Q_{t-1}) - \theta_t(\varepsilon) \left(c_t^* \gamma_P^t + \Gamma(\gamma_Q^t)\right). && (44)
\end{aligned}$$

1.4 Proof of Theorem 6

We first note that for any $Q \in \mathcal{M}_\varepsilon$,

$$\begin{aligned}
\mathrm{KL}(P, Q) &= \int_{\mathcal{X}} \log\left(\frac{P}{Q}\right) dP && (45)\\
&= \int_{\mathcal{X}} \log\left(\frac{P}{Q_0}\frac{Q_0}{Q}\right) dP && (46)\\
&= \int_{\mathcal{X}} \log\left(\frac{P}{Q_0}\right) dP - \int_{\mathcal{X}} \log\left(\frac{Q_0}{Q}\right) dP && (47)\\
&\geq \mathrm{KL}(P, Q_0) - \int_{\mathcal{X}} \frac{\varepsilon}{2} dP && (48)\\
&\geq \mathrm{KL}(P, Q_0) - \frac{\varepsilon}{2}, && (49)
\end{aligned}$$

which completes the proof of the upperbound To show (13), we have that

$$\begin{aligned}
\mathrm{KL}(P, Q_t) &\leq \mathrm{KL}(P, Q_{T-1}) - \theta_t(\varepsilon) \cdot \Lambda_t && (50)\\
&\leq \mathrm{KL}(P, Q_0) - \sum_{t=1}^{T-1} \theta_t(\varepsilon) \cdot \Lambda_t && (51)\\
&= \mathrm{KL}(P, Q_0) - \sum_{t=1}^{T-1} \theta_t(\varepsilon) \cdot \left(c_t^* \gamma_P^t + \Gamma(\gamma_Q^t)\right) && (52)\\
&\leq \mathrm{KL}(P, Q_0) - \sum_{t=1}^{T-1} \theta_t(\varepsilon) \cdot (\log 2 \cdot \gamma_P + \Gamma(\gamma_Q)) && (53)\\
&\leq \mathrm{KL}(P, Q_0) - (\log 2 \cdot \gamma_P + \log 2 \cdot \gamma_Q) \cdot \sum_{t=1}^{T-1} \theta_t(\varepsilon) && (54)\\
&\leq \mathrm{KL}(P, Q_0) - (\log 2 \cdot \gamma_P + \log 2 \cdot \gamma_Q) \cdot \sum_{t=1}^{T-1} \theta_t(\varepsilon) && (55)\\
&= \mathrm{KL}(P, Q_0) - \log 2 \cdot (\gamma_P + \gamma_Q) \cdot \theta_1(\varepsilon) \cdot \left(\frac{1 - \theta_t(\varepsilon)}{1 - \theta_1(\varepsilon)}\right) && (56)\\
&= \mathrm{KL}(P, Q_0) - \varepsilon \cdot \left(\frac{\gamma_P + \gamma_Q}{4}\right) \cdot (1 - \theta_t(\varepsilon)), && (57)
\end{aligned}$$

where we used the fact that $\Gamma(x) \geq \log 2 \cdot x$ and explicit geometric summation expression.

1.5 Proof of Theorems 7

We start by a general Lemma.

Lemma 7 *For any region of the support B, we have that*

$$\int_B dQ_t \geq \int_B dP - \int_B \log\left(\frac{P}{Q_t}\right) dP \tag{58}$$

Proof By first noting that for any region B,

$$\int_B (dP - dQ_t) = \int_B \left(1 - \frac{dQ_t}{dP}\right) dP \tag{59}$$

we then use the inequality $1 - x \leq \log(1/x)$ to get

$$\int_B (dP - dQ_t) = \int_B \left(1 - \frac{dP}{dQ_t}\right) dP \leq \int_B \log\left(\frac{dP}{dQ_t}\right) dP = \int_B \log\left(\frac{P}{Q_t}\right) dP \tag{60}$$

Re-arranging the above inequality gives us the bound. ■

Lemma 7 allows us to understand the relationship between two distributions P and Q_t in terms regions they capture. The general goal is to show that for a given region B (which includes the highly dense mode regions), the amount of mass captured by the model $\int_B dQ_t$, is lower bounded by the target mass $\int_B dP$, and some small quantity. The inequality in Lemma 7 comments on this precisely with the small difference being a term that looks familiar to the KL-divergence - rather one that is bound to the specific region B. Though, this term can be understood to be small since by Theorem 5, we know that the global KL decreases, we give further refinements to show the importance of privacy parameters ε. We show that the term $\int_B \log(P/Q_t)dP$ can be decomposed in different ways, leading to our two Theorems to prove.

Lemma 8

$$\int_B \log\left(\frac{P}{Q_t}\right) dP \leq \int_B \log\left(\frac{P}{Q_0}\right) dP - \Delta + \frac{\varepsilon}{2}\left(1 - \int_B dP\right). \tag{61}$$

where $\Delta = KL(P, Q_0) - KL(P, Q_t)$

Proof We decompose the space $\mathcal{X}$ into B and the complement B^c to get

$$\int_B \log\left(\frac{P}{Q_t}\right) dP = \int_{\mathcal{X}} \log\left(\frac{P}{Q_t}\right) dP - \int_{B^c} \log\left(\frac{P}{Q_t}\right) dP \tag{62}$$

$$= \mathrm{KL}(P, Q_t) - \int_{B^c} \log\left(\frac{P}{Q_t}\right) dP \tag{63}$$

$$\leq \mathrm{KL}(P, Q_0) - \Delta - \int_{B^c} \log\left(\frac{P}{Q_t}\right) dP, \tag{64}$$

where we used Theorem 5, and letting $\theta = \theta(\varepsilon)$ for brevity, we also have

$$\int_{B^c} \log\left(\frac{P}{Q_t}\right) dP = \int_{B^c} \log\left(\frac{P}{Q_0 \exp\left(\langle\theta, c\rangle - \varphi(\theta)\right)}\right) dP \tag{65}$$

$$= \int_{B^c} \log\left(\frac{P}{Q_0}\right) dP - \int_{B^c} \exp\left(\langle\theta, c\rangle - \varphi(\theta)\right) dP \tag{66}$$

$$\geq \int_{B^c} \log\left(\frac{P}{Q_0}\right) dP - \int_{B^c} \frac{\varepsilon}{2} dP \tag{67}$$

$$= \int_{B^c} \log\left(\frac{P}{Q_0}\right) dP - \frac{\varepsilon}{2}\left(1 - \int_B dP\right) \tag{68}$$

Combining these inequalities together gives us:

$$\int_B \log\left(\frac{P}{Q_t}\right) dP \leq \mathrm{KL}(P, Q_0) - \Delta - \left(\int_{B^c} \log\left(\frac{P}{Q_0}\right) dP - \frac{\varepsilon}{2}\left(1 - \int_B dP\right)\right) \tag{69}$$

$$= \int_{\mathcal{X}} \log\left(\frac{P}{Q_0}\right) dP - \int_{B^c} \log\left(\frac{P}{Q_0}\right) dP - \Delta + \frac{\varepsilon}{2}\left(1 - \int_B dP\right) \tag{70}$$

$$= \int_B \log\left(\frac{P}{Q_0}\right) dP - \Delta + \frac{\varepsilon}{2}\left(1 - \int_B dP\right) \tag{71}$$

■

We are now in a position to prove Theorem 7. Using Lemma 8 into the inequality in Lemma 7 yields

$$\int_B dQ_t \geq \int_B dP - \left(\int_B \log\left(\frac{P}{Q_0}\right) dP - \Delta + \frac{\varepsilon}{2}\left(1 - \int_B dP\right)\right) \tag{72}$$

$$= \left(1 + \frac{\varepsilon}{2}\right)\int_B dP - \frac{\varepsilon}{2} - \int_B \log\left(\frac{P}{Q_0}\right) + \Delta. \tag{73}$$

Reorganising and using the Theorem's notations, we get

$$\textsc{m}(B, Q) \ \geq \ \textsc{m}(B, P) - KL(P, Q_0; B) + \frac{\varepsilon}{2} \cdot J(P, Q; B, \varepsilon), \tag{74}$$

where we recall that $J(P, Q; B, \varepsilon) \doteq \textsc{m}(B, P) + \frac{2\Delta(Q)}{\varepsilon} - 1$. Theorem 6 says that we have in the high boosting regime $2\Delta(Q_T)/\varepsilon \geq (\gamma_P + \gamma_Q)/2 - \theta_T(\varepsilon) \cdot (\gamma_P + \gamma_Q)/2$. Letting $\overline{\gamma} \doteq (\gamma_P + \gamma_Q)/2$ and $K \doteq 4 \log 2$, we have from MBDE in the high boosting regime:

$$\begin{aligned}\frac{2\Delta(Q)}{\varepsilon} \ &\geq \ \overline{\gamma} \cdot \left(1 - \left(\frac{1}{1 + \frac{K}{\varepsilon}}\right)^T\right) \\ &\geq \ \overline{\gamma} \cdot \left(1 - \frac{1}{1 + \frac{TK}{\varepsilon}}\right) \\ &= \overline{\gamma} \cdot \frac{TK}{TK + \varepsilon}.\end{aligned} \tag{75}$$

To have $J(P, Q; B, \varepsilon) \geq -(2/\varepsilon) \cdot \alpha \textsc{m}(B, P)$, it is thus sufficient that

$$\begin{aligned}\textsc{m}(B, P) \ &\geq \ \frac{1}{1 + \frac{2\alpha}{\varepsilon}} \cdot \left(1 - \overline{\gamma} \cdot \frac{TK}{TK + \varepsilon}\right) \\ &= \varepsilon \cdot \frac{\varepsilon + (1 - \overline{\gamma})TK}{(\varepsilon + 2\alpha)(\varepsilon + TK)}.\end{aligned} \tag{76}$$

In this case, we check that we have from (74)

$$\textsc{m}(B, Q) \ \geq \ (1 - \alpha)\textsc{m}(B, P) - KL(P, Q_0; B), \tag{77}$$

as claimed.

1.6 Additional formal results

One might ask what such a strong model of privacy allows to keep from the accuracy standpoint in general. Perhaps paradoxically at first sight, it is not hard to show that privacy can bring approximation guarantees on learning: *if* we learn Q_ε within an ε-mollifier $\mathcal{M}$ (hence, we get ε-privacy for sampling from Q_ε), *then* each time *some* Q_ε in $\mathcal{M}$ accurately fits P, we are guaranteed that the one *we learn* also accurately fits P — albeit eventually more moderately —. We let $Q_\varepsilon(; .)$ denote the density learned, where . is the dataset argument.

Lemma 9 *Suppose* $\exists$ ε*-mollifier* $\mathcal{M}$ *s.t.* $Q_\varepsilon \in \mathcal{M}$*, then* $(\exists P, D', \delta : \textit{KL}(P, Q_\varepsilon(; P')) \leq \delta) \Rightarrow (\forall D, \textit{KL}(P, Q_\varepsilon(; P)) \leq \delta + \varepsilon)$.

Proof The proof is straightforward; we give it for completeness: for any dataset D, we have

$$\begin{aligned}
\mathrm{KL}(P, Q_\varepsilon(; P)) &= \int_{\mathcal{X}} \log\left(\frac{P}{Q_\varepsilon(; P)}\right) dP & (78)\\
&= \int_{\mathcal{X}} \log\left(\frac{P}{Q_\varepsilon(; P')}\right) dP + \int_{\mathcal{X}} \log\left(\frac{Q_\varepsilon(; P)}{Q_\varepsilon(; P')}\right) dP & (79)\\
&\leq \int_{\mathcal{X}} \log\left(\frac{P}{Q_\varepsilon(; P')}\right) dP + \varepsilon \cdot \int_{\mathcal{X}} dP & (80)\\
&= \mathrm{KL}(P, Q_\varepsilon(; P')) + \varepsilon & (81)\\
&\leq \delta + \varepsilon, & (82)
\end{aligned}$$

from which we derive the statement of Lemma 9 assuming $\mathcal{A}$ is ε-IP (the inequalities follow from the Lemma's assumption). ■

In the jargon of (computational) information geometry [1], we can summarize Lemma 9 as saying that if there exists an eligible[1] density in a small KL-ball relatively to P, we are guaranteed to find a density also in a small KL-ball relatively to P. This result is obviously good when the premises hold true, but it does not tell the full story when they do not. In fact, when there exists an eligible density outside a big KL-ball relatively to P, it is not hard to show using the same arguments as for the Lemma that we *cannot* find a good one, and this is not a feature of MBDE: this would hold regardless of the algorithm. This limitation is intrinsic to the likelihood ratio constraint of differential privacy, as the following Lemma shows. In the context of ε-DP, we assume that all input datasets have the same size, say m.

Lemma 10 *Let* $\mathcal{A}$ *denote an algorithm learning an* ε*-differentially private density. Denote* $D \sim P$ *an input of the algorithm and* $\mathcal{Q}_\varepsilon(D)$ *the set of all densities that can be the output of* $\mathcal{A}$ *on input* D*, taking in considerations all internal randomisations of* $\mathcal{A}$*. Suppose there exists an input* D' *for which one of these densities is far from the target:* $\exists D', \exists Q \in \mathcal{Q}_\varepsilon(D') : \textit{KL}(P, Q(; D')) \geq \Delta$ *for some "big"* $\Delta > 0$*. Then the output* Q *of* $\mathcal{A}$ *obtained from* ***any*** *input* $D \sim P$ *satisfies:* $\textit{KL}(P, Q(; D)) \geq \Delta - m\varepsilon$.

Proof Denote D the actual input of $\mathcal{A}$. There exists a sequence $\mathcal{D}$ of datasets of the same size, whose length is at most m, which transforms D into D' by repeatedly changing one observation in the current dataset: call it $\mathcal{D} = \{D, D_1, D_2, ..., D_k, D'\}$, with $k \leq m - 1$. Denote $Q(; D'')$ any

[1] Within the chosen ε-mollifier.

element of $\mathcal{Q}_\varepsilon(D'')$ for $D'' \in \mathcal{D}$. Since $\mathcal{A}$ is ε-differentially private, we have:

$$\begin{aligned}
\Delta &\leq \mathrm{KL}(P, Q(; D')) && (83)\\
&= \int_{\mathcal{X}} \log\left(\frac{P}{Q(; D')}\right) dP && (84)\\
&= \int_{\mathcal{X}} \log\left(\frac{P}{Q(; D)}\right) dP + \int_{\mathcal{X}} \log\left(\frac{Q(; D)}{Q(; D_1)}\right) dP + \sum_{j=1}^{k-1} \int_{\mathcal{X}} \log\left(\frac{Q(; D_j)}{Q(; D_{j+1})}\right) dP + \int_{\mathcal{X}} \log\left(\frac{Q(; D_k)}{Q(; D')}\right) dP && (85)\\
&= \mathrm{KL}(P, Q(; D)) + \int_{\mathcal{X}} \log\left(\frac{Q(; D)}{Q(; D_1)}\right) dP + \sum_{j=1}^{k-1} \int_{\mathcal{X}} \log\left(\frac{Q(; D_j)}{Q(; D_{j+1})}\right) dP + \int_{\mathcal{X}} \log\left(\frac{Q(; D_k)}{Q(; D')}\right) dP && (86)\\
&\leq \mathrm{KL}(P, Q(; D)) + m\varepsilon, && (87)
\end{aligned}$$

from which we derive the statement of Lemma 10. ■

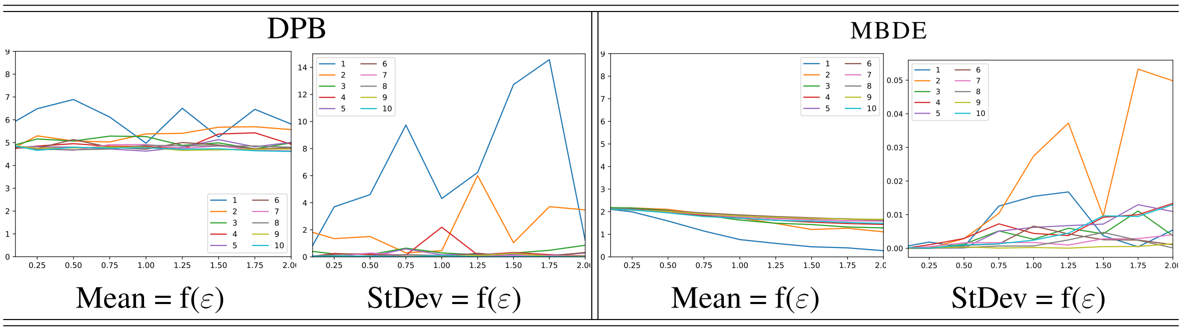


Figure 1: NLL metrics (mean and standard deviation) on the 1D random Gaussian problem for DPB (left pane) and MBDE (right pane), for a varying number of $m = 1, \ldots, 10$ random Gaussians. The lower the better on each metric. Remark the different scales for StDev (see text).

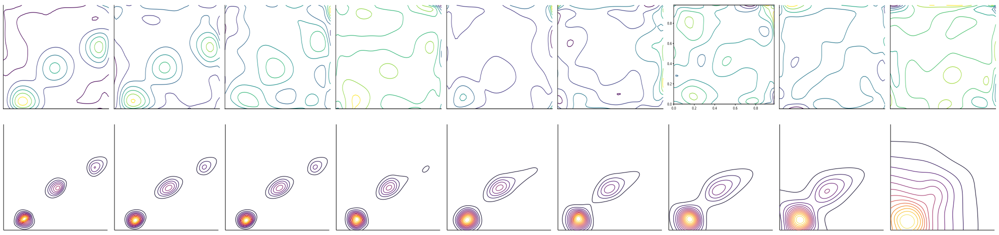


Figure 2: Randomly placed Gaussian convergence comparison for DPB (upper) against MBDE (lower).

2 Additional experiments

We provide here additional results to the main file. Figure 1 provides NLL values for the random 1D Gaussian problem. Figure 2 displays that picking Q_0 a standard Gaussian does not prevent to obtain good results — and beat DPB — when sampling random Gaussians.

References

[1] J.-D. Boissonnat, F. Nielsen, and R. Nock. Bregman voronoi diagrams. *DCG*, 44(2):281–307, 2010.

Chapter 5

Distributional Robustness with IPMs and Links to GANs and Autoencoders

This chapter will begin the study on the robustification aspects of regularization. There are links between Lipschitz and variance regularization to Distributional Robust Optimization (DRO) in the form of upper bounds. These results are extremely compelling as they explain much empirical work and naturally begin a bridge between regularization and robustness. Two questions remain largely unanswered, however:

- Does a similar result hold for other regularization schemes beyond Lipschitz and variance?
- How tight are the upper bounds for previous results?

In this work, our aim is to answer the above two questions and discovered a general result between regularization penalties and DRO, linked with Integral Probability Metrics (IPMs). For example, this includes generalized variance penalties and manifold regularization. For the latter question, equality conditions are characterized, which happens to be intimately related to regularized binary classification. In particular, we draw out this link into an application to GANs, making another contribution to generative models. It is discovered that regularizing discriminators, similar to the previous papers, benefits the performance of GANs and therefore provides foundations for several existing works such as MMD-, Sobelov- and Fisher-GAN.

Distributional Robustness with IPMs and links to Regularization and GANs

Hisham Husain
The Australian National University & Data61
hisham.husain@anu.edu.au

Abstract

Robustness to adversarial attacks is an important concern due to the fragility of deep neural networks to small perturbations and has received an abundance of attention in recent years. Distributional Robust Optimization (DRO), a particularly promising way of addressing this challenge, studies robustness via divergence-based uncertainty sets and has provided valuable insights into robustification strategies such as regularisation. In the context of machine learning, majority of existing results have chosen f-divergences, Wasserstein distances and more recently, the Maximum Mean Discrepancy (MMD) to construct uncertainty sets. We extend this line of work for the purposes of understanding robustness via regularization by studying uncertainty sets constructed with Integral Probability Metrics (IPMs) - a large family of divergences including the MMD, Total Variation and Wasserstein distances. Our main result shows that DRO under *any* choice of IPM corresponds to a family of regularization penalties, which recover and improve upon existing results in the setting of MMD and Wasserstein distances. Due to the generality of our result, we show that other choices of IPMs correspond to other commonly used penalties in machine learning. Furthermore, we extend our results to shed light on adversarial generative modelling via f-GANs, constituting the first study of distributional robustness for the f-GAN objective. Our results unveil the inductive properties of the discriminator set with regards to robustness, allowing us to give positive comments for a number of existing penalty-based GAN methods such as Wasserstein-, MMD- and Sobolev-GANs. In summary, our results intimately link GANs to distributional robustness, extend previous results on DRO and contribute to our understanding of the link between regularization and robustness at large.

1 Introduction

Robustness to adversarial attacks is an important concern due to the fragility of deep neural networks to small perturbations and has received an abundance of attention in recent years [21, 50, 31]. Distributionally Robust Optimization (DRO), a particularly promising way of addressing this challenge, studies robustness via divergence-based uncertainty sets and considers robustness against shifts in distributions. To see this more clearly, for some space Ω, model $h : \Omega \to \mathbb{R}$ and training data $\hat{P}$ with empirical loss $\mathbb{E}_{x\sim\hat{P}}[l_f]$, DRO when applied to machine learning studies the objective $\sup_{Q\in\mathcal{U}} \mathbb{E}_{x\sim Q}[l_f]$ where $\mathcal{U} = \left\{Q : d(Q, \hat{P}) \leq \varepsilon\right\}$ for a given divergence d and $\varepsilon > 0$ that characterize the adversary. Work along this line has shown that this objective is upper bounded by the empirical loss $\mathbb{E}_{x\sim\hat{P}}[l_f]$ plus a penalty term that plays the role of a regularizer, consequently providing formal connections and valuable insights into regularization as a robustification strategy [22, 27, 36, 5, 14, 11].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

The choice of d is crucial as it highlights the strength and nature of robustness we desire, and different choices yield differing penalties. It has been shown that minimizing the distributionally robust objective when d is chosen to be an f-divergence is roughly equivalent to variance regularization [22, 27, 36]. However, there is a problem with this choice of d, as highlighted in [48]: every distribution in the uncertainty set is required to be absolutely continuous with respect to P. This is particularly problematic in the case when P is empirical since every distribution in $\mathcal{U}$ will be finitely supported, meaning that the population distribution will not be contained as it is typically continuous.

Choosing the Wasserstein distance as d is a typical antidote for this problem, and much work has been invested in this direction, explicating connections to Lipschitz regularization [20, 10, 44, 42, 11]. More recently, uncertainty sets based on the kernel Maximum Mean Discrepancy (MMD) were investigated to address concerns with the f-divergence and discovered links to regularization with Hilbert space norms. Both the Wasserstein distance and MMD are part of a larger family of divergences referred to as Integral Probability Metrics (IPM) [35], which are characterized by a set of functions $\mathcal{F}$, and include other metrics such as the Total Variation distance and the Dudley Metric [47].

In this work, we generalize these results and study DRO for uncertainty sets induced by the Integral Probability Metric (IPM) for *any* set of functions $\mathcal{F}$. We present an identity which links distributional robustness under these uncertainty sets $\mathcal{U}_{\mathcal{F}}$, to regularization under a new penalty $\Lambda_{\mathcal{F}}$. Our identity takes the form

$$\boxed{\sup_{Q \in \mathcal{U}_{\mathcal{F}}} \int_{\Omega} h dQ = \int_{\Omega} h dP + \Lambda_{\mathcal{F}}(h)} \tag{1}$$

The appeal of this result is that it reduces the infinite-dimensional optimization on the left-hand side into a penalty-based regularization problem on the right-hand side. We study properties of this penalty and show that it can be upper bounded by another term, $\Theta_{\mathcal{F}}$, which recovers and improves upon existing penalties when $\mathcal{F}$ is chosen to coincide with the MMD and Wasserstein distances. Our result, however, holds in much more generality, allowing us to derive new penalties by considering other IPMs such as the Total Variation, Fisher IPM [33], and Sobelov IPM [32]. We find that these new penalties are related to existing penalties in regularized critic losses [51] and manifold regularization [4], permitting us to provide untried robustness perspectives for existing regularization schemes. Furthermore, most work in this direction takes the form of upper bounds, and although working with $\Theta_{\mathcal{F}}$ reduces (1) into an inequality, we present a necessary and sufficient condition such that $\Lambda_{\mathcal{F}}$ coincides with $\Theta_{\mathcal{F}}$, yielding equality. This condition reveals an intimate connection between distributional robustness and regularized binary classification.

We then apply our result to understanding the distributional robustness of Generative Adversarial Networks (GANs), a popular method for modelling distributions that learn a model Q by utilizing a set of discriminators D that try to distinguish Q from P (the training data). This is particularly relevant for the robustness community since lines of work [53, 9, 58, 57, 28, 26, 39, 45, 46, 24, 55, 40] implement GANs as a robustifying mechanism by training a binary classifier on the learned GAN distribution. Our analysis applies to the f-GAN objective [37] - a loss that subsumes many existing GAN losses. This is, to the best of our knowledge, the first analysis of robustness for f-GANs with respect to divergence-based uncertainty sets. The main insight of our result is the advocation of regularized discriminators when training GANs. In particular, we show that the generative distribution learned using regularized discriminators gives guarantees on the worst-case perturbed distribution (robustness). Our findings complement existing empirical benefits of regularized discriminators such as the MMD-GAN [29, 2, 6], Wasserstein-GAN [3, 23], Sobelov-GAN [32], Fisher-GAN [33] and other penalty-based GANs [51].

Our contributions come in three Theorems, where the first two concern DRO with IPMs (Section 3) and the third is an extension to understanding GANs (Section 4):
▷ **(Theorem 1)** An identity for distributional robustness using uncertainty sets induced by any IPM. Our result tells us that this is *exactly* equal to regularization with a penalty $\Lambda_{\mathcal{F}}$. We show that this penalty can be upper bounded by another penalty $\Theta_{\mathcal{F}}$ which recovers existing work when the IPM is set to the MMD and Wasserstein distance, tightening these results. Since our result holds in much more generality, we derive penalties for other IPMs such as the Total Variation, Fisher IPM, and Sobelov IPM, and draw connections to existing methods.
▷ **(Theorem 2)** A necessary and sufficient condition under which the penalties $\Lambda_{\mathcal{F}}$ and $\Theta_{\mathcal{F}}$ coincide. It turns out this condition is linked to regularized binary classification and is related to critic losses

appearing in penalty-based GANs. This allows us to give positive results for work in this direction, along with drawing a link between regularized binary classification and distributional robustness.
▷ **(Theorem 3)** A result that characterizes the distributional robustness of the f-GAN objective showing that the discriminator set plays an important part for the robustness of a GAN. This is, to the best of our knowledge, the first result on divergence-based distributional robustness of f-GANs. Our result allows us to provide a novel perspective for several existing penalty-based GAN methods such as Wasserstein-, MMD-, and Sobelov-GANs.

2 Preliminaries

2.1 Notation

We will use Ω to denote a compact Polish space and denote Σ as the standard Borel σ-algebra on Ω and $\mathbb{R}$ will denote the real numbers. We use $\mathscr{F}(\Omega, \mathbb{R})$ to denote the set of all bounded and measurable functions mapping from Ω into $\mathbb{R}$ with respect to Σ, $\mathscr{B}(\Omega)$ to be the set of finite signed measures and the set $\mathscr{P}(\Omega) \subset \mathscr{B}(\Omega)$ will denote the set of probability measures. For any additive monoid X, a function $f : X \to \mathbb{R}$ is subadditive if $f(x + x') \leq f(x) + f(x')$ and the *infimal convolution* between two functions $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ is another function given by $(f \barwedge g)(x) = \inf_{x' \in X} (f(x') + g(x - x'))$. For any proposition $\mathscr{I}$, the inversion bracket is $[\![\mathscr{I}]\!] = 1$ if $\mathscr{I}$ is true and 0 otherwise. We say a set of functions $\mathcal{F}$ is even if $h \in \mathcal{F}$ implies $-h \in \mathcal{F}$. For a function $h \in \mathscr{F}(\Omega, \mathbb{R})$ and metric $c : \Omega \times \Omega \to \mathbb{R}$, the Lipschitz constant of h (w.r.t c) is $\operatorname{Lip}_c(h) = \sup_{\omega, \omega' \in \Omega} |h(\omega) - h(\omega')| / c(\omega, \omega')$ and $\|h\|_\infty := \sup_{\omega \in \Omega} |h(\omega)|$. For any set of functions $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, we use $\overline{\operatorname{co}}(\mathcal{F})$ to denote the closed convex hull of $\mathcal{F}$. For a function $h \in \mathscr{F}(\Omega, \mathbb{R})$ and measure $\mu \in \mathscr{P}(\Omega)$, we use $\operatorname{Var}_\mu(h) = \mathbb{E}_\mu[h^2] - \mathbb{E}_\mu[h]^2$ to denote the variance of h under μ.

2.2 Background and Related Work

We will focus our discussion around Distributionally Robust Optimization (DRO) [41] and its use for understanding machine learning. For a given reference distribution P, which is typically the training data in machine learning, the neighbourhood takes the form $\{Q : d(Q, P) \leq \varepsilon\}$ for some divergence d and $\varepsilon > 0$ that characterize the nature and budget of robustness. In the context of machine learning, the most popular choices of d studied thus far are the f-divergences [5, 13, 27], Wasserstein distance [16, 1, 7] and the kernel Maximum Mean Discrepancy (MMD) [48]. For two distributions P, Q, the f-divergence is $d_f(P, Q) = \int_\Omega f(dP/dQ)dQ$ and the main advancement regarding f-divergences, centered around χ^2-divergence, is the connection to variance regularization [22, 27, 36]. This is appealing since it reflects the classical bias-variance trade-off. In contrast, variance regularization also appears in our results, under the choice of μ-Fisher IPM. One of the drawbacks of using f-divergences as pointed out in [48], is that the uncertainty set induced by f-divergences contains only those distributions that share support (since we require absolute continuity) and thus will typically not include the population distribution. The Wasserstein distance is commonly antidotal for these problems since it is defined between distributions that do not share support and DRO results have been developed for this direction, with the main results showing links to Lipschitz regularization [20, 10, 44, 42, 11]. Another distance used to remedy this problem is the Maximum Mean Discrepancy, which has been studied in [48] and shown connections to Hilbert space norm regularization and kernel ridge regression. Since both of these are Integral Probability Metrics (IPMs) [35], it is natural to study uncertainty sets generated by general IPMs:

Definition 1 (Integral Probability Metric) *For any $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, the ($\mathcal{F}$-)Integral Probability Metric between $P, Q \in \mathscr{P}(\Omega)$ is*

$$d_{\mathcal{F}}(P, Q) := \sup_{h \in \mathcal{F}} \left(\int_\Omega h dP - \int_\Omega h dQ \right).$$

The IPM is characterized by a set $\mathcal{F}$ and if $\mathcal{F}$ is even, then $d_{\mathcal{F}}$ is symmetric. One should note that we have an intersection between IPMs and f-divergence when $\mathcal{F} = \{h : \|h\|_\infty \leq 1\}$ and $f(t) = |t - 1|$, which corresponds to the Total Variation. Other cases when they intersect have been thoroughly pursued in [47]. Another interesting case is the 1-Wasserstein distance, which is realized when

Table 1

IPM	$\mathcal{F}$	$\Theta_{\mathcal{F}}(h)$
Wasserstein Distance	$\{h : \mathrm{Lip}_c(h) \leq 1\}$	$\mathrm{Lip}_c(h)$
Maximum Mean Discrepancy	$\{h : \left\Vert h \right\Vert_k \leq 1\}$	$\left\Vert h \right\Vert_k$
Total Variation	$\{h : \left\Vert h \right\Vert_\infty \leq 1\}$	$\left\Vert h \right\Vert_\infty$
Dudley Metric	$\{h : \left\Vert h \right\Vert_\infty + \mathrm{Lip}_c(h) \leq 1\}$	$\left\Vert h \right\Vert_\infty + \mathrm{Lip}_c(h)$
μ-Sobelov IPM	$\left\{h : \mathbb{E}_{\mu(X)}\left[\left\Vert \nabla h(x) \right\Vert^2\right] \leq 1\right\}$	$\sqrt{\mathbb{E}_{\mu(X)}\left[\left\Vert \nabla h(X) \right\Vert^2\right]}$
μ-Fisher IPM	$\left\{h : \mathbb{E}_{\mu(X)}\left[h^2(X)\right] \leq 1\right\}$	$\sqrt{\mathbb{E}_{\mu(X)}\left[h^2(X)\right]}$

$\mathcal{F} = \{h : \mathrm{Lip}_c(h) \leq 1\}$ for some ground metric $c : \Omega \times \Omega \to \mathbb{R}$ [52]. Table 1 contains other known choices of IPMs. As the IPM can be viewed as matching moments specified by $\mathcal{F}$, there is similar work which considers uncertainty sets that match the first and second moment such as [12]. In the context of machine learning our work is, to the best of our knowledge, the first study of the general IPM to understand regularization. Outside this realm, there exist pursuits to study structural properties of IPM-based uncertainty sets such as invariance [43]. While these are important to understand, they, however, do not give immediate consequences for machine learning.

3 Distributional Robustness

In this section, we first introduce the uncertainty set and two complexity measures that form building blocks of the main penalty term $\Lambda_{\mathcal{F}}$ (as appearing in Equation 1), then proceed to the main distributional robustness Theorem.

Definition 2 *For any $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, $P \in \mathscr{P}(\Omega)$, the $\mathcal{F}$-ball centered at P with radius ε is defined to be $B_{\varepsilon,\mathcal{F}}(P) = \{Q \in \mathscr{P}(\Omega) : d_{\mathcal{F}}(Q, P) \leq \varepsilon\}$.*

We now introduce a complexity measure that will be of central importance when defining the penalty: For a function set $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ and function $h \in \mathscr{F}(\Omega, \mathbb{R})$, we set $\Theta_{\mathcal{F}}(h) := \inf\{\lambda > 0 : h \in \lambda \cdot \overline{\mathrm{co}}(\mathcal{F})\}$. This quantity represents the smallest lambda that multiplicatively stretches the set $\overline{\mathrm{co}}(\mathcal{F})$ until it contains h. We illustrate this geometrically in Figure 1 for a non-convex case of $\mathcal{F}$ and present examples of $\Theta_{\mathcal{F}}$ in Table 1.

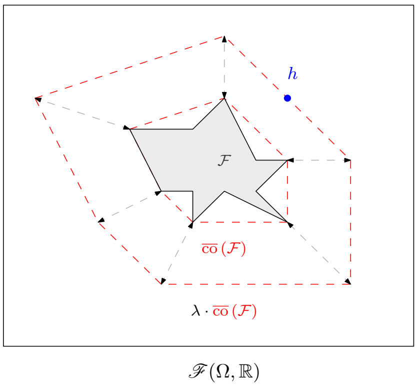

h
$\mathcal{F}$
$\overline{\mathrm{co}}(\mathcal{F})$
$\lambda \cdot \overline{\mathrm{co}}(\mathcal{F})$
$\mathscr{F}(\Omega, \mathbb{R})$

Figure 1: $\Theta_{\mathcal{F}}(h)$ is the smallest multiplicative factor λ required to stretch the convex hull of $\mathcal{F}$ until h is contained.

The second complexity measure depends on a distribution $P \in \mathscr{P}(\Omega, \mathbb{R})$ and is defined as $J_P(h) = \sup_{\nu \in \mathscr{P}(\Omega)} \int_\Omega h d\nu - \int_\Omega h dP$. Note that if h reaches its maximum at some $\omega^* \in \Omega$ then $J_P(h)$ will be smaller if P is concentrated around ω^*. We now present the main penalty, which is infimal convolution of these two complexity measures.

Definition 3 ($\mathcal{F}$-Penalty) *For any $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, $h \in \mathscr{F}(\Omega, \mathbb{R})$ and $\varepsilon > 0$, the $\mathcal{F}$-penalty $\Lambda_{\mathcal{F},\varepsilon} : \mathscr{F}(\Omega, \mathbb{R}) \to [0, \infty]$ is*

$$\Lambda_{\mathcal{F},\varepsilon}(h) = \left(J_P \,\bar{\star}\, \varepsilon\Theta_{\mathcal{F}}\right)(h),$$

where $J_P(h) = \sup_{\nu \in \mathscr{P}(\Omega)} \int_\Omega h d\nu - \int_\Omega h dP$ and $\bar{\star}$ is the infimal convolution operator.

The infimal convolution is central in convex analysis since it is the analogue of addition in the convex dual space [49]. We now present the main theorem, which links this penalty to distributional robustness via $\mathcal{F}$-uncertainty sets and discuss further the role of this penalty.

Theorem 1 *Let $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ and $P \in \mathscr{P}(\Omega)$. For any $h \in \mathscr{F}(\Omega, \mathbb{R})$ and for all $\varepsilon > 0$*

$$\sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \int_\Omega h dQ = \int_\Omega h dP + \Lambda_{\mathcal{F},\varepsilon}(h).$$

Proof (Sketch, full proof in Supplementary material) We can rewrite the constraint over $B_{\varepsilon,\mathcal{F}}(P)$ with the use of a dual variable which leads to a min-max equation. Using generalized minimax theorems [17] and compactness of the set of probability measures, we are able to swap the min-max and solve the inner min using classical results in convex analysis [38], yielding the statement of the theorem. ■

The result allows us to turn the infinite-dimensional optimization on the left-hand side into a familiar penalty-based regularization objective, and we remark that there is no restriction on the choice of $\mathcal{F}$. To see the effect of $\Lambda_{\mathcal{F},\varepsilon}$, notice that by definition of $\barwedge$ we have

$$\Lambda_{\mathcal{F},\varepsilon}(h) = \inf_{\substack{h_1, h_2 \\ h_1 + h_2 = h}} \left(J_P(h_1) + \varepsilon \Theta_{\mathcal{F}}(h_2) \right),$$

which means this penalty finds a decomposition of h into h_1, h_2 so that the two penalties $J_P(h_1)$ and $\varepsilon\Theta_{\mathcal{F}}(h_2)$ are controlled. Notice that any decomposition gives an upper bound, and this is precisely how we will show links and tighten existing results. We will then present a necessary and sufficient condition under which $\Lambda_{\mathcal{F},\varepsilon}(h) = \varepsilon\Theta_{\mathcal{F}}(h)$. This condition plays a fundamental role in linking robustness to regularization and unlike majority of existing results, yields an *equality*.

To see the applicability of the result, consider the supervised learning setup: We have an input space $\mathcal{X}$, output space $\mathcal{Y}$, and a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ which measures performance of a hypothesis $g : \mathcal{X} \to \mathcal{Y}$ on a sample (x, y) with $l(g(x), y)$. In this case, we set $\Omega = \mathcal{X} \times \mathcal{Y}$, P to be the available data, and $h = l(g(x), y)$:

$$\sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \int_\Omega l(g(x), y) dQ(x, y) = \underbrace{\int_\Omega l(g(x), y) dP(x, y)}_{\text{data fitting term}} + \underbrace{\Lambda_{\mathcal{F},\varepsilon}(l(g(x), y))}_{\text{robustness penalty}}.$$

The first term is interpreted as a data fitting term, while the second term is a penalty term that ensures robustness of g. We remark that upper bounds are still favourable in the application of supervised learning, which we will now discuss.

To generate our first upper bound, consider the following decomposition: $h_1 = b$ and $h_2 = h - b$ for some $b \in \mathbb{R}$, yielding the following Corollary.

Corollary 1 *Let $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ and $P \in \mathscr{P}(\Omega)$. For any $h \in \mathscr{F}(\Omega, \mathbb{R})$ and for all $\varepsilon > 0$*

$$\sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \int_\Omega h dQ \leq \int_\Omega h dP + \varepsilon \inf_{b \in \mathbb{R}} \Theta_{\mathcal{F}}(h - b).$$

We will show that Corollary 1 recovers or tightens main results, and holds in much more generality since we may choose *any* set $\mathcal{F}$. The choice of $\mathcal{F}$ is important to our notion of uncertainty as it captures the moments we are interested in, and there is a natural trade-off between picking $\mathcal{F}$ to be too large or too small, which we illustrate with extreme cases. Consider the largest possible set $\mathcal{F} = \mathscr{F}(\Omega, \mathbb{R})$, under which the uncertainty set of distributions, $B_{\varepsilon,\mathcal{F}}(P) = \{P\}$ is a singleton for all $\varepsilon > 0$. This is indeed reflected on the right hand side of Corollary 1, noting that such a strong set $\mathcal{F}$ yields $\Theta_{\mathcal{F}}(h) = 0$ for any $h \in \mathscr{F}(\Omega, \mathbb{R})$. On the other hand, if we pick $\mathcal{F} = \{f(x) = k : k \in \mathbb{R}\}$ to be the set of constants, which is a rather restrictive set, then the uncertainty ball of distributions is the largest it can be $B_{\varepsilon,\mathcal{F}} = \mathscr{P}(\Omega)$ since $d_{\mathcal{F}}(Q, P) = 0$ for all $Q \in \mathscr{P}(\Omega)$. We now focus on non-trivial settings of $\mathcal{F}$, showing that $\Theta_{\mathcal{F}}$ recovers and improves upon familiar existing penalties.

(a) **(Wasserstein Distance)** $\mathcal{F} = \{h : \operatorname{Lip}_c(h) \leq 1\}$. The penalty is $\Theta_{\mathcal{F}}(h) = \operatorname{Lip}_c(h)$, and Corollary 1 recovers the intuition of Lipschitz regularized networks as presented in [20, 10, 44, 42, 8, 11]. However, the penalty in the original theorem $\Lambda_{\mathcal{F},\varepsilon}$ is tighter. To see this by example, consider $\Omega = \mathbb{R}$, P a normal distribution centered at 0 with variance $\sigma > 0$, $h(t) = \sin 2t + t$ and $\varepsilon = 1$. Note that $\varepsilon \operatorname{Lip}_c(h) = 3$ however h can be decomposed into $h_1 = \sin 2t$ and $h_2 = t$ with $J_P(h_1) = 1$ and $\varepsilon \operatorname{Lip}_c(h_2) = 1$. Hence we have $\Lambda_{\mathcal{F},\varepsilon}(h) \leq 2 < 3 = \varepsilon \operatorname{Lip}_c(h)$.

(b) **(Maximum Mean Discrepancy)** $\mathcal{F} = \{h : \|h\|_k \leq 1\}$ where $k : \Omega \times \Omega \to \mathbb{R}$ is a positive definite characteristic kernel and $\|\cdot\|_k$ is the Reproducing Kernel Hilbert Space (RKHS) norm induced by k [34]. For h in the RKHS, the penalty can be bounded by $\Lambda_{\mathcal{F},\varepsilon}(h) \leq \inf_{b \in \mathbb{R}} \|h - b\|_k$. This tightens the existing work on MMD DRO [48, Corollary 3.2] when $b = 0$.

(c) **(Total Variation)** $\mathcal{F} = \{h : \|h\|_\infty \leq 1\}$. Our result tells us that the penalty upper bounded with $\Lambda_{\mathcal{F},\varepsilon}(h) \leq \inf_{b \in \mathbb{R}} \|h - b\|_\infty$, which is tighter than taking $\|h\|_\infty$.

(d) **(μ-Fisher IPM)** $\mathcal{F} = \left\{h : \mathbb{E}_{\mu(X)}\left[h^2(X)\right] \leq 1\right\}$ for some $\mu \in \mathscr{P}(\Omega)$ [33]. The penalty is $\Theta_{\mathcal{F}}(h) = \sqrt{\mathbb{E}_{\mu(X)}\left[h^2(X)\right]}$, however we can solve the infimum in Corollary 1 to get $\inf_{b \in \mathbb{R}} \Theta_{\mathcal{F}}(h - b) = \sqrt{\mathrm{Var}_\mu(h)}$ (Lemma 9 in Supplementary). This is interesting since the variance of h as a penalty has appeared in work studying f-divergence uncertainty sets. Note that when $\mu = (P+Q)/2$ for some $P, Q \in \mathscr{P}(\Omega)$ then $d_{\mathcal{F}}(P, Q)$ is related to the χ^2-divergence, the central f-divergence in these lines of work. In this setting, Corollary 1 extends the interpretation of variance regularization as a robustification strategy for any $\mu \in \mathscr{P}(\Omega)$.

Another interesting choice of $\mathcal{F}$ is the μ-Sobelov IPM which we show in Table 1, whereby the resulting penalty is similar to those existing in manifold regularization [4]. All IPMs considered so far are of the form $\{h : \zeta(h) \leq 1\}$ for some $\zeta : \mathscr{F}(\Omega, \mathbb{R}) \to [0, \infty]$, and the resulting $\Theta_{\mathcal{F}}(h)$ closely resembles $\zeta(h)$. We derive $\Theta_{\mathcal{F}}$ for this general form with some assumptions on ζ.

Lemma 1 *Let $\zeta : \mathscr{F}(\Omega, \mathbb{R}) \to [0, \infty]$ be such that for some $k > 0$, $\zeta(a \cdot h) = a^k \cdot \zeta(h)$ for any $h \in \mathscr{F}(\Omega, \mathbb{R}), a > 0$. If $\mathcal{F} = \{h : \zeta(h) \leq 1\}$, then $\Theta_{\mathcal{F}}(h) \leq \sqrt[k]{\zeta(h)}$ with equality if ζ is convex.*

Our examples presented in Table 1 have convex choices of ζ with either $k = 1$ or $k = 2$. Using this Lemma, we may also interpret the case of two penalties added together, such as the Dudley metric in Table 1. Furthermore, Lemma 1 can be used for future applications of our work to elucidate robustness perspectives of methods using penalties of the form $\sqrt[k]{\zeta(h)}$.

We now return to the discussion on how closely related $\Lambda_{\mathcal{F},\varepsilon}$ is to $\varepsilon\Theta_{\mathcal{F}}$. Consider now two decompsitions of h for the infimal convolution: $h_1 = 0, h_2 = h$ and $h_1 = h, h_2 = 0$, so we have $\Lambda_{\mathcal{F},\varepsilon}(h) \leq \varepsilon\Theta_{\mathcal{F}}(h)$ and $\Lambda_{\mathcal{F},\varepsilon}(h) \leq J_P(h)$ respectively. This yields $\Lambda_{\mathcal{F},\varepsilon}(h) \leq \min\left(J_P(h), \varepsilon\Theta_{\mathcal{F}}(h)\right)$, and we illustrate the tightness of this inequality through the following lemma.

Lemma 2 *The mapping $h \mapsto \Lambda_{\mathcal{F},\varepsilon}(h)$ is subadditive and $\Lambda_{\mathcal{F},\varepsilon}(h)$ is the largest subadditive function that minorizes $\min\left(J_P(h), \varepsilon\Theta_{\mathcal{F}}(h)\right)$.*

The consequence of Lemma 2 is that if $\min\left(J_P(h), \varepsilon\Theta_{\mathcal{F}}(h)\right)$ is subadditive then $\Lambda_{\mathcal{F},\varepsilon}(h) = \min\left(J_P(h), \varepsilon\Theta_{\mathcal{F}}(h)\right)$ since a function always minorizes itself. In the proof of Lemma 2, we show that both J_P and $\varepsilon\Theta_{\mathcal{F}}$ are subadditive and so if $\min\left(J_P, \varepsilon\Theta_{\mathcal{F}}\right)$ is consistently equal to either J_P or $\varepsilon\Theta_{\mathcal{F}}$ for some ε then we have equality.

We now present a necessary and sufficient condition for a function $h : \Omega \to \mathbb{R}$ so that $\Lambda_{\mathcal{F},\varepsilon}(h) = \varepsilon\Theta_{\mathcal{F}}(h)$ for all $\varepsilon > 0$. In doing so, not only do we lead to a better understanding of distributional robustness, we also contribute to understanding tightness of previous results and inequalities subsumed by Corollary 1. It turns out rather surprisingly that the characterization is directly related to penalty-regularized critic losses.

Theorem 2 *A function $h \in \mathscr{F}(\Omega, \mathbb{R})$ satisfies $\Lambda_{\mathcal{F},\varepsilon}(h) = \Theta_{\mathcal{F}}(h)$ if and only if*

$$h \in \operatorname*{arg\,inf}_{\hat{h} \in \mathscr{F}(\Omega, \mathbb{R})} \left(\mathbb{E}_P[\hat{h}] - \mathbb{E}_\mu[\hat{h}] + \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right), \tag{2}$$

for some $\mu \in \mathscr{P}(\Omega)$.

First, note that this characterization holds for any h as long as one can find a μ that satisfies Equation (2). In particular, when $\mu = P$, then the minimizers of Equation (2) are constant functions. Furthermore, Equation (2) can be viewed as a regularized binary classification objective in the following way: Ω is the input space, $Y = \{-1, +1\}$ is the label space, $\hat{h} : \Omega \to \mathbb{R}$ is the classifier, $\Theta_{\mathcal{F}}$ is a penalty with weight ε, and P (resp. μ) corresponds to the -1 (resp. $+1$) class conditional distribution. In particular, this is precisely the objective for the discriminator in penalty-based GANs [23, 51], referred to as the critic loss where P is the fake data generated by a model and μ is the real data. Intuitively, the discriminator function will assign negative values to regions of μ and positive values to regions of P. The discriminator function is then used to guide learning of the model generator by focusing on moving μ to where h assigns higher values. In conjunction with Theorem 1, this discriminator is robust to shifts to the distribution P and we outline the consequence more clearly in the following Corollary.

Corollary 2 *Let $P_+, P_- \in \mathscr{P}(\Omega)$ and suppose $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ is even. If*

$$h^* \in \operatorname*{arg\,inf}_{\hat{h} \in \mathscr{F}(\Omega,\mathbb{R})} \left(\mathbb{E}_{P_-}[\hat{h}] - \mathbb{E}_{P_+}[\hat{h}] + \varepsilon \Theta_{\mathcal{F}}(\hat{h}) \right), \tag{3}$$

then we have

$$\inf_{Q \in B_{\varepsilon,\mathcal{F}}(P_+)} \int_\Omega h^* dQ = \int_\Omega h^* dP_+ - \varepsilon \Theta_{\mathcal{F}}(h^*)$$
$$\sup_{Q \in B_{\varepsilon,\mathcal{F}}(P_-)} \int_\Omega h^* dQ = \int_\Omega h^* dP_- + \varepsilon \Theta_{\mathcal{F}}(h^*).$$

The implication of this corollary is that the classifier learned by solving Equation (3) is still positive (resp. negative) around $B_{\varepsilon,\mathcal{F}}$ neighborhoods of P_+ (resp. P_-). In the context of GANs, P_+ and P_- will be the real and fake distributions. This is a rather intuitive result since the classifier h^* is penalized against $\Theta_{\mathcal{F}}$ however the above Corollary gives formal perspectives along with interpretations to the weighting ε and the choice of penalty (induced by $\mathcal{F}$). We write this Corollary in a more general form since we believe it can be useful for other studies of robustness. An example of this is robustness certification: Given a binary classifier and reference distribution ρ, one can compute $\mathbb{E}_{\rho(X)}[h(X)] - \varepsilon\Theta_{\mathcal{F}}(-h)$ and check if this value is ≥ 0. Using Definition 2.2 of [14] and Corollary 1 of our work, if this value is ≥ 0 then this certifies that the classifier is robust to $\mathcal{F}$-IPM perturbations around ρ. This follows from the fact that Corollary 1 (using $-h$) implies $\mathbb{E}_{\rho(X)}[h(X)] - \varepsilon\Theta_{\mathcal{F}}(-h) \leq \inf_{Q \in B_{\varepsilon,\mathcal{F}}(\rho)} \mathbb{E}_{Q(X)}[h(X)]$ and positivity of the term on the right is precisely the condition laid out in Definition 2.2 of [15]. Corollary 2 uses the fact that the condition outlined in Theorem 2 is sufficient; however, we emphasize that it is also necessary, suggesting an intimate link between regularized binary-classification and distributional robustness.

4 Distributional Robustness of f-GANs

In this section, we show how our main theorem can naturally be applied into the robustness for f-GANs more generally. This is particularly relevant for the robustness community since as mentioned in the introduction, GANs are implemented as a robustifying mechanism for training binary classifiers. In this setting, Ω will typically be a high dimensional Euclidean space to represent the set of images and $P \in \mathscr{P}(\Omega)$ will be an empirical distribution that we are interested in modelling. The model distribution, also referred to as the generative distribution denoted as $\mu \in \mathscr{P}(\Omega)$, is learned by minimizing a divergence between P and μ. We now introduce the f-GAN objective, which is a central divergence in the GAN paradigm.

Definition 4 (f-GAN, [37]) *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a lower semicontinuous convex function with $f(1) = 0$ and $\mathcal{H} \subset \mathscr{F}(\Omega, \operatorname{dom} f^\star)$ be a set of discriminators. The GAN objective for data $P \in \mathscr{P}(\Omega)$ and model $\mu \in \mathscr{P}(\Omega)$ is*

$$\mathrm{GAN}_{f,\mathcal{H}}(\mu; P) = \sup_{h \in \mathcal{H}} \left(\int_\Omega h dP - \int_\Omega f^\star(h) d\mu \right),$$

where $f^\star(y) = \sup_{x \in \mathbb{R}} (x \cdot y - f(x))$ is the convex conjugate.

We are interested in minimizing the above objective with respect to μ, which results in a min-max objective due to the supremum taken over $\mathcal{H}$. One should note that there are two components of this objective that characterize it, the function f and discriminator set $\mathcal{H}$. In practice, the discriminator set is often restricted, and so the resulting objective is not a divergence; however, empirical studies have observed convergence [19], which warrants an investigation into the effects of a restricted discriminator on model performance. Existing theoretical work has hinted the benefits of a restricted discriminator, for example, [56] show that generalization is related to the Rademacher complexity of the discriminator set and suggest a discrimination-generalization trade-off. Other work has suggested that the particular setting of Lipschitz discriminators leads to improvements for both practical [56, 19, 59, 54, 18] and theoretical purposes [25, 18, 30]. It is clear that the discriminator set is a key character in the tale of success of GANs; however, the existing literature is silent on what it means for robustness, a particular application that GANs have posed successful in, and this is precisely the link we establish with the following Theorem.

Theorem 3 *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex lower semi-continuous function with $f(1) = 0$, $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ and $\mathcal{H} \subseteq \mathscr{F}(\Omega, \operatorname{dom}(f^\star))$. For any model and data distributions $\mu, P \in \mathscr{P}(\Omega)$ respectively, we have for all $\varepsilon > 0$*

$$\sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \mathrm{GAN}_{f,\mathcal{H}}(\mu; Q) \leq \mathrm{GAN}_{f,\mathcal{H}}(\mu; P) + \varepsilon \sup_{h \in \mathcal{H}} \Theta_{\mathcal{F}}(h).$$

This Theorem tells us that the robust version of the GAN objective can be upper bounded by the standard GAN objective plus a term that quantifies the complexity of the discriminator set. Note that the robustness parameters (ε and $\mathcal{F}$) interact only with the discriminator set and not the generative model μ, revealing the importance of choosing a regularized discriminator set $\mathcal{H}$. To see this more clearly, consider the setting $\mathcal{F} = \mathcal{H}$, and since $\Theta_{\mathcal{H}}(h) \leq 1$, we have

$$\sup_{Q \in B_{\varepsilon,\mathcal{H}}(P)} \mathrm{GAN}_{f,\mathcal{H}}(\mu; Q) \leq \mathrm{GAN}_{f,\mathcal{H}}(\mu; P) + \varepsilon, \tag{4}$$

for all $\varepsilon > 0$. The key insight is that training GANs using discriminators $\mathcal{H}$ yields guarantees on the robust GAN objective for adversaries who pick Q from $B_{\varepsilon,\mathcal{H}}(P)$. Note that if one picks discriminators $\mathcal{H}$ that are too strong then the ball $B_{\varepsilon,\mathcal{H}}(P)$ will shrink and become singleton $\{P\}$ when $\mathcal{H} = \mathscr{F}(\Omega, \mathbb{R})$. On the other hand, if $\mathcal{H}$ is chosen to be smaller then the uncertainty set is larger; however, the first term $\mathrm{GAN}_{f,\mathcal{H}}$ will be a weaker divergence, since the discriminator set determines the strength of the objective [30]. Hence, there is a trade-off between discrimination and robustness, that complements and parallels the discrimination-generalization story described in [56].

We now discuss the particular settings of $\mathcal{F}$ and how our theorem gives a perspective of distributional robustness on existing GAN methods. First, consider choices of $\mathcal{F}$ so that $d_{\mathcal{F}}$ corresponds to MMD, Fisher IPM and Sobelov IPM which translates to the MMD-GAN, Fisher-GAN and Sobelov GAN respectively, allowing us to view these methods from a robustness perspective in light of Theorem 3 and Equation (4). Furthermore, our result also contributes to the positive commentary under the popular choice of Lipschitz regularized discriminators, guarantees against adversaries selecting from Wasserstein uncertainty sets. It should be noted that recently, a method that regularizes discriminators by minimizing a penalty referred to as 0-GP [51] has proven convergence and generalization guarantees. It can be easily shown that this penalty satisfies the conditions of Lemma 1 for $k = 2$ due to its resemblance to the Sobelov IPM, allowing us to present a robustness interpretation for this penalty. Our main insight from this perspective reveals the theoretical benefits of regularized discriminators. In light of our results, learning a binary classifier using a GAN (trained with regularized discriminators) as a downstream task implies this classifier will consequently be robust.

5 Conclusion

Our results extend the Distributionally Robust Optimization (DRO) framework to IPMs, which reveal further importance of the role regularization plays for robustness and machine learning at large. Unlike most DRO applications to machine learning, we present equality and show that achieving this is fundamentally rooted in regularized binary classification. We then show that DRO can be extended to understand GANs and unveil the role of discrimination regularization in these frameworks. The results will also help DRO explain regularization penalties through the lens of robustness in the future. Our contributions are modular and pave the way to build on related areas, one such example being robustness certification, which we leave for the subject of future work.

Broader Impact

From the perspective of impact, the main contribution of our work is understanding how regularization, a commonly used technique in machine training, gives benefits for robustness. We show this for different areas of machine learning, such as supervised learning and generative adversarial networks. The ultimate goal of such work is to develop further our understanding of these methods and how their performances can be improved. Our work does not have a focused application use-case under which we can discuss specific ethical considerations since it contributes more generally to the advancements of performance. In this sense, ethical considerations are subject to the application of these methods.

Acknowledgments and Disclosure of Funding

We would like to thank Jeremias Knoblauch and anonymous reviewers for comments regarding the focus and clarity of presentation. This work was funded by the Australian Government Research Training Program and Data61.

References

[1] Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[2] Michael Arbel, Dougal Sutherland, Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for mmd gans. In *Advances in Neural Information Processing Systems*, pages 6700–6710, 2018.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[4] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[5] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[6] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[7] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[8] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[9] Jeremy Charlier, Aman Singh, Gaston Ormazabal, Radu State, and Henning Schulzrinne. Syngan: Towards generating synthetic network attacks using gans. *arXiv preprint arXiv:1908.09899*, 2019.

[10] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 854–863. JMLR. org, 2017.

[11] Zac Cranko, Zhan Shi, Xinhua Zhang, Richard Nock, and Simon Kornblith. Generalised lipschitz regularisation equals distributional robustness. *arXiv preprint arXiv:2002.04197*, 2020.

[12] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[13] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[14] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *FOCS*, 2013.

[15] KD Dvijotham, J Hayes, B Balle, Z Kolter, C Qin, A Gyorgy, K Xiao, S Gowal, and P Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020.

[16] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[17] Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39(1):42, 1953.

[18] Farzan Farnia and David Tse. A convex duality framework for gans. In *Advances in Neural Information Processing Systems*, pages 5254–5263, 2018.

[19] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.

[20] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[21] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[22] Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. Robust empirical optimization is almost the same as mean–variance optimization. *Operations research letters*, 46(4):448–452, 2018.

[23] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[24] Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 43–49. IEEE, 2018.

[25] Hisham Husain, Richard Nock, and Robert C Williamson. A primal-dual link between gans and autoencoders. In *Advances in Neural Information Processing Systems*, pages 413–422, 2019.

[26] Ajil Jalal, Andrew Ilyas, Constantinos Daskalakis, and Alexandros G Dimakis. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

[27] Henry Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

[28] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with gan. *arXiv preprint arXiv:1705.03387*, 2017.

[29] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.

[30] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5545–5553, 2017.

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[32] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

[33] Youssef Mroueh and Tom Sercu. Fisher gan. In *Advances in Neural Information Processing Systems*, pages 2513–2523, 2017.

[34] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016.

[35] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[36] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in neural information processing systems*, pages 2971–2980, 2017.

[37] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

[38] Jean-Paul Penot. *Calculus without derivatives*, volume 266. Springer Science & Business Media, 2012.

[39] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018.

[40] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018.

[41] Herbert E Scarf. A min-max solution of an inventory problem. Technical report, RAND CORP SANTA MONICA CALIF, 1957.

[42] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[43] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

[44] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2, 2017.

[45] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

[46] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, pages 8312–8323, 2018.

[47] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

[48] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, pages 9131–9141, 2019.

[49] Thomas Strömberg. *A study of the operation of infimal convolution*. PhD thesis, Luleå tekniska universitet, 1994.

[50] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[51] Hoang Thanh-Tung, Truyen Tran, and Svetha Venkatesh. Improving generalization and stability of generative adversarial networks. *arXiv preprint arXiv:1902.03984*, 2019.

[52] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[53] Huaxia Wang and Chun-Nam Yu. A direct approach to robust deep learning using adversarial networks. *arXiv preprint arXiv:1905.09591*, 2019.

[54] Bingzhe Wu, Shiwan Zhao, ChaoChao Chen, Haoyang Xu, Li Wang, Xiaolu Zhang, Guangyu Sun, and Jun Zhou. Generalization in generative adversarial networks: A novel perspective from privacy protection. In *Advances in Neural Information Processing Systems*, pages 306–316, 2019.

[55] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

[56] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

[57] He Zhao, Trung Le, Paul Montague, Olivier De Vel, Tamas Abraham, and Dinh Phung. Perturbations are not enough: Generating adversarial examples with spatial distortions. *arXiv preprint arXiv:1910.01329*, 2019.

[58] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017.

[59] Zhiming Zhou, Jiadong Liang, Yuxuan Song, Lantao Yu, Hongwei Wang, Weinan Zhang, Yong Yu, and Zhihua Zhang. Lipschitz generative adversarial nets. *arXiv preprint arXiv:1902.05687*, 2019.

Supplementary Material for “Distributional Robustness with IPMs and links to Regularization and GANs”

Hisham Husain
The Australian National University & Data61
hisham.husain@anu.edu.au

Abstract

This document contains the proofs for the main results of the submission “Distributional Robustness with IPMs and links to Regularization and GANs”.

Appendix: Table of Contents

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

1 Proofs of Main Results

Before we begin, we introduce some notation that will be used to prove the main results that is exclusive to the Appendix. We will be invoking general convex analysis on the space $\mathscr{F}(\Omega, \mathbb{R})$, in the same fashion as [2], noting that $\mathscr{F}(\Omega, \mathbb{R})$ is a Hausdorff locally convex space (through the uniform norm). We use $\mathscr{B}(\Omega)$ to denote the denote the set of all bounded and finitely additive signed measures over Ω (with a given σ-algebra). For any set $D \subseteq \mathscr{B}(\Omega)$ and $h \in \mathscr{F}(\Omega, \mathbb{R})$, we use $\sigma_D(h) = \sup_{\nu \in D} \langle h, \nu \rangle$ and $\delta_D(\nu) = \infty \cdot [\![\nu \notin D]\!]$ to denote the *support* and *indicator* functions such as in [5]. We introduce the conjugate specific to these spaces

Definition 1 ([6]) *For any proper convex function $F : \mathscr{F}(\Omega, \mathbb{R}) \to (-\infty, \infty)$, we have for any $\mu \in \mathscr{B}(\Omega)$ we define*

$$F^\star(\mu) = \sup_{h \in \mathscr{F}(\Omega,\mathbb{R})} \left(\int_\Omega h d\mu - F(h) \right)$$

and for any $h \in \mathscr{F}(\Omega, \mathbb{R})$ we define

$$F^{\star\star}(h) = \sup_{\mu \in \mathscr{B}(\Omega)} \left(\int_\Omega h d\mu - F^\star(\mu) \right).$$

Theorem 1 ([8] Theorem 2.3.3) *If X is a Hausdorff locally convex space, and $F : X \to (-\infty, \infty]$ is a proper convex lower semi-continuous function then $F^{\star\star} = F$.*

There is an additional robustness result which we will deploying for several proofs which holds for any space A that admits Polish topology.

Lemma 1 *For any $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, we have that*

$$d_{\mathcal{F}}(P, \mu) = d_{\overline{\text{co}}(\mathcal{F})}(P, \mu).$$

Proof Let $\Delta_n := \{\alpha \in [0,1]^n : \sum_{i=1}^n \alpha = 1\}$ Note that we have

$$\begin{aligned}
d_{\text{co}(\mathcal{F})}(P, \mu) &= \sup_{n \in \mathbb{N}, \alpha \in \Delta_n, f_i \in \mathcal{F} \forall i = 1, \ldots, n} \left\{ \mathbb{E}_P \left[\sum_{i=1}^n \alpha_i f_i \right] - \mathbb{E}_\mu \left[\sum_{i=1}^n \alpha_i f_i \right] \right\} \\
&= \sup_{n \in \mathbb{N}, \alpha \in \Delta_n, f_i \in \mathcal{F} \forall i = 1, \ldots, n} \sum_{i=1}^n \alpha_i \left\{ \mathbb{E}_P [f_i] - \mathbb{E}_\mu [f_i] \right\} \\
&= \sup_{n \in \mathbb{N}, \alpha \in \Delta_n} \sum_{i=1}^n \alpha_i \sup_{f_i \in \mathcal{F}} \left\{ \mathbb{E}_P [f_i] - \mathbb{E}_\mu [f_i] \right\} \\
&= \sup_{n \in \mathbb{N}, \alpha \in \Delta_n} \sum_{i=1}^n \alpha_i d_{\mathcal{F}}(P, \mu) \\
&= d_{\mathcal{F}}(P, \mu)
\end{aligned}$$

It is also closed under taking the closure since $d_{\mathcal{F}}$ is the supremum of continuous (linear) functions and the supremum over a set with a linear objective is equal to taking the supremum over the closure of that set. ■

Definition 2 *For any $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, we define the functional $R_{\mathcal{F}} : \mathscr{F}(\Omega, \mathbb{R}) \to [0, \infty]$ as*

$$R_{\mathcal{F}}(h) := \int_\Omega h dP + \delta_{\overline{\text{co}}(\mathcal{F})}(h).$$

Lemma 2 *For any $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, $R_{\mathcal{F}}$ is proper convex and lower semi-continuous.*

Proof The mapping $h \mapsto \int_\Omega hdP$ is clearly convex and lower semi-continuous. Since $\overline{\text{co}}\,(\mathcal{F})$ is a closed and convex set, the indicator function $\delta_{\overline{\text{co}}(\mathcal{F})}(h)$ is proper convex and lower semi-continuous and thus the result follows. ■

Lemma 3 *The mappings $\nu \mapsto d_{\mathcal{F}}(\nu, P)$ and $h \mapsto R_{\mathcal{F}}(h)$ are convex conjugates*

Proof Note first that for any $\nu \in \mathscr{B}(\Omega)$

$$\begin{aligned} R^\star_{\mathcal{F}}(\nu) &= \sup_{h \in \mathscr{F}(\Omega,\mathbb{R})} \left\{ \int_\Omega hd\nu - \int_\Omega hdP - \delta_{\overline{\text{co}}(\mathcal{F})}(h) \right\} \\ &= \sup_{h \in \overline{\text{co}}(\mathcal{F})} \left\{ \int_\Omega hd\nu - \int_\Omega hdP \right\} \\ &= d_{\overline{\text{co}}(\mathcal{F})}(\nu, P) \\ &\overset{(1)}{=} d_{\mathcal{F}}(\nu, P), \end{aligned}$$

where (1) is due to Lemma 1. We also have that

$$\begin{aligned} \left(d_{\mathcal{F}}(\cdot, P)\right)^\star(h) &= \sup_{\nu \in \mathscr{B}(\Omega)} \left\{ \int_\Omega hd\nu - d_{\mathcal{F}}(\nu, P) \right\} \\ &\overset{(1)}{=} \sup_{\nu \in \mathscr{B}(\Omega)} \left\{ \int_\Omega hd\nu - R^\star_{\mathcal{F}}(\nu) \right\} \\ &\overset{(2)}{=} R^{\star\star}_{\mathcal{F}}(\nu) \\ &\overset{(3)}{=} R_{\mathcal{F}}(\nu), \end{aligned}$$

where (1) holds due to the above, (2) holds by definition of conjugate and (3) holds by a combination of Lemma 1 and Lemma 2. ■

We also present a lemma which will prove to be useful in proving the main results.

Lemma 4 *For any $\mathcal{F} \subset \mathscr{F}(\Omega, \mathbb{R})$, the mapping $h \mapsto \Theta_{\mathcal{F}}(h)$ is convex.*

Proof First notice that for any $t > 0$ and $h \in \mathscr{F}(\Omega, \mathbb{R})$ we have that $\Theta_{\mathcal{F}}(t \cdot h) = t \cdot \Theta_{\mathcal{F}}(h)$. For any $t \in [0, 1]$ and $h, h' \in \mathscr{F}(\Omega, \mathbb{R})$, consider the element $\tilde{h} := t \cdot h + (1 - t) \cdot h'$. Since $t \cdot h \in t\Theta_{\mathcal{F}}(h) \cdot \overline{\text{co}}\,(\mathcal{F})$ and $(1 - t)h \in (1 - t)\Theta_{\mathcal{F}}(h) \cdot \overline{\text{co}}\,(\mathcal{F})$, we have that

$$\begin{aligned} &\tilde{h} \in t\Theta_{\mathcal{F}}(h) \cdot \overline{\text{co}}\,(\mathcal{F}) + (1 - t)\Theta_{\mathcal{F}}(h) \cdot \overline{\text{co}}\,(\mathcal{F}) \\ \iff &\tilde{h} \in \left(t\Theta_{\mathcal{F}}(h) + (1 - t)\Theta_{\mathcal{F}}(h')\right) \cdot \overline{\text{co}}\,(\mathcal{F}), \end{aligned}$$

which in turn implies that $\Theta_{\mathcal{F}}(\tilde{h}) \leq t\Theta_{\mathcal{F}}(h) + (1 - t)\Theta_{\mathcal{F}}(h')$, proving convexity of $\Theta_{\mathcal{F}}$. ■

1.1 Proof of Theorem 1

Theorem 2 *Let $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ and $P \in \mathscr{P}(\Omega)$. For any $h \in \mathscr{F}(\Omega, \mathbb{R})$ and for all $\varepsilon > 0$*

$$\sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \int_\Omega hdQ = \int_\Omega hdP + \Lambda_{\mathcal{F},\varepsilon}(h).$$

Proof We first require two lemmata.

Lemma 5 *For any $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, $P \in \mathscr{P}(\Omega)$, $\lambda \geq 0$ and $h \in \mathscr{F}(\Omega, \mathbb{R})$, we have*

$$\sup_{Q \in \mathscr{P}(\Omega)} \left(\int_\Omega hdQ - \lambda d_{\mathcal{F}}(Q, P) \right) = R_{\lambda\mathcal{F}} \,\bar{\star}\, \sigma_{\mathscr{P}(\Omega)}(h)$$

Proof We use a standard result from convex analysis which states that the convex conjugate of the sum of two functions is the infimal convolution of their conjugates. Hence we have

$$\begin{aligned}\sup_{Q\in\mathscr{P}(\Omega)}\left(\int_\Omega hdQ-\lambda d_{\mathcal{F}}(Q,P)\right) &= \sup_{Q\in\mathscr{B}(\Omega)}\left(\int_\Omega hdQ-\lambda d_{\mathcal{F}}(Q,P)-\delta_{\mathscr{P}(\Omega)}(Q)\right)\\ &= \left(\lambda d_{\mathcal{F}}(Q,P)+\delta_{\mathscr{P}(\Omega)}(Q)\right)^\star\\ &= \left(\lambda d_{\mathcal{F}}(Q,P)\right)^\star \mathbin{\bar\star} \left(\delta_{\mathscr{P}(\Omega)}(Q)\right)^\star\\ &= R_{\lambda\mathcal{F}}\mathbin{\bar\star}\sigma_{\mathscr{P}(\Omega)}(h),\end{aligned}$$

which follows from Lemma 3 and the fact that support functions are conjugates of indicator functions [4, Section 3.4.1, Example (a)]. ■

Lemma 6 *For any $\mathcal{F}\subseteq\mathscr{F}(\Omega,\mathbb{R})$, $P\in\mathscr{P}(\Omega)$, and $h\in\mathscr{F}(\Omega,\mathbb{R})$, we have*

$$\inf_{\lambda\geq 0}\left(R_{\lambda\mathcal{F}}\mathbin{\bar\star}\sigma_{\mathscr{P}(\Omega)}(h)+\lambda\varepsilon\right)=\int_\Omega hdP+J_P\mathbin{\bar\star}\varepsilon\Theta_{\mathcal{F}}(h)$$

Proof Using the definition of infimal convolution, we have

$$\begin{aligned}&\inf_{\lambda\geq 0}\left(R_{\lambda\mathcal{F}}\mathbin{\bar\star}\sigma_{\mathscr{P}(\Omega)}(h)+\lambda\varepsilon\right)\\ &=\inf_{\lambda\geq 0}\left(\inf_{h'\in\mathscr{F}(\Omega,\mathbb{R})}\left(\int_\Omega(h-h')dP+\delta_{\overline{\mathrm{co}}(\lambda\mathcal{F})}(h-h')+\sigma_{\mathscr{P}(\Omega)}(h)\right)+\lambda\varepsilon\right)\\ &=\inf_{\lambda\geq 0}\inf_{h'\in\mathscr{F}(\Omega,\mathbb{R})}\left(\int_\Omega hdP-\int_\Omega h'dP+\delta_{\overline{\mathrm{co}}(\lambda\mathcal{F})}(h-h')+\sigma_{\mathscr{P}(\Omega)}(h')+\lambda\varepsilon\right)\\ &=\int_\Omega hdP+\inf_{h'\in\mathscr{F}(\Omega,\mathbb{R})}\left(-\int_\Omega h'dP+\inf_{\lambda\geq 0}\left(\delta_{\overline{\mathrm{co}}(\lambda\mathcal{F})}(h-h')+\lambda\varepsilon\right)+\sigma_{\mathscr{P}(\Omega)}(h')\right)\\ &=\int_\Omega hdP+\inf_{h'\in\mathscr{F}(\Omega,\mathbb{R})}\left(\sigma_{\mathscr{P}(\Omega)}(h')-\int_\Omega h'dP+\inf_{\lambda\geq 0}\left(\delta_{\overline{\mathrm{co}}(\lambda\mathcal{F})}(h-h')+\lambda\varepsilon\right)\right)\\ &=\int_\Omega hdP+\inf_{h'\in\mathscr{F}(\Omega,\mathbb{R})}\left(\sigma_{\mathscr{P}(\Omega)}(h')-\int_\Omega h'dP+\inf_{\lambda\geq 0}\left(\infty\cdot[\![h-h'\notin\lambda\cdot\overline{\mathrm{co}}\left(\mathcal{F}\right)]\!]+\lambda\varepsilon\right)\right)\\ &=\int_\Omega hdP+\inf_{h'\in\mathscr{F}(\Omega,\mathbb{R})}\left(J_P(h')+\varepsilon\Theta_{\mathcal{F}}(h-h')\right)\\ &=\int_\Omega hdP+J_P\mathbin{\bar\star}\varepsilon\Theta_{\mathcal{F}}(h).\end{aligned}$$

■

We are now ready to prove the Theorem. By introducing a dual variable $\lambda>0$ that penalizes the ball constraint, we have

$$\begin{aligned}\sup_{Q\in B_{\varepsilon,\mathcal{F}}(P)}\int_\Omega hdQ &= \sup_{Q\in\mathscr{P}(\Omega):d_{\mathcal{F}}(Q,P)\leq\varepsilon}\int_\Omega hdQ\\ &= \sup_{Q\in\mathscr{P}(\Omega)}\inf_{\lambda\geq 0}\left(\int_\Omega hdQ+\lambda\left(\varepsilon-d_{\mathcal{F}}(Q,P)\right)\right)\\ &\overset{(1)}{=}\inf_{\lambda\geq 0}\sup_{Q\in\mathscr{P}(\Omega)}\left(\int_\Omega hdQ+\lambda\left(\varepsilon-d_{\mathcal{F}}(Q,P)\right)\right)\\ &= \inf_{\lambda\geq 0}\left(\sup_{Q\in\mathscr{P}(\Omega)}\left(\int_\Omega hdQ-\lambda d_{\mathcal{F}}(Q,P)\right)+\lambda\varepsilon\right)\\ &\overset{(2)}{=}\inf_{\lambda\geq 0}\left(R_{\lambda\mathcal{F}}\mathbin{\bar\star}\sigma_{\mathscr{P}(\Omega)}(h)+\lambda\varepsilon\right)\\ &\overset{(3)}{=}\int_\Omega hdP+J_P\mathbin{\bar\star}\varepsilon\Theta_{\mathcal{F}}(h),\end{aligned}$$

where (2) and (3) hold due to Lemma 5 and 6 respectively. To see why (1) holds, first note that the mapping $Q \mapsto \int_\Omega hdQ + \lambda\left(\varepsilon - d_{\mathcal{F}}(Q,P)\right)$ is concave and lower semicontinuous since $d_{\mathcal{F}}$ is the supremum of linear functions. Next we have by an application of the Banach-Alaogu Theorem that $\mathscr{P}(\Omega)$ is compact [2, Lemma 27 (b)]. Hence by [1, Theorem 2], (1) follows. ■

1.2 Proof of Corollary 1

Corollary 1 *Let $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ and $P \in \mathscr{P}(\Omega)$. For any $h \in \mathscr{F}(\Omega, \mathbb{R})$ and for all $\varepsilon > 0$*

$$\sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \int_\Omega hdQ \leq \int_\Omega hdP + \varepsilon \inf_{b \in \mathbb{R}} \Theta_{\mathcal{F}}(h-b).$$

Proof By definition of the infimal convolution we can consider a decomposition of the form $h_1 = b$ and $h_2 = h - b$ for some $b \in \mathbb{R}$. notice that $J_P(b) = 0$ and by taking the smallest possible $b \in \mathbb{R}$ yields

$$\Theta_{\mathcal{F},\varepsilon}(h) \leq \varepsilon \inf_{b \in \mathbb{R}} \Theta_{\mathcal{F}}(h-b),$$

which completes the proof. ■

1.3 Proof of Lemma 1

Lemma 7 *Let $\zeta : \mathscr{F}(\Omega, \mathbb{R}) \to [0, \infty]$ be a penalty such that $\zeta(a \cdot h) = a^k \cdot \zeta(h)$ for any $h \in \mathscr{F}(\Omega, \mathbb{R})$, $k, a > 0$. Let $\mathcal{F} = \{h : \zeta(h) \leq 1\}$ then we have $\Theta_{\mathcal{F}}(h) \leq \sqrt[k]{\zeta(h)}$ with equality if ζ is convex.*

Proof Let us consider the non-convex case so that $\mathcal{F}$ is not necessarily convex. We then have for any $\mathcal{F} \subseteq \mathscr{F}(\Omega, \mathbb{R})$

$$\begin{aligned} h \in \overline{\text{co}}\left(\lambda\mathcal{F}\right) &\iff h \in \lambda\overline{\text{co}}\left(\mathcal{F}\right) \\ &\iff \frac{h}{\lambda} \in \overline{\text{co}}\left(\mathcal{F}\right) \end{aligned}$$

For a fixed $h \in \mathscr{F}(\Omega, \mathbb{R})$, set $\lambda = \sqrt[k]{\zeta(h)}$ and notice that

$$\begin{aligned} \zeta\left(\frac{h}{\lambda}\right) &= \zeta\left(\frac{h}{\sqrt[k]{\zeta(h)}}\right) \\ &= \left(\frac{1}{\sqrt[k]{\zeta(h)}}\right)^k \zeta\left(h\right) \\ &= \zeta\left(h\right), \end{aligned}$$

and so we have $\Theta_{\mathcal{F}}(h) \leq \sqrt[k]{\zeta(h)}$. In the case when the penalty is convex, we have that $\mathcal{F}$ will be convex and so

$$\begin{aligned} h \in \lambda\overline{\text{co}}\left(\mathcal{F}\right) &\iff \frac{h}{\lambda} \in \overline{\text{co}}\left(\mathcal{F}\right) \\ &\iff \frac{h}{\lambda} \in \mathcal{F} \\ &\iff \zeta\left(\frac{h}{\lambda}\right) \leq 1 \\ &\iff \frac{1}{\lambda^k}\zeta(h) \leq 1 \\ &\iff \zeta(h) \leq \lambda^k \\ &\iff \sqrt[k]{\zeta(h)} \leq \lambda. \end{aligned}$$

Hence we have $\Theta_{\mathcal{F}}(h) = \inf_{\sqrt[k]{\zeta(h)} \leq \lambda} \lambda = \sqrt[k]{\zeta(h)}$. ■

1.4 Proof of Lemma 2

Lemma 8 *The mapping* $h \mapsto \Lambda_{\mathcal{F},\varepsilon}(h)$ *is subadditive and* $\Lambda_{\mathcal{F},\varepsilon}(h)$ *is the largest subadditive function that minorizes* $\min\left(J_P(h), \varepsilon\Theta_{\mathcal{F}}(h)\right)$.

Proof Since $\Theta_{\mathcal{F}}(h)$ is convex (Lemma 4) and $\Theta_{\mathcal{F}}(t \cdot h) = t \cdot \Theta_{\mathcal{F}}(h)$ for $t > 0$, it follows that $\Theta_{\mathcal{F}}(h)$ is subadditive. Next notice that J_P is subadditive since for any $h, h' \in \mathscr{F}(\Omega, \mathbb{R})$

$$\begin{aligned} J_P(h + h') &= \sup_{\omega\in\Omega} h(\omega) + h'(\omega) - \int_\Omega h dP - \int_\Omega h' dP \\ &\leq \sup_{\omega\in\Omega} h(\omega) - \int_\Omega h dP + \sup_{\omega\in\Omega} h'(\omega) - \int_\Omega h' dP \\ &= J_P(h) + J_P(h'). \end{aligned}$$

Next notice that $J_P(0) = 0$ and $\varepsilon\Theta_{\mathcal{F}}(0) = 0$. By [7, Theorem 2.5(c)] we have that $\Lambda_{\mathcal{F},\varepsilon}$ is sub-additive and that it is the largest subadditive function that minorizes $\min\left(J_P(h), \varepsilon\Theta_{\mathcal{F}}(h)\right)$. ■

1.5 Proof of Theorem 2

Theorem 3 *A function* $h \in \mathscr{F}(\Omega, \mathbb{R})$ *satisfies* $\Lambda_{\mathcal{F},\varepsilon}(h) = \Theta_{\mathcal{F}}(h)$ *if and only if*

$$h \in \operatorname*{arg\,inf}_{\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})} \left(\mathbb{E}_P[\hat{h}] - \mathbb{E}_\mu[\hat{h}] + \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right),$$

for some $\mu \in \mathscr{P}(\Omega)$.

Proof To prove this Theorem, we use the conditions for an optimal decomposition of an infimal convolution as shown in [3, Lemma 1]. First note that J_P and $\Theta_{\mathcal{F}}$ are convex (Lemma 4). Note that the property is equivalent to showing that the decomposition $h_1 = 0$ and $h_2 = h$ is optimal. By [3, Lemma 1], this decomposition is optimal if and only if there exists a measure $\nu^* \in \mathscr{B}(\Omega)$ such that

$$J_P(0) = \langle \nu^*, 0\rangle - J_P^\star(\nu^*) \tag{1}$$

$$\varepsilon\Theta_{\mathcal{F}}(h) = \langle \nu^*, h\rangle - (\varepsilon\Theta_{\mathcal{F}})^\star(\nu^*) \tag{2}$$

First note that $J_P(h) = \sigma_{\mathscr{P}(\Omega)}(h) + \sigma_{\{-P\}}(h)$ and using properties of infimal convolutions, we have for any $\nu \in \mathscr{P}(\Omega)$

$$\begin{aligned} J_P^\star(\nu) &= \left(\sigma_{\mathscr{P}(\Omega)} + \sigma_{\{-P\}}\right)^\star(\nu) \\ &= \left(\sigma^\star_{\mathscr{P}(\Omega)} \bar{\star}\, \sigma^\star_{\{-P\}}\right)(\nu) \\ &= \left(\delta_{\mathscr{P}(\Omega)} \bar{\star}\, \delta_{\{-P\}}\right)(\nu) \\ &= \inf_{\nu'\in\mathscr{B}(\Omega)} \left(\delta_{\mathscr{P}(\Omega)}(\nu') + \delta_{\{-P\}}(\nu - \nu')\right) \\ &= \inf_{\nu'\in\mathscr{P}(\Omega)} \delta_{\{-P\}}(\nu - \nu') \\ &= \infty \cdot [\![P + \nu \notin \mathscr{P}(\Omega)]\!] \\ &= \infty \cdot [\![\nu \notin \mathscr{P}(\Omega) - P]\!]. \end{aligned}$$

Since $J_P(0) = \langle \nu, 0\rangle = 0$ for any $\nu \in \mathscr{B}(\Omega)$, this tells us that a ν^* satisfies the condition of Equation 1 if and only if ν^* is of the form $\mu - P$ where μ is any element of $\mathscr{P}(\Omega)$. We can re-arrange Equation 2 into

$$\langle \nu^*, h\rangle - \varepsilon\Theta_{\mathcal{F}}(h) = (\varepsilon\Theta_{\mathcal{F}})^\star(\nu^*),$$

and by definition since $(\varepsilon\Theta_{\mathcal{F}})^\star(\nu^*) = \sup_{\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})}\left(\left\langle \nu^*, \hat{h}\right\rangle - \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right)$, Equation 2 setting $\nu^* = \mu - P$ becomes

$$\begin{aligned}
\langle \nu^*, h\rangle - \varepsilon\Theta_{\mathcal{F}}(h) &= \sup_{\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})}\left(\left\langle \nu^*, \hat{h}\right\rangle - \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right) \qquad (3)\\
\iff \langle \mu - P, h\rangle - \varepsilon\Theta_{\mathcal{F}}(h) &= \sup_{\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})}\left(\left\langle \mu - P, \hat{h}\right\rangle - \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right)\\
\iff \mathbb{E}_\mu[h] - \mathbb{E}_P[h] - \varepsilon\Theta_{\mathcal{F}}(h) &= \sup_{\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})}\left(\mathbb{E}_\mu[\hat{h}] - \mathbb{E}_P[\hat{h}] - \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right)\\
\iff h \in \operatorname*{arg\,sup}_{\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})} &\left(\mathbb{E}_\mu[\hat{h}] - \mathbb{E}_P[\hat{h}] - \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right)\\
\iff h \in \operatorname*{arg\,inf}_{\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})} &\left(\mathbb{E}_P[\hat{h}] - \mathbb{E}_\mu[\hat{h}] + \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right). \qquad (4)
\end{aligned}$$

Hence the decomposition $h_1 = 0$ and $h_2 = h$ is optimal if and only if h satisfies Equation 4 for some $\mu \in \mathscr{P}(\Omega)$, which is precisely the statement of the Theorem. ■

1.6 Proof of Corollary 2

Corollary 2 *Let $P_+, P_- \in \mathscr{P}(\Omega)$ and suppose $\mathcal{F} \subseteq \mathscr{F}(\Omega,\mathbb{R})$ is even. If*

$$h^* \in \operatorname*{arg\,inf}_{\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})}\left(\mathbb{E}_{P_-}[\hat{h}] - \mathbb{E}_{P_+}[\hat{h}] + \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right),$$

then we have

$$\begin{aligned}
\inf_{Q\in B_{\varepsilon,\mathcal{F}}(P_+)}\int_\Omega h^* dQ &= \int_\Omega h^* dP_+ - \varepsilon\Theta_{\mathcal{F}}(h^*)\\
\sup_{Q\in B_{\varepsilon,\mathcal{F}}(P_-)}\int_\Omega h^* dQ &= \int_\Omega h^* dP_- + \varepsilon\Theta_{\mathcal{F}}(h^*)
\end{aligned}$$

Proof Applying Theorem 2 with $P = P_-$ and $\mu = P_+$ and using Theorem 1 yields the result on $B_{\varepsilon,\mathcal{F}}(P_-)$. Notice that $\mathcal{F}$ is even, which means that $\Theta_{\mathcal{F}}(h) = \Theta_{\mathcal{F}}(-h)$ and so we have

$$\begin{aligned}
h^* &\in \operatorname*{arg\,inf}_{\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})}\left(\mathbb{E}_{P_-}[\hat{h}] - \mathbb{E}_{P_+}[\hat{h}] + \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right)\\
\iff -h^* &\in \operatorname*{arg\,inf}_{-\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})}\left(-\mathbb{E}_{P_-}[\hat{h}] + \mathbb{E}_{P_+}[\hat{h}] + \varepsilon\Theta_{\mathcal{F}}(-\hat{h})\right)\\
\iff -h^* &\in \operatorname*{arg\,inf}_{-\hat{h}\in\mathscr{F}(\Omega,\mathbb{R})}\left(\mathbb{E}_{P_+}[\hat{h}] - \mathbb{E}_{P_-}[\hat{h}] + \varepsilon\Theta_{\mathcal{F}}(\hat{h})\right).
\end{aligned}$$

We can then apply Theorem 2 to $-h^*$ which means $\Lambda_{\varepsilon,\mathcal{F}}(-h^*) = \varepsilon\Theta_{\mathcal{F}}(-h^*) = \varepsilon\Theta_{\mathcal{F}}(h^*)$. Putting this together and applying Theorem 1 to $-h^*$ gives

$$\sup_{Q\in B_{\varepsilon,\mathcal{F}}(P_+)}\int_\Omega -h^* dQ = \int_\Omega -h^* dP_+ + \varepsilon\Theta_{\mathcal{F}}(h^*),$$

and multiplying both sides by -1 concludes the proof. ■

1.7 Proof of Theorem 3

Theorem 4 *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex lower semi-continuous function with $f(1) = 0$, $\mathcal{F} \subseteq \mathscr{F}(\Omega,\mathbb{R})$ and $\mathcal{H} \subseteq \mathscr{F}(\Omega, \operatorname{dom}(f^\star))$. For any model and data distributions $\mu, P \in \mathscr{P}(\Omega)$ respectively, we have for all $\varepsilon > 0$*

$$\sup_{Q\in B_{\varepsilon,\mathcal{F}}(P)} \mathrm{GAN}_{f,\mathcal{H}}(\mu; Q) \le \mathrm{GAN}_{f,\mathcal{H}}(\mu; P) + \varepsilon\sup_{h\in\mathcal{H}}\Theta_{\mathcal{F}}(h)$$

Proof We have

$$\begin{aligned}
\sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \mathrm{GAN}_{f,\mathcal{H}}(\mu; Q) &= \sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \sup_{h \in \mathcal{H}} \left(\int_\Omega h dQ - \int_\Omega f^\star(h) d\mu \right) \\
&\overset{(1)}{=} \sup_{h \in \mathcal{H}} \sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \left(\int_\Omega h dQ - \int_\Omega f^\star(h) d\mu \right) \\
&= \sup_{h \in \mathcal{H}} \left(\sup_{Q \in B_{\varepsilon,\mathcal{F}}(P)} \int_\Omega h dQ - \int_\Omega f^\star(h) d\mu \right) \\
&\overset{(2)}{=} \sup_{h \in \mathcal{H}} \left(\int_\Omega h dP + \Lambda_{\mathcal{F},\varepsilon}(h) - \int_\Omega f^\star(h) d\mu \right) \\
&\overset{(3)}{\leq} \sup_{h \in \mathcal{H}} \left(\int_\Omega h dP + \varepsilon \Theta_{\mathcal{F}}(h) - \int_\Omega f^\star(h) d\mu \right) \\
&\overset{(4)}{\leq} \sup_{h \in \mathcal{H}} \left(\int_\Omega h dP - \int_\Omega f^\star(h) d\mu \right) + \varepsilon \sup_{h \in \mathcal{H}} \Theta_{\mathcal{F}}(h) \\
&= \mathrm{GAN}_{f,\mathcal{H}}(\mu; P) + \varepsilon \sup_{h \in \mathcal{H}} \Theta_{\mathcal{F}}(h),
\end{aligned}$$

where (1) holds since we can exchange supremums, (2) is due to Theorem 1, (3) holds since $\Lambda_{\mathcal{F},\varepsilon} \leq \varepsilon \Theta_{\mathcal{F}}(h)$ and finally (4) holds since we can upper bound by taking out supremums. ■

Lemma 9 *For any $\mu \in \mathscr{P}(\Omega)$, $h \in \mathscr{F}(\Omega, \mathbb{R})$ we have*

$$\inf_{b\in\mathbb{R}} \sqrt{\mathbb{E}_{\mu(X)}[(h(X)-b)^2]} = \sqrt{\mathrm{Var}_\mu(h)}$$

Proof Let $\varphi(b) = \mathbb{E}_{\mu(X)}[(h(X)-b)^2]$ and $S(b) = \sqrt{\varphi(b)}$ and using simple calculus we have

$$S'(b) = \frac{\varphi'(b)}{2\sqrt{\varphi(b)}},$$

and noting that $\varphi(b) > 0$, we can find the minima by solving $\varphi'(b) = 0$ by first noting that

$$\varphi(b) = \mathbb{E}_{\mu(X)}[h^2(X)] - 2b\mathbb{E}_{\mu(X)}[h(X)] + b^2,$$

and so we have

$$\begin{aligned}\varphi'(b) = 0 &\iff -2 \cdot \mathbb{E}_{\mu(X)}[h(X)] + 2b = 0\\ &\iff b = \mathbb{E}_{\mu(X)}[h(X)].\end{aligned}$$

Putting this together yields

$$\begin{aligned}\inf_{b\in\mathbb{R}} \sqrt{\mathbb{E}_{\mu(X)}[(h(X)-b)^2]} &= \inf_{b\in\mathbb{R}} S(b)\\ &= S\left(\mathbb{E}_{\mu(X)}[h(X)]\right)\\ &= \mathbb{E}_{\mu(X)}\left[\left(h(X) - \mathbb{E}_{\mu(X)}[h(X)]\right)^2\right]\\ &= \sqrt{\mathrm{Var}_\mu(h)}\end{aligned}$$

■

References

[1] Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39(1):42, 1953.

[2] Shuang Liu and Kamalika Chaudhuri. The inductive bias of restricted f-gans. *arXiv preprint arXiv:1809.04542*, 2018.

[3] Japhet Niyobuhungiro. *Optimal decomposition for infimal convolution on Banach Couples*. Linköping University Electronic Press, 2013.

[4] Jean-Paul Penot. *Calculus without derivatives*, volume 266. Springer Science & Business Media, 2012.

[5] R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

[6] Ralph Rockafellar. Integrals which are convex functionals. *Pacific journal of mathematics*, 24(3):525–539, 1968.

[7] Thomas Strömberg. *A study of the operation of infimal convolution*. PhD thesis, Luleå tekniska universitet, 1994.

[8] Constantin Zalinescu. *Convex analysis in general vector spaces*. World scientific, 2002.

Chapter 6

Regularized Policies are Reward Robust

In this chapter, the narrative that regularization is a robustification strategy is extended to the particular setting of Reinforcement Learning (RL). Such an application is relevant since regularisation is commonly applied with much heuristic motivation and has also demonstrated empirical success. The most popular choice is causal-entropy, which is incorporated in popular frameworks such as Soft-Actor-Critic [Haarnoja et al., 2018]. We show that various forms of regularization correspond to being robust from the perspective of reward perturbations. Due to our result's generality, we apply our framework to various other forms of RL settings that include imitation learning from experts (and their regularized counterparts) and reward-free entropic maximization.

We encounter many new dualities that generalize existing results and serve independent technical interest to arrive at the result. We also find fascinating relationships between deep Q-learning, a technique that poses successful and is of great practical significance, to both regularization and robustness.

Regularized Policies are Reward Robust

Hisham Husain
The Australian National University
CSIRO Data61

Kamil Ciosek[*]
Spotify Research

Ryota Tomioka
Microsoft Research Cambridge

Abstract

Entropic regularization of policies in Reinforcement Learning (RL) is a commonly used heuristic to ensure that the learned policy explores the state-space sufficiently before overfitting to a local optimal policy. The primary motivation for using entropy is for exploration and disambiguating optimal policies; however, the theoretical effects are not entirely understood. In this work, we study the more general regularized RL objective and using Fenchel duality; we derive the dual problem which takes the form of an adversarial reward problem. In particular, we find that the optimal policy found by a regularized objective is precisely an optimal policy of a reinforcement learning problem under a worst-case adversarial reward. Our result allows us to reinterpret the popular entropic regularization scheme as a form of robustification. Furthermore, due to the generality of our results, we apply to other existing regularization schemes. Our results thus give insights into the effects of regularization of policies and deepen our understanding of exploration through robust rewards at large.

1 Introduction

Reinforcement Learning (RL) is a paradigm of algorithms which learn policies that maximize the expected discounted reward specified by a Markov Decision Process (MDP) (Sutton and Barto, 2018). The formulation of an MDP is well-posed with links in utility theory (Russell and Norvig, 2002) and specifies a reward function where the solution can be found precisely in a deterministic form. However, in practice, the reward function is typically an idealization, and it turns out that an optimal policy in this model will cope terribly when presented to unseen or uncertain situations. Intuitively, it is anticipated that there exist multiple policies that are near-optimal to this reward yet exhibit more robust and diversified behaviour. In particular, having multiple solutions of this form would be preferred since they can help the practitioner in understanding the environment and problem better.

Finding near-optimal policies in this sense requires balancing between ensuring that the policy is optimal for the given reward and demonstrates some form of robustness or diversity. This is commonly recollected as the *exploration* vs *exploitation* trade-off[1]. One of the most effective ways in ensuring this balance is by altering the objective of the MDP to include a form of penalty so that the resulting policy reflects characteristics of diversified behaviour. Causal entropy (Ziebart, 2010) is a popular example of this, where the policy is penalized for being deterministic in favour of exploration and disambiguating optimal policies. This has lead to the MaxEnt framework (Haarnoja et al., 2018c) and shown compelling relations to probabilistic inference (Dayan and Hinton, 1997; Neumann et al., 2011; Todorov, 2007; Kappen, 2005; Toussaint, 2009; Rawlik et al., 2013; Theodorou et al., 2010; Ziebart, 2010) whilst maintaining empirically superior performance on several tasks (Haarnoja et al., 2018c,b), including robustness in the face of uncertainty (Haarnoja et al., 2018a). In the case where the reward function is not specified, the entropy alone as an objective is also prevalent to ensure exploration (Hazan et al., 2019). Similar forms of regularization have appeared in Wu et al. (2019), which ensure that the policy is stabilized in accordance with a pre-determined behaviour and other forms of diversifying schemes using policy regularization have been developed in (Hong et al., 2018). Furthermore, the benefits of regularizers have also been observed in adversarial imitation learning

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

[*] Work done while at Microsoft Research Cambridge.

[1] Traditionally, this refers to the sequential behavior where one is interested in finding better policies at each timestep.

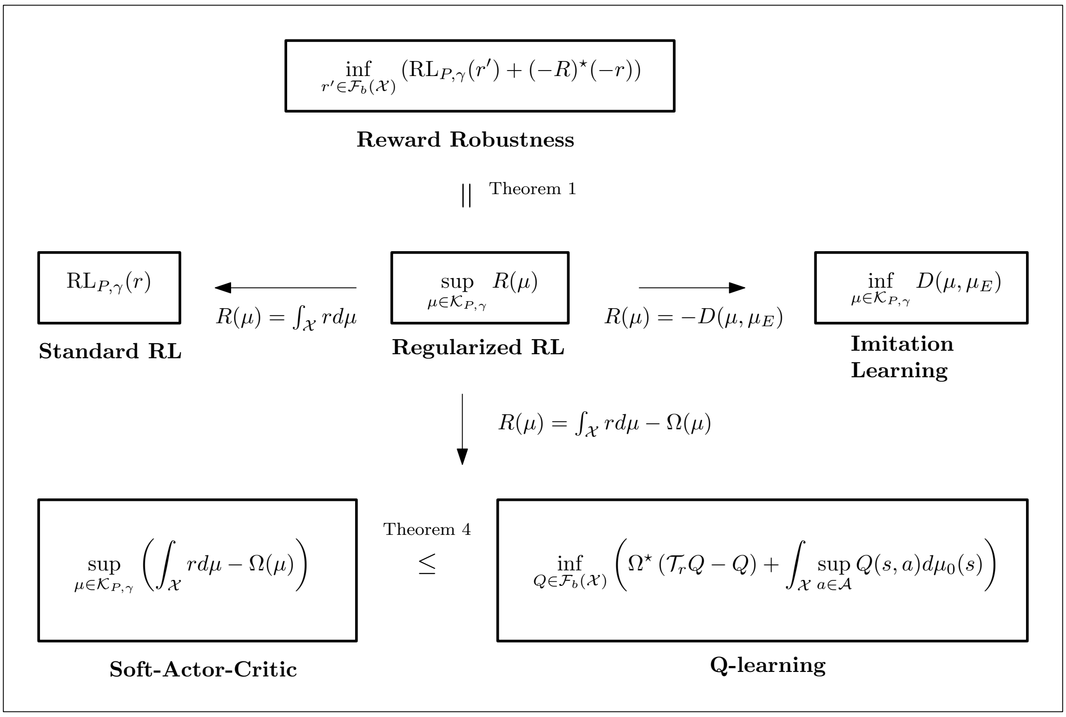

$\inf_{r' \in \mathcal{F}_b(\mathcal{X})} (\mathrm{RL}_{P,\gamma}(r') + (-R)^\star(-r))$
Reward Robustness
$||$ Theorem 1
$\mathrm{RL}_{P,\gamma}(r)$
Standard RL
$R(\mu) = \int_{\mathcal{X}} r d\mu$
$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu)$
Regularized RL
$R(\mu) = -D(\mu, \mu_E)$
$\inf_{\mu \in \mathcal{K}_{P,\gamma}} D(\mu, \mu_E)$
Imitation Learning
$R(\mu) = \int_{\mathcal{X}} r d\mu - \Omega(\mu)$
$\sup_{\mu \in \mathcal{K}_{P,\gamma}} \left(\int_{\mathcal{X}} r d\mu - \Omega(\mu) \right)$
Soft-Actor-Critic
Theorem 4
$\leq$
$\inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\Omega^\star \left(\mathcal{T}_r Q - Q \right) + \int_{\mathcal{X}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) \right)$
Q-learning

Figure 1: Our main is to provide a unified view of existing objectives in Reinforcement Learning and relate them to a reward robustness problem as highlighted above through Theorem 1. Additionally, we show another link between regularized policies and Q-learning in Theorem 4.

methods (Ho and Ermon, 2016; Li et al., 2017).

While the empirical success should rejoice, it is somewhat unsettling that changing the objective deviates from the MDP set-up, which was initially motivated through the axioms of utility theory (Russell and Norvig, 2002). In particular, it is not clear what kind of policy these regularized objectives are learning from the perspective of the original reward maximization problems, especially since it is apparent that regularized policies pose successfully in these schemes. On this front, there exists work that shows entropic regularization smoothens the optimization landscape (Ahmed et al., 2019) and induces sparse policies when considering a larger class of policy regularizers (Yang et al., 2019). While these works advocate the effects of policy regularization, the benefits of regularization from an accuracy or robustness perspective and not very well understood. This is especially relevant since in machine learning more generally, regularization has shown strong links to generalization and robustness (Duchi et al., 2016; Sinha et al., 2017; Husain, 2020). The first attempt is (Eysenbach and Levine, 2019), which shows that MaxEnt performs explicitly well on a robust reward problem. This approach however, is limited to only the MaxEnt and cannot apply to other schemes such as regularized imitation learning.

In this work, we tackle this precisely and focus on the problem specified by finding a policy that maximizes an objective R that is concave in the space of state-action visitation distributions. This objective includes the standard reward objective and subsumes other popular objectives such as the MaxEnt framework and imitation learning. Our main insight is that the policy learned using a concave objective R is *robust* against rewards chosen by an adversary, where R determines the nature of the adversary. We find that the policy is precisely a maximizer against the worst-case reward r'. Moreover, we characterize the analytic form of r' (using a technical assumption on R), which delivers more insight onto the nature of robustness. Our results thus allow us to reinterpret entropic regularization and exploration more generally as a robustifying mechanism and add to the advocation for using such methods in practice. In summary, our contributions are

1. A duality result linking generalized RL objectives as adversarial reward problems[2], which allows us to reinterpret the extant MaxEnt framework, among others, as a robustifying mechanism.

[2] We remark that this is not the same as conventional adversarial training, as found in supervised learning.

2. Characterization of the adversarial reward solved by these regularized policy objectives. In doing so, we derive a generalized value function interpretation of entropic regularization.

3. A primal-dual link between the regularized policy objective and Q-learning loss. This allows us to reinterpret the mean-squared error Q-learning as a form regularization of policies and robustification against rewards in light of our main result.

4. Deriving the robust-reward problem for other popular frameworks such as imitation learning and model-free entropic optimization. This allows us to compare and unify these separate problems under reward-robustness. We illustrate this diagrammatically in Figure 1.

2 Preliminaries

Reinforcement Learning We use a compact set $\mathcal{S}$ to denote the state space, $\mathcal{A}$ the action space and set $\mathcal{X} = \mathcal{S} \times \mathcal{A}$. We assume these spaces are Polish and furthermore use $\mathscr{P}(\mathcal{S})$, $\mathscr{P}(\mathcal{A})$ and $\mathscr{P}(\mathcal{X})$ to denote the set of Borel probability measures. Similarly, we use $\mathcal{F}_b(\mathcal{S})$, $\mathcal{F}_b(\mathcal{A})$ and $\mathcal{F}_b(\mathcal{X})$ to denote the set of bounded and measurable functions on the sets $\mathcal{S}, \mathcal{A}$ and $\mathcal{X}$ respectively. A reward function is a mapping $r : \mathcal{X} \to \mathbb{R}$, a transition kernel is specified as $P : \mathcal{X} \to \mathscr{P}(\mathcal{S})$ and a policy is a mapping $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$. Let $\gamma > 0$ be an implicit fixed discount parameter. It can be shown that each $\mathcal{S}$, $\mathcal{A}$, P, initial distribution μ_0 and policy π uniquely define a Markov chain $\{(S_t, A_t)\}_{t=1}^{\infty} \subseteq \mathcal{X}$. We denote the underlying probability space as $(\mathcal{X}, \mathscr{T}, P_{\mu_0,\pi})$ where $P_{\mu_0,\pi} \in \mathscr{P}(\mathcal{X})$ is referred to as the state-action visitation distribution. We refer the reader to (Meyn and Tweedie, 2012, Chapter 3) and (Revuz, 2008, Chapter 2) for more detailed constructions. The goal in RL is to find a policy that maximizes expected return over the state-action pairs visited, which can be concretely summarized in the optimization problem:

$$\sup_{\pi:\mathcal{S}\to\mathscr{P}(\mathcal{A})} \mathbb{E}_{P_{\mu_0,\pi}(s,a)}\left[r(s,a)\right]. \tag{1}$$

This objective is linear in the space of state-action visitation distributions and thus is equivalent to the linear program $\max_{\mu\in\mathcal{K}_{P,\gamma}} \int_{\mathcal{X}} r(s,a)d\mu(s,a)$ where

$$\mathcal{K}_{P,\gamma} = \left\{\mu \in \mathscr{P}(\mathcal{X}) : \int_{\mathcal{A}} \mu(s,a)da = (1-\gamma)\mu_0(s) + \gamma \int_{\mathcal{X}} P(s \mid s', a')d\mu(s', a')\right\}.$$

In particular, for any policy π, we have that $P_{\mu_0,\pi} \in \mathcal{K}_{P,\gamma}$ and that for any element $\mu \in \mathcal{K}_{P,\gamma}$, we can construct the corresponding policy $\pi_\mu(s) = \mu(s,a)/\int_{\mathcal{A}} \mu(s,a)da$. We will now introduce notation to formally write the reinforcement learning problem described in 1 since it will serve useful for the remainder of the paper.

Definition 1 *For a reward function $r : \mathcal{X} \to \mathbb{R}$, we define*

$$\mathrm{RL}_{P,\gamma}(r) := \sup_{\mu\in\mathcal{K}_{P,\gamma}} \int_{\mathcal{X}} r(s,a)d\mu(s,a)$$
$$M_{P,\gamma}(r) := \arg\sup_{\mu\in\mathcal{K}_{P,\gamma}} \int_{\mathcal{X}} r(s,a)d\mu(s,a)$$

In the above, $\mathrm{RL}_{P,\gamma}(r)$ is the same as (1) and represents the maximum expected reward possible under an environment P, discount factor γ and reward function r. The set $M_{P,\gamma}(r) \subseteq \mathscr{P}(\mathcal{X})$ represent the solutions that achieve maximal expected reward.

Convex Analysis and Legendre-Fenchel Duality We use $\mathscr{B}(\mathcal{X})$ to denote the set of finitely-additive measures and denote its topological dual to be $\mathcal{F}_b(\mathcal{X})$, the set of measurable and bounded functions mapping from $\mathcal{X}$ to $\mathbb{R}$. For any functional $F : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$, we define the Legendre-Fenchel dual, for any $h \in \mathcal{F}_b(\mathcal{X})$ as

$$F^\star(h) = \sup_{\mu\in\mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} h(x)d\mu(x) - F(\mu)\right).$$

For a set of functions $\mathcal{F} \subseteq \mathcal{F}_b(\mathcal{X})$, we use $\iota_{\mathcal{F}}(h)$ to denote the convex indicator function defined which is 0 if $h \in \mathcal{F}$ and $+\infty$ otherwise. For any two measures $\mu, \nu \in \mathscr{B}(\mathcal{X})$, we define the f-divergence between μ and ν to be $D_f(\mu,\nu) = \int_{\mathcal{X}} f(d\mu/d\nu)d\nu - \int_{\mathcal{X}} d\nu + 1$ where $f : \mathbb{R} \to (-\infty, \infty]$ is a lower semicontinuous convex function with $f(1) = 0$. In particular, the setting of $f(t) = t\log t$ is the popular Kullback-Leiber divergence, which we denote by $\mathrm{KL}(\mu,\nu) = D_f(\mu,\nu)$.

3 Related Work

Our main contribution is a reinterpretation of regularized policy maximization as robustifying mechanisms and so we discuss developments at understanding these methods along with similar results existing in machine learning at large. The idea of using causal entropy (Ziebart, 2010) is guided by the intuition of encouraging curious and diversified behavior. Further developed in (Haarnoja et al., 2018c), empirical success of using this penalty has been apparent. In particular, regularized policies unlike standard policies have illustrated robust behavior in the face of uncertainty and diversified behavior in finite sample schemes. Despite the empirical success, there is not much work studying these benefits from a formal perspective. The main existing results show that regularized objectives include smoothen the optimization landscape (Ahmed et al., 2019) and yield sparse policies (Yang et al., 2019). (Eysenbach and Levine, 2019) focuses on the MaxEnt framework and relates the optimal policy to solving a variable reward problem, which is line with our findings. Their results in contrast to ours, cannot be applied to other policy regularizers or other schemes that use causal entropy in the absence of reward functions such as adversarial imitation learning (Li et al., 2017).

In the realm of machine learning more generally, regularization has been principally established as a robustifying strategy. In supervised learning, various forms of robustness have shown connections to a number of regularization penalties such as Lipschitzness (Blanchet and Murthy, 2019; Sinha et al., 2017; Cranko et al., 2020; Husain, 2020), variance (Duchi et al., 2016) and Hilbert space norms (Staib and Jegelka, 2019). In Optimal Transport (OT), it has also been shown that entropic regularization is linked to ground cost robustness (Paty and Cuturi, 2020). Our result thus extends and develops these narratives for RL. (Zhang et al., 2020) also uses technical tools similar to our work such as Fenchel duality however for their purposes and findings are for quite different purposes.

4 Reward Robust Reinforcement Learning

We will be focusing on the problem specified by

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu),$$

where $R : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ is a concave upper semicontinuous function. Note that when a reward function $r : \mathcal{X} \to \mathbb{R}$ is given, setting $R(\mu) = \int_{\mathcal{X}} r(x) d\mu(x)$ recovers the standard maximum expected reward problem. Furthermore, the above subsumes other developments of RL in the case where the reward is unknown and R is chosen to be the entropy (Hazan et al., 2019) or imitation learning when $R(\mu) = -D(\mu, \mu_E)$ where μ_E is some expert demonstration and D is a divergence between probability measures (Ghasemipour et al., 2019). We present the main result which shows the above as a reward robust RL problem.

Theorem 1 *For any concave upper semicontinuous function $R : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$, we have*

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) = \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') + (-R)^{\star}(-r') \right)$$

Proof (Sketch) The key part of the proof is to rewrite R in terms of the convex conjugate of $-R$, which is well-defined since $-R$ is lower semicontinuous and convex, by assumptions on R. The proof then concludes by moving the supremum over μ inside by an application of a generalized minimax theorem. ■

The key point from the above is that the value of the maximal policy over R is exactly equal to the problem of finding an adversarial reward. In particular, the adversarial reward problem seeks to find a reward r' that makes the maximally achievable reward $\mathrm{RL}_{P,\gamma}$ as small as possible while paying the penalty $(-R)^{\star}(-r')$, where $(-R)^{\star}$ is a convex function. We remark that this is a one-party problem involving only an adversary. The conventional notion of robustness would relate this to the optimal model μ. We do this precisely by presenting a result that links the optimal μ and adversarial reward r':

Theorem 2 *Let μ^* and r^* be the optimal solution to the problems specified in Theorem 1, then we have that $\mu^* \in M_{P,\gamma}(r^*)$.*

This result tell us that an optimal policy found by solving the regularized objective is precisely an optimal policy of the Reinforcement Learning problem specified by the adversarial reward r^*. This is particularly striking since it tells us that though we are maximizing some concave R, which may be motivated for separate purposes, we can always guarantee that the policy learned is optimal for some reward r' in the axiomatic utility theory sense. In particular, this reward r^* is chosen to be the worst-case for this environment. The strength of robustness and nature of the adversarial reward clearly depends on the choice of R, as this is what budgets the adversarial reward r'. We will show that under a technical assumption on R, we can characterize the form r^* takes, which happens to depend on a single state-dependent mapping $V \in \mathcal{F}_b(\mathcal{S})$. The particular technical assumption on $(-R)^{\star}$ is that it is *increasing* by which we mean $r(x) \geq r'(x)$ for every $x \in \mathcal{X}$ implies $(-R)^{\star}(r) \geq (-R)^{\star}(r')$. We first introduce a result.

Theorem 3 *Suppose R is concave upper semicontinuous and let $\mathscr{I}$ be the value of the optimization problem*

$$\inf_{V \in \mathcal{F}_b(\mathcal{S}), r \in \mathcal{F}_b(\mathcal{X})} \left((1-\gamma) \int_{\mathcal{S}} V(s) d\mu_0(s) + (-R)^{\star}(-r) \right), \tag{2}$$

$$\text{s.t.}\, V(s) \geq r(s,a) + \gamma \int_{\mathcal{S}} V(s') dP(s' \mid s, a).$$

It then holds that $\mathscr{I} = \sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu)$.

It should be first noted that the above is a strong duality Theorem and indeed is a generalized version of the standard linear programming duality between policy maximization and value function minimization as described in (Agarwal et al., 2019), which is recovered when $R(\mu) = \int_{\mathcal{X}} r(x)d\mu(x)$ for some reward r. We will now show that the optimal value function of this objective gives the optimal reward. In particular, note that by solving the above constraint for the reward yields

$$r_V(s,a) := V(s) - \gamma \cdot \int_{\mathcal{S}} V(s')dP(s' \mid s,a). \tag{3}$$

We then have the following result

Lemma 1 *Suppose* $(-R)^\star$ *is increasing and* V^* *is the optimal solution of* (2) *then* r_{V^*} *is the optimal adversarial reward.*

The main consequence of the above Lemma is that it characterizes the shape of the adversarial reward chosen. In particular, it tells us that as long as as R satisfies the technical assumption ($(-R)^\star$ is increasing), the adversarial reward will be of the form r_V for some V. This is insightful since it tells us that the adversarial reward relates rewards between states through the dynamics of P. For example, note that if a particular state-action pair (s,a) yields the same state s then $r_V(s,a) = (1-\gamma)V(s)$. This technical condition on R can be satisfied for any R with a simple reparametrization, which we lay out in Lemma 1 in the supplementary material, and exploit when deriving $(-R)^\star$ for Soft-Actor-Critic. Moreover, we will show that the common choices of R which are motivated for smoothing or other empirical benefits naturally satisfy this technical assumption.

Generalized Soft-Actor-Critic Regularization Consider the case of having an available reward and using a convex penalty $\Omega : \mathscr{B}(\mathcal{X}) \times \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ for the policy so we select $R = R_\Omega$ of the form

$$R_\Omega(\mu) = \int_{\mathcal{X}} r(s,a)d\mu(s,a) - \varepsilon \cdot \Omega(\mu),$$

for some $\varepsilon > 0$. It can easily be shown (see Appendix) that $(-R)^\star(-r') = \varepsilon\Omega^\star\left(\frac{r-r'}{\varepsilon}\right)$, so that we have the following.

Corollary 1 *Let* $\Omega : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ *be a convex penalty then for any* $\varepsilon > 0$ *we have*

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R_\Omega(\mu) = \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') + \varepsilon\Omega^\star\left(\frac{r-r'}{\varepsilon}\right)\right).$$

The above tells us that the adversarial reward problem pays a price for deviating from the given reward r due to the second term $\varepsilon\Omega^\star\left(\frac{r-r'}{\varepsilon}\right)$. In the Soft-Actor-Critic (SAC) method, this corresponds to selecting (upto some constant) $\Omega_{\mathrm{SAC}}(\mu) = \mathbb{E}_{\mu(s,a)}\left[\mathrm{KL}(\pi_\mu(\cdot \mid s), U)\right]$, where π_μ is the policy induced by μ and U is the uniform distribution over $\mathcal{A}$. We presented Corollary 1 with a general Ω, which we believe will be useful for future developments. In this work, we consider the causal policy entropy along with 2-Tsallis entropy in the next next section. For the SAC case, we have the following result

Lemma 2 (Soft-Actor-Critic) *For any* $\varepsilon > 0$ *and* $r, r' \in \mathcal{F}(\mathcal{X})$*, we have*

$$\begin{aligned}&\varepsilon\Omega^\star_{\mathrm{SAC}}\left(\frac{r-r'}{\varepsilon}\right)\\ &= \varepsilon \cdot \sup_{s \in \mathcal{S}} \left(\int_{\mathcal{X}} \exp\left(\frac{r(s,a) - r'(s,a)}{\varepsilon}\right) dU(a) - 1\right)\end{aligned}$$

If one reasons about how the adversary behaves, the first incentive is to make $\mathrm{RL}_{P,\gamma}(r')$ small by selecting very small rewards across the environment. However, we can see that for the case of entropic regularization, the adversary pays a big price for selecting r' to be far from the original reward r for any given state. Note that in this case, we have $(-R)^\star$ is increasing and so in light of the concrete insight found in Lemma 1, we are able to reason about the SAC policy maximizing a reward of the worst-case reward of the form (3). This is striking since it tells us that the adversarial reward r' will respect the environment dynamics across the action space even if the ground reward r does not.

Derivation of Q-learning through robust learning In this subsection, we derive Q-learning through the reward-robust RL framework. In this context, learning a policy that is robust to a small variation in the reward corresponds to allowing a small violation of the Bellman equation with respect to the original reward function. For any Q-function $Q \in \mathcal{F}_b(\mathcal{X})$, we define the bellman operator $\mathcal{T}_r : \mathcal{F}_b(\mathcal{X}) \to \mathcal{F}_b(\mathcal{X})$ as

$$\mathcal{T}_r Q(s,a) = r(s,a) + \gamma \int_{\mathcal{X}} \sup_{a' \in \mathcal{A}} Q(s',a')dP(s' \mid s,a)$$

The maximum reward problem can be restated as

$$\mathrm{RL}_{P,\gamma}(r) = \inf_{Q \geq \mathcal{T}_r Q} \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a)d\mu_0(s), \tag{4}$$

where the optimal $Q^* \in \mathcal{F}_b(\mathcal{X})$ from the above is a contraction of $\mathcal{T}_r$ meaning that $\mathcal{T}_r Q^* = Q^*$. As it is difficult to find this contraction, one method known as *deep Q-learning* tackles this by parametrizing Q with a deep neural network and uses regression in the supervised learning sense to match $\mathcal{T}_r Q$ to Q (Sutton and Barto, 2018). This will deviate from the original objective since it relaxes this constraint $Q = \mathcal{T}_r\mathcal{Q}$ into the term appearing in the objective, which will naturally introduce bias. We now show quite a remarkable connection that doing so is related to policy regularization and by virtue of Corollary 1, linked to reward robustness.

Theorem 4 *For any $\varepsilon > 0$ and convex Ω such that $\Omega^\star$ is increasing, we have*

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R_\Omega(\mu) = \inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\varepsilon \Omega^\star \left(\frac{\mathcal{T}_r Q - Q}{\varepsilon} \right) + \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) \right).$$

We remark that the above is an inequality if $\Omega^\star$ is not increasing which results in *weak duality.* First note that the Theorem is precisely a relaxed *unconstrained* version of *constraint* objective appearing in (4). The most notable aspect of this result is that it links the regularized objective to finding a Q-function that minimizes the difference in the Bellman update $\varepsilon\Omega^\star\left(\frac{\mathcal{T}_rQ-Q}{\varepsilon}\right)$, depending on the choice of Ω. There exists work that show a relationship between gradients in entropy regularization and Q-learning (Schulman et al., 2017), however we state a more generalized result and bridge it to reward robustness. To see how this relates to the existing losses used in Q-learning, let us consider both the finite and continuous case. In the finite case, we can pick $\Omega(\mu) = \sum_{x \in \mathcal{X}} \mu(x)^2$, which is the 2-Tsallis entropy. One can easily derive the dual $\Omega^\star(r) = \frac{1}{4}\sum_{x\in\mathcal{X}} r(x)^2$ and thus the right side of Theorem 4 becomes (setting $\varepsilon = 1$)

$$\inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\frac{1}{4} \sum_{(s,a) \in \mathcal{X}} (\mathcal{T}_r Q(s,a) - Q(s,a))^2 + \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) \right).$$

The variational problem above is a regression problem between Q and $\mathcal{T}_rQ$ using the squared loss, which is the typical objective in deep Q-learning. The consequence of our result is that using this particular choice of loss to learn the Q function is related to learning a policy with the 2-Tsallis entropy, which is rather striking. Furthermore, the 2-Tsallis entropy behaves similar to the Shannon entropy in the sense that it is maximized when μ is uniform and minimized when μ is degenerate. In the continuous case, a buffer distribution $\nu \in \mathscr{P}(\mathcal{X})$ is used for the loss by defining the mean-squared error as L^2 norm with respect to ν between $\mathcal{T}_rQ$ and Q: given by $\|\mathcal{T}_rQ - Q\|^2_{L^2(\nu)}$. In this case, it can be shown that if $\Omega(\mu) = \frac{1}{4}\int_{\mathcal{X}}\left(\frac{d\mu}{d\nu}\right)^2 d\nu$ when $\mu \ll \nu$ and $+\infty$ otherwise then $\Omega^\star(h) = \|h\|^2_{L^2(\nu)}$.

Imitation Learning One method of learning a policy is to imitate expert data which comes in the form of a given distribution $\mu_E \in \mathscr{P}(\mathcal{X})$. Unlike the regularized schemes above, there is no specified reward function. Using the unified perspective provided in (Ghasemipour et al., 2019), where imitation learning is cast as divergence minimization, we can write these methods into our framework by selecting $R(\mu) = -D(\mu, \mu_E)$ (for each corresponding divergence). In particular, our goal is to not only derive the corresponding robust-reward problem but also show that $(-R)^\star$ will be increasing for these cases. We delegate the technical derivations to the Supplementary Section 1.8 and only present the results here. First, we focus on Adversarial Inverse Reinforcement Learning (AIRL) (Fu et al., 2017) selecting $R(\mu) = -\operatorname{KL}(\mu, \mu_E)$ in which case we have

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) = \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') + \int_{\mathcal{X}} \exp\left(-r'(x)\right) d\mu_E(x) - 1 \right),$$

noting that $(-R)^\star$ is increasing. We show the more general result that when $R(\mu) = -D_f(\mu, \mu_E)$ where D_f is an f-divergence then $(-R)^\star$ will be increasing. Using this choice of R corresponds to f-MAX (Ghasemipour et al., 2019). Another method for imitation learning is to use a discriminator based divergence as employed in InfoGAIL (Li et al., 2017). In this setting we assume we have a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and denoting the Lipschitz constant of a function $h \in \mathcal{F}_b(\mathcal{X})$ as $\operatorname{Lip}_d(h) := \sup_{x,x' \in \mathcal{X}} |h(x) - h(x')| / d(x,x')$, we set

$$R(\mu) = -\sup_{h : \operatorname{Lip}_d(h) \le L} \left(\int_{\mathcal{X}} h(x) d\mu(x) - \int_{\mathcal{X}} h(x) d\mu_E(x) \right),$$

where $L > 0$ is chosen as a hyperparameter. In this case, we have

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) = \inf_{r' : \operatorname{Lip}_d(r') \le L} \left(\mathrm{RL}_{P,\gamma}(r') - \int_{\mathcal{X}} r' d\mu_E \right).$$

It is clear from the above that the adversarial reward seeks to ensure $\mathrm{RL}_{P,\gamma}$ is as low as possible while maintaining that r' is large around the expert trajectory due to the second term. It should also be noted that the choice of L reflects as the budget of the adversary. We do not have $(-R)^\star$ increasing for this choice of R. On the other hand, it is typical in practice that an

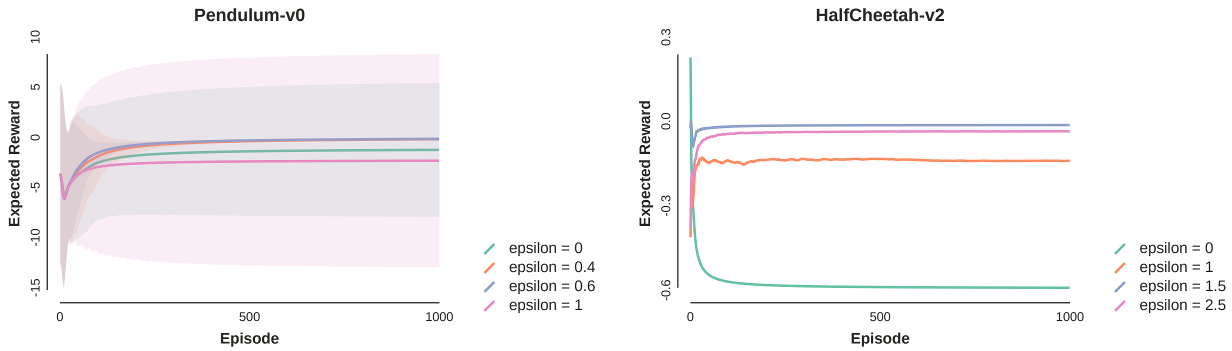


Figure 2: Expected reward over 1000 episodes of policies returned by SAC trained on an adversarial reward r_{adv} and tested on the true reward using different weighting ε for entropy.

entropy term is included in this term:

$$R(\mu) = -\sup_{h:\text{Lip}_d(h)\leq L}\left(\int_{\mathcal{X}} h(x)d\mu(x) - \int_{\mathcal{X}} h(x)d\mu_E(x)\right) - \varepsilon\mathbb{E}_{\mu(s,a)}\left[\text{KL}(\pi_\mu(\cdot \mid s), U_A)\right],$$

for some $\varepsilon > 0$ where U_A is the uniform distribution over $\mathcal{A}$. Under this setting, it turns out that $(-R)^\star$ is now increasing, in which case Lemma 1 applies. It is rather intriguing that the role of entropy here ensures that the reward that the InfoGAIL policy maximizes is worst-case, of high value around trajectories from the expert, and attains the familiar shape in Equation (3). This further advocates for the use of entropy regularization.

Entropic Exploration We now consider the case where there is no reward function or expert distribution specified and the only objective to maximize is entropy. For such a scheme, there exists efficient algorithms (Hazan et al., 2019). More specifically, we have $R(\mu) = -\text{KL}(\mu, U_{\mathcal{X}})$ where $U_{\mathcal{X}}$ is the uniform distribution over $\mathcal{X}$. We then have that

$$\sup_{\mu\in\mathcal{K}_{P,\gamma}} R(\mu) = \inf_{r'\in\mathcal{F}_b(\mathcal{X})}\left(\text{RL}_{P,\gamma}(r') + \int_{\mathcal{X}} \exp\left(-r'(x)\right) dU_{\mathcal{X}}(x) - 1\right),$$

and similar to the other choices of R, we have that $(-R)^\star$ is increasing. We would like to remark that if one defines KL to be $+\infty$ when μ is not a probability measure then $(-R)^\star(r) = \log\int_{\mathcal{X}}\exp(r(x))dU_{\mathcal{X}}(x)$ (Ruderman et al., 2012).

5 Experiments

The main practical ramification of our work is to advocate the use of regularized policies by highlighting the robustification aspect, for which we derived a strong theoretical link. There exists extensive empirical evidence for which our work provides foundation for. However, we will show some brief yet illustrative examples which focus on the reward adversarial aspect of regularized policies, as illustrated by our main result Theorem 1. Our goal is thus to see the performance of regularized policies on rewards they are not trained on and analyze their behavior based on the robustness parameter ε. First we consider the Pendulum-v0 environment and train the Soft-Actor-Critic (SAC) method on a reward that has been altered with. We do so by constructing an adversarial reward r_{adv} using

$$r_{\text{adv}} = \begin{cases} r(s,a) + \delta & \text{if } r(s,a) \leq -5 \\ r(s,a) & \text{otherwise} \end{cases}$$

where δ is drawn from a normal distribution centered at 5 with variance 0.1. In doing so, initial states of the pendulum will be favored and easier to reach however the maximal reward will still be attained at the inverted position. We train SAC for various values of ε and test their performance on the true reward in Figure 2 (left). We find that the effect of increasing ε yields better performance than no entropy however adding too much entropy (in the case of $\varepsilon = 1$) damages performance. We repeat a similar experiment for HalfCheetah-v2 however using an adversarial reward specified by

$$r_{\text{adv}} = \begin{cases} r(s,a) + \delta & \text{if } r(s,a) \leq 0 \\ r(s,a) & \text{otherwise} \end{cases}$$

where δ is drawn from a normal distribution centered at 3 with variance 0.1. We plot the performance under the expected reward in Figure 2 (right). It can also be seen that adding entropy surpasses the non-regularized policy $\varepsilon = 0$ and that increasing ε higher will worsen performance (as seen by $\varepsilon = 2.5$).

6 Conclusion

Our results allow us to reason about regularization of policies and the regression Q-learning objective from the perspective of robustness. This is not surprising given the advancements in machine learning more generally pointing at the link between regularization and robustness along with the impressive empirical evidence of these schemes. Regularized objectives, however, offer other benefits that are inherently sample based phenomenon such as smoothened objectives or stable training. While our results do not directly target this, we have built a connection between two objectives which will pose modular for future developments.

Acknowledgements

We would like to thank Zakaria Mhammedi for useful feedback regarding the technical analysis. HH is funded by the Australian Government Research Training Program and Data61.

References

Agarwal, A., Jiang, N., and Kakade, S. M. (2019). Reinforcement learning: Theory and algorithms. Technical report, Technical Report, CS Department, UW Seattle.

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. (2019). Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pages 151–160.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Cranko, Z., Shi, Z., Zhang, X., Nock, R., and Kornblith, S. (2020). Generalised lipschitz regularisation equals distributional robustness. *arXiv preprint arXiv:2002.04197.*

Dayan, P. and Hinton, G. E. (1997). Using expectation-maximization for reinforcement learning. *Neural Computation*, 9(2):271–278.

Duchi, J., Glynn, P., and Namkoong, H. (2016). Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425.*

Eysenbach, B. and Levine, S. (2019). If maxent rl is the answer, what is the question? *arXiv preprint arXiv:1910.01913.*

Fu, J., Luo, K., and Levine, S. (2017). Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248.*

Ghasemipour, S. K. S., Zemel, R., and Gu, S. (2019). A divergence minimization perspective on imitation learning methods. *arXiv preprint arXiv:1911.02256.*

Haarnoja, T., Ha, S., Zhou, A., Tan, J., Tucker, G., and Levine, S. (2018a). Learning to walk via deep reinforcement learning. *arXiv preprint arXiv:1812.11103.*

Haarnoja, T., Pong, V., Zhou, A., Dalal, M., Abbeel, P., and Levine, S. (2018b). Composable deep reinforcement learning for robotic manipulation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6244–6251. IEEE.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018c). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290.*

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019). Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573.

Hong, Z.-W., Shann, T.-Y., Su, S.-Y., Chang, Y.-H., Fu, T.-J., and Lee, C.-Y. (2018). Diversity-driven exploration strategy for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 10489–10500.

Husain, H. (2020). Distributional robustness with ipms and links to regularization and gans. *Advances in Neural Information Processing Systems*, 33.

Kappen, H. J. (2005). Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011.

Li, Y., Song, J., and Ermon, S. (2017). Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, pages 3812–3822.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability.* Springer Science & Business Media.

Neumann, G. et al. (2011). Variational inference for policy search in changing situations. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 817–824.

Paty, F.-P. and Cuturi, M. (2020). Regularized optimal transport is ground cost adversarial. *arXiv preprint arXiv:2002.03967.*

Rawlik, K., Toussaint, M., and Vijayakumar, S. (2013). On stochastic optimal control and reinforcement learning by approximate inference. In *Twenty-third international joint conference on artificial intelligence*.

Revuz, D. (2008). *Markov chains*. Elsevier.

Ruderman, A., Reid, M., García-García, D., and Petterson, J. (2012). Tighter variational representations of f-divergences via restriction to probability measures. *arXiv preprint arXiv:1206.4664*.

Russell, S. and Norvig, P. (2002). Artificial intelligence: a modern approach.

Schulman, J., Chen, X., and Abbeel, P. (2017). Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*.

Sinha, A., Namkoong, H., and Duchi, J. (2017). Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2.

Staib, M. and Jegelka, S. (2019). Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, pages 9131–9141.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Theodorou, E., Buchli, J., and Schaal, S. (2010). A generalized path integral control approach to reinforcement learning. *The Journal of Machine Learning Research*, 11:3137–3181.

Todorov, E. (2007). Linearly-solvable markov decision problems. In *Advances in neural information processing systems*, pages 1369–1376.

Toussaint, M. (2009). Robot trajectory optimization using approximate inference. In *Proceedings of the 26th annual international conference on machine learning*, pages 1049–1056.

Wu, Y., Tucker, G., and Nachum, O. (2019). Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

Yang, W., Li, X., and Zhang, Z. (2019). A regularized approach to sparse optimal policy in reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5940–5950.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151*.

Ziebart, B. D. (2010). Modeling purposeful adaptive behavior with the principle of maximum causal entropy.

Supplementary Material for “Regularized Policies are Reward Robust”

1 Proofs of Main Results

We first introduce some notation that will be used exclusively for the Appendix. For any function $R : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$, we define $R_+(\mu) = R(\mu) + \iota_{\mathscr{P}}(\mu)$ and $R_-(\mu) = R(\mu) - \iota_{\mathscr{P}}(\mu)$. Indeed, it should noted that if R is upper semi-continuous concave then R_- is upper semi-continuous concave and $-R_-$ is proper convex. The central benefit of rewriting R in this is way is due to

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) = \sup_{\mu \in \mathcal{K}_{P,\gamma}} R_-(\mu).$$

First we will show a technical result.

Lemma 1 *If $R : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ is upper semicontinuous and concave then $(-R_-)^\star$ is increasing.*

Proof Let $r, r' \in \mathcal{F}_b(\mathcal{X})$ such that $r \leq r'$ and let

$$\nu \in \operatorname*{arg\,sup}_{\mu \in \mathscr{P}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) d\mu(x) + R(\mu) \right),$$

noting that ν exists since the mapping $\mu \mapsto \int_{\mathcal{X}} r(x) d\mu(x) + R(\mu)$ is concave, upper semicontinuous and $\mathscr{P}(\mathcal{X})$ is compact. Next we have

$$\begin{aligned}
&(-R_-)^\star(r) - (-R_-)^\star(r') \\
&= \sup_{\mu \in \mathscr{P}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) d\mu(x) + R(\mu) \right) - \sup_{\mu \in \mathscr{P}(\mathcal{X})} \left(\int_{\mathcal{X}} r'(x) d\mu(x) + R(\mu) \right) \\
&\leq \int_{\mathcal{X}} r(x) d\nu(x) + R(\nu) - \int_{\mathcal{X}} r'(x) d\nu(x) - R(\nu) \\
&= \int_{\mathcal{X}} \left(r(x) - r'(x) \right) d\nu(x) \\
&\leq 0
\end{aligned}$$

■

We also recall some classical results regarding Fenchel duality between the spaces $\mathcal{F}_b(\mathcal{X})$ and $\mathscr{B}(\mathcal{X})$.

Definition 1 (Rockafellar (1968)) *For any proper convex function $F : \mathcal{F}_b(\mathcal{X}) \to (-\infty, \infty]$ and $\mu \in \mathscr{B}(\mathcal{X})$ we define*

$$F^\star(\mu) = \sup_{h \in \mathcal{F}_b} \left(\int_{\mathcal{X}} h d\mu - F(h) \right)$$

and for any $h \in \mathcal{F}_b(\Omega)$ we define

$$F^{\star\star}(h) = \sup_{\mu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} h d\mu - F^\star(\mu) \right).$$

Theorem 1 (Zalinescu (2002) Theorem 2.3.3) *If X is a Hausdorff locally convex space, and $F : X \to (-\infty, \infty]$ is a proper convex lower semi-continuous function then $F^{\star\star} = F$.*

1.1 Proof of Theorem 1

We have

$$\begin{aligned}
\sup_{\mu\in\mathcal{K}_{P,\gamma}} R(\mu) &= \sup_{\mu\in\mathcal{K}_{P,\gamma}} -(-R(\mu)) \\
&\overset{(1)}{=} \sup_{\mu\in\mathcal{K}_{P,\gamma}} -(-R(\mu))^{\star\star} \\
&\overset{(2)}{=} \sup_{\mu\in\mathcal{K}_{P,\gamma}} - \sup_{r'\in\mathcal{F}_b(\mathcal{X})} \left(\int_{\mathcal{X}} r'(x)d\mu(x) - (-R)^{\star}(r')\right) \\
&= \sup_{\mu\in\mathcal{K}_{P,\gamma}} \inf_{r'\in\mathcal{F}_b(\mathcal{X})} \left(\int_{\mathcal{X}} (-r'(x))\, d\mu(x) + (-R)^{\star}(r')\right) \\
&\overset{(3)}{=} \inf_{r'\in\mathcal{F}_b(\mathcal{X})} \sup_{\mu\in\mathcal{K}_{P,\gamma}} \left(\int_{\mathcal{X}} (-r'(x))\, d\mu(x) + (-R)^{\star}(r')\right) \\
&\overset{(4)}{=} \inf_{r'\in\mathcal{F}_b(\mathcal{X})} \left(\sup_{\mu\in\mathcal{K}_{P,\gamma}} \int_{\mathcal{X}} r'(x)d\mu(x) + (-R)^{\star}(-r')\right) \\
&\overset{(5)}{=} \inf_{r'\in\mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') + (-R)^{\star}(-r')\right)
\end{aligned}$$

where (1) holds since $-R$ is proper convex, (2) is the definition of the conjugate, (3) is an application of Ky Fan's minimax theorem (Fan, 1953, Theorem 2) noting that the set $\mathcal{K}_{P,\gamma}$ is compact, and that the mapping $r \mapsto \int_{\mathcal{X}} (-r'(x))\, d\mu(x) + (-R)^{\star}(r')$ is concave and the mapping $\mu \mapsto \int_{\mathcal{X}} (-r'(x))\, d\mu(x)$ is linear. (4) holds by negating r' since $-\mathcal{F}_b(\mathcal{X}) = \mathcal{F}_b(\mathcal{X})$ and (5) holds by definition.

1.2 Proof of Theorem 2

By definition, we have $\mathrm{RL}_{P,\gamma}(r^*) - \langle r^*, \mu^*\rangle \geq 0$. To show the other direction, it follows that

$$\begin{aligned}
\mathrm{RL}_{P,\gamma}(r^*) - \langle r^*, \mu^*\rangle &= (\mathrm{RL}_{P,\gamma}(r^*) + (-R)^{\star}(-r^*)) - (\langle r^*, \mu^*\rangle + (-R)^{\star}(-r^*)) \\
&\overset{(1)}{=} \inf_{r'\in\mathcal{F}_b(\mathcal{X})} (\mathrm{RL}_{P,\gamma}(r') + (-R)^{\star}(-r')) - (\langle r^*, \mu^*\rangle + (-R)^{\star}(-r^*)) \\
&\overset{(2)}{=} \sup_{\mu\in\mathcal{K}_{P,\gamma}} R(\mu) - (\langle r^*, \mu^*\rangle + (-R)^{\star}(-r^*)) \\
&\overset{(3)}{=} R(\mu^*) - (\langle r^*, \mu^*\rangle + (-R)^{\star}(-r^*)) \\
&= \langle -r^*, \mu^*\rangle - (-R)(\mu^*) - (-R)^{\star}(-r^*) \\
&\overset{(4)}{\leq} 0,
\end{aligned}$$

where (1) follows via optimality of r^*, (2) is due to the duality result, (3) follows via optimality of μ^* and (4) is an application of the Fenchel-Young inequality on the convex function $-R$. Finally, we have $\mathrm{RL}_{P,\gamma}(r^*) = \langle r^*, \mu^*\rangle$, which implies optimality of μ^* and concludes the proof.

1.3 Proof of Theorem 3

Using the classic linear programming duality result, we have

$$\mathrm{RL}_{P,\gamma}(r) = (1-\gamma) \inf_{V\in\mathcal{V}_{P,r,\gamma}} \int_{\mathcal{S}} V(s)d\mu_0(s), \tag{1}$$

where

$$\mathcal{V}_{P,r,\gamma} = \left\{ V \in \mathcal{F}_b(\mathcal{S}) : V(s) \geq r(s,a) + \gamma \int_{\mathcal{S}} V(s')dP(s' \mid s,a), \forall (s,a) \in \mathcal{X} \right\},$$

and define

$$r_V(s,a) := V(s) - \gamma \int_{\mathcal{S}} V(s')dP(s' \mid s,a). \tag{2}$$

It then holds that

$$\begin{aligned}
\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) &\overset{(1)}{=} \inf_{r' \in \mathcal{F}_b(\mathcal{X})} (\mathrm{RL}_{P,\gamma}(r') + (-R)^\star(-r')) \\
&\overset{(2)}{=} \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left((1-\gamma) \inf_{V \in \mathcal{V}_{P,r',\gamma}} \int_{\mathcal{S}} V(s)d\mu_0(s) + (-R)^\star(-r') \right) \\
&= \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \inf_{V \in \mathcal{F}_b(\mathcal{S})} \left((1-\gamma) \int_{\mathcal{S}} V(s)d\mu_0(s) + (-R)^\star(-r') + \iota_{\mathcal{V}_{P,r',\gamma}}(V) \right) \\
&= \inf_{V \in \mathcal{F}_b(\mathcal{S})} \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left((1-\gamma) \int_{\mathcal{S}} V(s)d\mu_0(s) + (-R)^\star(-r') + \iota_{\mathcal{V}_{P,r',\gamma}}(V) \right) \\
&= \inf_{V \in \mathcal{F}_b(\mathcal{S})} \inf_{r' \leq r_V} \left((1-\gamma) \int_{\mathcal{S}} V(s)d\mu_0(s) + (-R)^\star(-r') \right),
\end{aligned}$$

where (1) is due to Theorem 1, (2) is due to (1) and noting that $r \leq r_V$ implies $V(s) \geq r(s,a) + \gamma \int_{\mathcal{S}} V(s')dP(s' \mid s,a)$ concludes the proof.

1.4 Proof of Lemma 1

First note that for any $\mu \in \mathcal{K}_{P,\gamma}$, we have

$$\begin{aligned}
&\int_{\mathcal{X}} r_V(s,a)d\mu(s,a) \\
&= \left(\int_{\mathcal{S}} V(s)d\mu(s,a) - \gamma \int_{\mathcal{X}} \int_{\mathcal{S}} V(s')dP(s' \mid s,a)d\mu(s,a) \right) \\
&= \left(\int_{\mathcal{S}} V(s)d\mu(s,a) - \int_{\mathcal{S}} V(s)d\mu(s,a) + (1-\gamma) \int_{\mathcal{S}} V(s)d\mu_0(s) \right) \\
&= (1-\gamma) \int_{\mathcal{S}} V(s)d\mu_0(s),
\end{aligned}$$

and so we can conclude for any $V \in \mathcal{F}_b(\mathcal{S})$, we have

$$\mathrm{RL}_{P,\gamma}(r_V) = (1-\gamma) \int_{\mathcal{S}} V(s)d\mu_0(s).$$

Next, we have

$$\begin{aligned}
\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) &= \inf_{V \in \mathcal{F}_b(\mathcal{S})} \left((1-\gamma) \int_{\mathcal{S}} V(s)d\mu_0(s) + (-R)^\star(-r_V) \right) \\
&= \inf_{V \in \mathcal{F}_b(\mathcal{S})} (\mathrm{RL}_{P,\gamma}(r_V) + (-R)^\star(-r_V)) \\
&\geq \inf_{r' \in \mathcal{F}_b(\mathcal{X})} (\mathrm{RL}_{P,\gamma}(r') + (-R)^\star\,(-r')) \\
&= \sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu),
\end{aligned}$$

and since the lower bound can achieve equality, it implies that the optimal r^* is of the form r_V.

1.5 Proof of Corollary 1

We have

$$\begin{aligned}
(-R)^\star(-r') &= \sup_{\mu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} -r'(x) d\mu(x) + R(\mu) \right) \\
&= \sup_{\mu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} -r'(x) d\mu(x) + \int_{\mathcal{X}} r(x) d\mu(x) - \varepsilon\Omega(\mu) \right) \\
&= \sup_{\mu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) - r'(x) d\mu(x) - \varepsilon\Omega(\mu) \right) \\
&= \varepsilon \sup_{\mu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} \frac{r(x) - r'(x)}{\varepsilon} d\mu(x) - \Omega(\mu) \right) \\
&= \varepsilon\Omega^\star \left(\frac{r - r'}{\varepsilon} \right),
\end{aligned}$$

which concludes the proof.

1.6 Proof of Theorem 4

First define the set

$$\mathcal{Q}_{P,r,\gamma} = \left\{ Q \in \mathcal{F}_b(\mathcal{X}) : Q(s,a) \geq r(s,a) + \gamma \int_{\mathcal{X}} \sup_{a' \in \mathcal{A}} Q(s',a') dP(s' \mid s,a) \right\},$$

and define

$$r_Q(s,a) = Q(s,a) - \gamma \int_{\mathcal{X}} \sup_{a' \in \mathcal{A}} Q(s',a') dP(s' \mid s,a)$$

Next we can write

$$\mathrm{RL}_{P,\gamma}(r) = \inf_{Q \in \mathcal{Q}_{P,r,\gamma}} \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s), \tag{A}$$

next we have

$$\begin{aligned}
\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) &\overset{(1)}{=} \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') + (-R)^\star(-r') \right) \\
&\overset{(2)}{=} \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\inf_{Q \in \mathcal{Q}_{P,r',\gamma}} \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) + (-R)^\star(-r') \right) \\
&= \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) + \iota_{\mathcal{Q}_{P,r',\gamma}}(Q) \right) + (-R)^\star(-r') \right) \\
&= \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) + (-R)^\star(-r') + \iota_{\mathcal{Q}_{P,r',\gamma}}(Q) \right) \\
&= \inf_{Q \in \mathcal{F}_b(\mathcal{X})} \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) + (-R)^\star(-r') + \iota_{\mathcal{Q}_{P,r',\gamma}}(Q) \right) \\
&= \inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) + \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left((-R)^\star(-r') + \iota_{\mathcal{Q}_{P,r',\gamma}}(Q) \right) \right) \\
&= \inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) + \inf_{r' \leq r_Q} (-R)^\star(-r') \right) \\
&\overset{(3)}{=} \inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) + (-R)^\star(-r_Q) \right),
\end{aligned}$$

where (1) is due to Theorem 1, (2) is due to (A), and (3) follows since $(-R)^\star$ is increasing by assumption. Next, noting that $(-R)^\star(-r_Q) = \varepsilon\Omega^\star\left(\frac{r-r_Q}{\varepsilon}\right)$, and that

$$\begin{aligned} r - r_Q &= r(s,a) - \frac{Q(s,a)}{1-\gamma} + \gamma \int_{\mathcal{X}} \sup_{a' \in \mathcal{A}} Q(s',a') dP(s' \mid s,a) \\ &= \left(r(s,a) + \gamma \int_{\mathcal{X}} \sup_{a' \in \mathcal{A}} Q(s',a') dP(s' \mid s,a) \right) - Q(s,a) \\ &= \mathcal{T}Q - Q, \end{aligned}$$

which is the difference between the Bellman operator. Putting this together yields

$$\begin{aligned} &\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) \\ &= \inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\varepsilon\Omega^\star\left(\frac{r-r_Q}{\varepsilon}\right) + \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) \right) \\ &= \inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\varepsilon\Omega^\star\left(\frac{\mathcal{T}Q - Q}{\varepsilon}\right) + \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) \right) \end{aligned}$$

1.7 Proof of Lemma 2

We first set $n = |A|$. Let $\mathcal{F}_b(\mathcal{S}, \mathbb{R}^n)$ denote the set of measurable and bounded functions mapping from $\mathcal{S}$ into $\mathbb{R}^n$. For any $\pi \in \mathcal{F}_b(\mathcal{S}, \mathbb{R}^n)$, we use $\pi(a \mid s)$ to denote the index corresponding to $a \in \mathcal{A}$ for the function π evaluated at $s \in \mathcal{S}$. Next, we define the following set:

$$\mathcal{B}_\times := \left\{ \mu(s,a) = \pi(a \mid s) \cdot \mu_S(s) \mid \mu_S \in \mathscr{P}(\mathcal{S}), \pi \in \mathcal{F}_b(\mathcal{S}, \mathbb{R}^n) \right\},$$

noting that $\mathcal{B}_\times \subseteq \mathscr{B}(\mathcal{X})$. We also have that $\mathscr{P}(\mathcal{X}) \subset \mathcal{B}_\times$ since this corresponds to having each $\pi(a \mid s)$ satisfy $\pi(a \mid s) \in [0,1]$ and $\sum_{a \in \mathcal{A}} \pi(a \mid s) = 1$. We then redefine

$$\Omega(\mu) = \begin{cases} \mathbb{E}_{\mu(s,a)}\left[\mathrm{KL}(\pi_\mu(\cdot \mid s), U)\right] & \text{if } \mu \in \mathcal{B}_\times \\ \infty & \text{if } \mu \notin \mathcal{B}_\times \end{cases}$$

We will first show that this choice of Ω is convex. First we need a Lemma that will make it easier.

Lemma 2 *The functional $F : \mathbb{R}^n \to \mathbb{R}$ defined as*

$$F(\mathbf{x}) = \sum_{i=1}^n x_i \cdot \log\left(\frac{x_i}{\sum_{j=1} x_j}\right)$$

is convex over its domain $\mathbb{R}^n_{>0}$.

Proof We derive the Hessian of F which can be verified to be:

$$HF(\mathbf{x}) = \operatorname{diag}\left(\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}\right) - \frac{1}{\sum_{i=1}^n x_i} \cdot \mathbf{1}^\intercal \mathbf{1}.$$

Next, we have for any vector $z \in \mathbb{R}^n$ and $x \in \operatorname{dom} F$:

$$\begin{aligned} z^\intercal HF(x) z &= z^\intercal \operatorname{diag}\left(\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}\right) z - \frac{1}{\sum_{i=1}^n x_i}\left(\sum_{i=1}^n z_i\right)^2 \\ &= \sum_{i=1}^n \frac{z_i^2}{x_i} - \frac{1}{\sum_{i=1}^n x_i}\left(\sum_{i=1}^n z_i\right)^2 \\ &= \frac{1}{\sum_{i=1}^n x_i}\left(\left(\sum_{i=1}^n x_i\right) \cdot \left(\sum_{i=1}^n \frac{z_i^2}{x_i}\right) - \left(\sum_{i=1}^n z_i\right)^2\right) \\ &\geq 0, \end{aligned}$$

where the last inequality follows by an application of Cauchy-Schwarz inequality noting that $x \in \operatorname{Dom} F = \mathbb{R}^n_{>0}$. Since the Hessian is positive semi-definite, it follows that F is convex. ■

First denote by $\mu_S(s) = \sum_{a\in\mathcal{A}} \mu(s,a)$ and note that $\pi_\mu(a \mid s) = \mu(s,a)/\mu_S(s)$. For any $\mu \in \operatorname{dom}\Omega$, we have

$$\begin{aligned}\Omega(\mu) &= \mathbb{E}_{\mu(s,a)}\left[\mathrm{KL}(\pi_\mu, U)\right] \\ &= \mathbb{E}_{\mu(s,a)}\left[\sum_{a\in\mathcal{A}} \pi_\mu(a \mid s)\cdot\log\left(\pi_\mu(a\mid s)\right) + \log n\right] \\ &= \mathbb{E}_{\mu_S(s)}\left[\sum_{a\in\mathcal{A}} \pi_\mu(a \mid s)\cdot\log\left(\pi_\mu(a\mid s)\right)\right] + \log n \\ &= \int_{\mathcal{S}}\sum_{a\in\mathcal{A}} \mu_S(s)\pi_\mu(a\mid s)\cdot\log\left(\pi_\mu(a\mid s)\right) ds + \log n \\ &= \int_{\mathcal{S}}\sum_{a\in\mathcal{A}} \mu(s,a)\cdot\log\left(\frac{\mu(s,a)}{\sum_{a'\in\mathcal{A}}\mu(s,a')}\right) ds + \log n,\end{aligned}$$

and convexity follows by the above Lemma. Before we proceed, we need to also show that $\mathcal{B}_\times$ is convex so that our redefining of Ω does not break convexity established above. Consider $\mu,\nu \in \mathcal{B}_\times$ and so there exists $\mu_S, \nu_S \in \mathscr{P}(\mathcal{S})$ and $\pi_\mu, \pi_\nu \in \mathcal{F}_b(\mathcal{S}, \mathbb{R}^n)$ with $\mu(s,a) = \pi_\mu(a\mid s)\cdot\mu_S(s)$ and $\nu(s,a) = \pi_\nu(a\mid s)\cdot\nu_S(s)$. For any $\lambda \in [0,1]$, we have (setting $P_{\mu,\nu}(s) = \frac{\mu_S(s)+\nu_S(s)}{2}$)

$$\begin{aligned}\lambda\cdot\mu(s,a) + (1-\lambda)\nu(s,a) &= \lambda\pi_\mu(a\mid s)\cdot\mu_S(s) + (1-\lambda)\cdot\pi_\nu(a\mid s)\cdot\nu_S(s) \\ &= P_{\mu,\nu}(s)\cdot\left(\lambda\pi_\mu(a\mid s)\cdot\frac{\mu_S(s)}{P_{\mu,\nu}(s)} + (1-\lambda)\cdot\pi_\nu(a\mid s)\cdot\frac{\nu_S(s)}{P_{\mu,\nu}(s)}\right).\end{aligned}$$

By construction, both μ_S and ν_S are absolutely continuous with respect to $P_{\mu,\nu}$ and thus the terms inside the bracket are bounded and well-defined. Moreover $P_{\mu,\nu} \in \mathscr{P}(\mathcal{S})$ and thus this element is in $\mathcal{B}_\times$, which concludes the convexity proof. We now proceed to derive the conjugate. For any $r' \in \mathcal{F}_b(\mathcal{X})$ we have

$$\begin{aligned}\Omega^\star(r') &= \sup_{\mu\in\mathscr{B}(\mathcal{X})}\left(\int_{\mathcal{X}} r'(s,a)d\mu(s,a) - \Omega(\mu)\right) \\ &\overset{(1)}{=} \sup_{\mu\in\mathcal{B}_\times}\left(\int_{\mathcal{X}} r'(s,a)d\mu(s,a) - \Omega(\mu)\right) \\ &= \sup_{\mu\in\mathcal{B}_\times}\left(\int_{\mathcal{X}} r'(s,a)d\mu(s,a) - \mathbb{E}_{\mu(s,a)}\left[\mathrm{KL}(\pi_\mu(\cdot\mid s), U)\right]\right) \\ &= \sup_{\mu\in\mathcal{B}_\times}\left(\int_{\mathcal{X}}\left(\int_{\mathcal{A}} r'(s,a)d\pi_\mu(a\mid s) - \mathrm{KL}(\pi_\mu(\cdot\mid s),U)\right)d\mu(s,a)\right) \\ &= \sup_{\mu_S\in\mathscr{P}(\mathcal{S})}\ \sup_{\pi_\mu(\cdot\mid s)\in\mathcal{F}_b(\mathcal{S},\mathbb{R}^n)}\left(\int_{\mathcal{X}}\left(\int_{\mathcal{A}} r'(s,a)d\pi_\mu(a\mid s) - \mathrm{KL}(\pi_\mu(\cdot\mid s),U)\right)d\mu_S(s)\right) \\ &\overset{(2)}{=} \sup_{\mu_S\in\mathscr{P}(\mathcal{S})}\int_{\mathcal{X}}\sup_{\pi_\mu\in\mathbb{R}^n}\left(\int_{\mathcal{A}} r'(s,a)d\pi_\mu(a) - \mathrm{KL}(\pi_\mu,U)\right)d\mu_S(s) \\ &\overset{(3)}{=} \sup_{\mu_S\in\mathscr{P}(\mathcal{S})}\int_{\mathcal{X}}\sup_{\pi_\mu\in\mathscr{P}(\mathcal{A})}\left(\int_{\mathcal{A}} r'(s,a)d\pi_\mu(a) - \mathrm{KL}(\pi_\mu,U)\right)d\mu_S(s) \\ &\overset{(4)}{=} \sup_{\mu_S\in\mathscr{P}(\mathcal{S})}\int_{\mathcal{X}}\exp\left(r'(s,a)\right)dU(a) - 1 \\ &\overset{(5)}{=} \sup_{s\in\mathcal{S}}\int_{\mathcal{X}}\exp\left(r'(s,a)\right)dU(a) - 1,\end{aligned}$$

where (1) holds since $\operatorname{dom}\Omega \subseteq \mathcal{B}_\times$. (2) holds from (Rockafellar and Wets, 2009, Theorem 14.60, p. 677) using the fact that $\mathcal{F}_b(\mathcal{S},\mathbb{R}^n)$ is trivially a decomposable space in definition (Rockafellar and Wets, 2009, Definition 14.59, p. 676). (3) holds since $\operatorname{dom}(\mathrm{KL}(\cdot,U)) \subseteq \mathscr{P}(\mathcal{A}) \subset \mathbb{R}^n$. (4) is due to (Feydy et al., 2019, Proposition 5) and (5) follows by noting that the optimal μ_S is concentrated around the supremum.

1.8 Imitation Learning

1.8.1 f-divergence

Note that for any $r \in \mathcal{F}_b(\mathcal{X})$ we have

$$\begin{aligned}(-R)^\star(r) &= \sup_{\nu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) d\nu(x) + R(\nu) \right) \\ &= \sup_{\nu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) d\nu(x) - \mathrm{KL}(\nu, \mu_E \right) \\ &\overset{(1)}{=} \int_{\mathcal{X}} r(x) d\mu_E(x) - 1,\end{aligned}$$

where (1) holds due to (Feydy et al., 2019, Proposition 5). We will now show that $(-R)^\star$ is increasing for any $R(\mu) = -D_f(\mu, \mu_E)$ where D_f is an f-divergence. First let

$$\nu \in \operatorname*{arg\,sup}_{\mu \in \mathscr{P}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) d\mu(x) + R(\mu) \right),$$

noting that ν exists since the mapping $\mu \mapsto \int_{\mathcal{X}} r(x) d\mu(x) + R(\mu)$ is concave, upper semicontinuous and $\mathscr{P}(\mathcal{X})$ is compact. For any $r' \geq r$

$$\begin{aligned}&(-R_-)^\star(r) - (-R_-)^\star(r') \\ &= \sup_{\mu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) d\mu(x) + R(\mu) \right) - \sup_{\mu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} r'(x) d\mu(x) + R(\mu) \right) \\ &\overset{(1)}{=} \sup_{\mu \in \mathscr{P}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) d\mu(x) + R(\mu) \right) - \sup_{\mu \in \mathscr{P}(\mathcal{X})} \left(\int_{\mathcal{X}} r'(x) d\mu(x) + R(\mu) \right) \\ &\leq \int_{\mathcal{X}} r(x) d\nu(x) + R(\nu) - \int_{\mathcal{X}} r'(x) d\nu(x) - R(\nu) \\ &= \int_{\mathcal{X}} (r(x) - r'(x)) \, d\nu(x) \\ &\leq 0,\end{aligned}$$

where (1) holds due to the fact that $\operatorname{dom}(D_f(\cdot, \mu_E)) \subseteq \mathscr{P}(\mathcal{X})$.

1.8.2 InfoGAIL

In this case, we exploit the fact that $-R(\mu)$ takes the form of an Integral Probability Metric between μ and μ_E. Let $\mathcal{H}_L$ the set of functions that are L-Lipschitz with respect to d. For any $r \in \mathcal{F}_b(\mathcal{X})$ we have

$$\begin{aligned}(-R)^\star(r) &= \sup_{\nu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) d\nu(x) - \sup_{h : \mathrm{Lip}_d(h) \leq L} \left(\int_{\mathcal{X}} h(x) d\nu(x) - \int_{\mathcal{X}} h(x) d\mu_E(x) \right) \right) \\ &\overset{(1)}{=} \int_{\mathcal{X}} r(x) d\mu_E(x) + \iota_{\mathcal{H}_L}(r),\end{aligned}$$

where (1) is due to (Husain, 2020, Lemma 5). Thus, it holds that

$$\begin{aligned}\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) &= \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') + \int_{\mathcal{X}} -r'(x) d\mu_E(x) + \iota_{\mathcal{H}_L}(-r') \right) \\ &\overset{(2)}{=} \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') - \int_{\mathcal{X}} r'(x) d\mu_E(x) + \iota_{\mathcal{H}_L}(r') \right) \\ &= \inf_{r' : \mathrm{Lip}_d \leq L} \left(\mathrm{RL}_{P,\gamma}(r') - \int_{\mathcal{X}} r'(x) d\mu_E(x) \right),\end{aligned}$$

where (2) holds since $\text{Lip}_d(-r) = \text{Lip}_d(r)$. We now show that adding an entropy term to

$$R(\mu) = - \sup_{h:\text{Lip}_d(h)\leq L} \left(\int_{\mathcal{X}} h(x) d\mu(x) - \int_{\mathcal{X}} h(x) d\mu_E(x) \right) - \varepsilon \mathbb{E}_{\mu(s,a)} \left[\text{KL}(\pi_\mu(\cdot \mid s), U_A) \right] \tag{3}$$

will ensure that $(-R)^\star$ is increasing. Using standard results from (Penot, 2012) that the conjugate of the sum of two functions is the infimal convolution between their conjugates mean we will convolve both (3) and entropy conjugate from Lemma 2 of the main file.:

$$\begin{aligned}(-R)^\star(r') &= \inf_{r\in\mathcal{F}_b(\mathcal{X})} \left(\sup_{s\in\mathcal{S}} \int_{\mathcal{X}} \exp\left(r'(s,a) - r(s,a)\right) dU(a) + \int_{\mathcal{X}} r d\mu_E + \iota_{\mathcal{H}_L}(r) \right) &(4)\\ &= \inf_{r\in\mathcal{H}_L} \left(\sup_{s\in\mathcal{S}} \int_{\mathcal{X}} \exp\left(r'(s,a) - r(s,a)\right) dU(a) + \int_{\mathcal{X}} r d\mu_E \right). &(5)\end{aligned}$$

Let $r'' \leq r'$ pointwise and define

$$r^* \in \arg\inf_{r\in\mathcal{H}_L} \left(\sup_{s\in\mathcal{S}} \int_{\mathcal{X}} \exp\left(r'(s,a) - r(s,a)\right) dU(a) + \int_{\mathcal{X}} r d\mu_E \right), \tag{6}$$

noting that since exists due to Weierstrass Theorem since $\mathcal{H}_L$ is compact and the mapping inside is convex and lower semicontinuous. Next, we have

$$\begin{aligned}&(-R)^\star(r'') - (-R)^\star(r') &(7)\\ &= \inf_{r\in\mathcal{H}_L} \left(\sup_{s\in\mathcal{S}} \int_{\mathcal{X}} \exp\left(r''(s,a) - r(s,a)\right) dU(a) + \int_{\mathcal{X}} r d\mu_E \right) - \inf_{r\in\mathcal{H}_L} \left(\sup_{s\in\mathcal{S}} \int_{\mathcal{X}} \exp\left(r'(s,a) - r(s,a)\right) dU(a) + \int_{\mathcal{X}} r d\mu_E \right) &(8)\\ &\leq \sup_{s\in\mathcal{S}} \int_{\mathcal{X}} \exp\left(r''(s,a) - r^*(s,a)\right) dU(a) + \int_{\mathcal{X}} r^* d\mu_E - \sup_{s\in\mathcal{S}} \int_{\mathcal{X}} \exp\left(r'(s,a) - r^*(s,a)\right) dU(a) - \int_{\mathcal{X}} r^* d\mu_E &(9)\\ &= \sup_{s\in\mathcal{S}} \int_{\mathcal{X}} \exp\left(r''(s,a) - r^*(s,a)\right) dU(a) - \sup_{s\in\mathcal{S}} \int_{\mathcal{X}} \exp\left(r'(s,a) - r^*(s,a)\right) dU(a) &(10)\\ &\leq 0, &(11)\end{aligned}$$

where the last inequality follows from the fact that $r'' \leq r'$ and thus this proves that $(-R)^\star$ is increasing.

1.9 Entropic Exploration

For any $r \in \mathcal{F}_b(\mathcal{X})$

$$\begin{aligned}(-R)^\star(r) &= \sup_{\mu\in\mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} r(x) d\mu(x) - \text{KL}(\mu, U_{\mathcal{X}}) \right)\\ &\overset{(1)}{=} \int_{\mathcal{X}} \exp\left(r(x)\right) dU_{\mathcal{X}}(x) - 1,\end{aligned}$$

where (1) follows from (Feydy et al., 2019, Proposition 5).

References

Fan, K. (1953). Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39(1):42.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690.

Husain, H. (2020). Distributional robustness with ipms and links to regularization and gans. *Advances in Neural Information Processing Systems*, 33.

Penot, J.-P. (2012). *Calculus without derivatives*, volume 266. Springer Science & Business Media.

Rockafellar, R. (1968). Integrals which are convex functionals. *Pacific journal of mathematics*, 24(3):525–539.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Zalinescu, C. (2002). *Convex analysis in general vector spaces*. World scientific.

Chapter 7

Conclusion

The study of regularization at an abstract level has proven to be useful in delivering results that bridge practice and theory and offer new pursuits for future directions. We summarize the main insights into two sections: generative models and model robustness.

Regarding the former, Chapter 3 reconciled two primary methods to address (1), which allowed us to explain phenomena such as the behaviour of WAE and provided foundations for Lipschitz regularized discriminators, which have demonstrated success in existing empirical work. More generally, Chapter 5 showed that regularized discriminators in other forms also yield benefits from the perspective of distributional robustness and compliments existing narratives in this direction. We illustrated a theoretical benefit of the equivalence between GANs and Autoencoders in the form of a generalization bound, which deepens our understanding of implicit generative models. Chapter 4 made an orthogonal contribution to generative models by presenting a privacy-compliant model based on discriminators in GANs. This model is proven to avoid major pitfalls of GANs such as mode-collapse and is guaranteed to converge while consistently outperforming state-of-the-art methods.

From the robustness perspective, our work built a reinterpretation of regularization as robustification for several learning settings. A connection of this type naturally addresses both (1) and (2) since it is both a reinterpretation and a beneficial guarantee on empirical performance. In particular, we first studied how regularization in the form of penalized objectives manifest as distributionally robust objectives as upper bounds to a great deal of generality in Chapter 5. The results provided insights on other forms of existing penalties beyond Lipschitz, variance and kernel norms, such as the manifold regularization penalty and generalized variance penalties such as those appearing in the fairness-based approaches. It is then shown how the result is tightened to equality, which has links to regularized binary classification objectives commonly appearing in GANs. The two learning settings these results are actively applied to include supervised learning (such as classification and regression) and GANs. Chapter 6 extended the bridge between regularization and robustness into the Reinforcement Learning (RL) setting. The main result addressed the robustification properties of entropic-regularized policies. In particular, the contribution comments on other RL policy maximization schemes such as a wide range of regularizers and imitation learning, amongst new RL duality results that generalize their standard counterparts. Moreover, we discovered an interesting connection between the deep Q-learning objective and regularization in policies, hinting at the benefits of such particularly successful schemes in practice.

Throughout each chapter, we derived and employed technical results that serve as additional contributions. In particular, the proof techniques are versatile and can be applied to different learning settings involving regularization, as appropriately discussed, beyond generative models and robustness to target the motivation of (1) and (2).

7.1 Future Work

We discuss some future work along three different directions based on the contributions mentioned above.

7.1.1 Taxonomy of Generative Models

The results of Chapter 3 established a relationship based on Fenchel-duality results, which we emphasize can be applied to different choices of regularizers and GAN objectives at large. An example of this includes translation based GANs such as CycleGAN [Zhu et al., 2017] and MAGAN [Amodio and Krishnaswamy, 2018]. These objectives operate with multiple different regularizers and often involve training double the number of discriminators and generators. By showing duality results, one can reduce and better understand the purpose of each appearing term. Other lines of future work in this direction include considering costs c that are not metrics and using the Wasserstein distance's general dual formulation to derive the GAN model. A similar result exists in [Farnia and Tse, 2018] where the c-transform will manifest; however, the results established in Chapter 3 can be used to relate this to autoencoder models.

7.1.2 Robust Adversarial Training

The results in Chapter 5 established a link between distributional level robustness and regularization. However, in the context of ML, robustness is commonly addressed from the lens of adversarial training - a study into the tolerance of classifiers against adversarial noise perturbations. While there are links between Wasserstein distributional robustness and adversarial training, our results' relevance with IPMs beyond the Wasserstein distance is unclear. There are two concrete directions where these results can be of use for the adversarial training community:

- Recently, the final distribution learned by a GAN has been utilized as training data for binary classifiers [Wang and Yu, 2019; Charlier et al., 2019; Zhao et al., 2017, 2019; Lee et al., 2017; Jalal et al., 2017; Poursaeed et al., 2018; Song et al., 2017, 2018; Hayes and Danezis, 2018; Xiao et al., 2018; Samangouei et al., 2018], which have observed to be empirically robust in the adversarial sense. There is little theory supporting these claims, and therefore, given the robustifying benefits established in Chapter 5 of regularized GANs, the work establishes the foundations to understand the robustness of these learned classifiers.

- An important topic used to measure classifiers' adversarial robustness is through a quantity referred to as a *robustness certificate*. This certificate outputs a number under which a practitioner can assert the degree of robustness for a given classifier. Recently, it has been shown that one can form a certificate with a distributionally robust framework for f-divergences [Dvijotham et al., 2020]. The results from Chapter 5 outline similar schemes for IPMs and, therefore, can be used to develop new certificates similar in contribution to that of [Dvijotham et al., 2020].

7.1.3 Duality for Practice

The main contributions of Chapters 3 and 5 can be viewed as duality theorems which allow us to reinterpret objectives as other existing objectives (such as GANs to Autoencoders) or completely new ones (such as reward-adversarial games). This thesis's main benefits have been conceptual, providing explicit theoretical gains and providing insights for existing practice. However, new algorithms can also be inspired through dualities. For example, the RL setting shows that policy learning is equivalent to a reward-adversarial problem, and therefore one can learn the optimal reward in this dual space and consequently recover the optimal policy by solving an MDP on this reward. The practitioner can also use the optimal adversarial reward in understanding the environment and problem set-up at large. Another example involves using the dual form proven between GANs and Autoencoders, where one can use the optimal discriminator to construct an optimal encoder. Such a result will practically realize the link between GANs and Autoencoders beyond the theory established in this thesis.

Bibliography

Ali, S. M. and Silvey, S. D., 1966. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28, 1 (1966), 131–142. (cited on page 12)

Amodio, M. and Krishnaswamy, S., 2018. Magan: Aligning biological manifolds. In *International Conference on Machine Learning*, 215–223. PMLR. (cited on pages 3 and 102)

Arbel, M.; Sutherland, D.; Bińkowski, M.; and Gretton, A., 2018. On gradient regularizers for mmd gans. In *Advances in Neural Information Processing Systems*, 6700–6710. (cited on page 6)

Bartlett, P. L. and Mendelson, S., 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3, Nov (2002), 463–482. (cited on page 5)

Belkin, M.; Niyogi, P.; and Sindhwani, V., 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7, Nov (2006), 2399–2434. (cited on page 6)

Blanchet, J.; Kang, Y.; and Murthy, K., 2019. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56, 3 (2019), 830–857. (cited on page 6)

Blanchet, J. and Murthy, K., 2019. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44, 2 (2019), 565–600. (cited on page 6)

Calvetti, D. and Reichel, L., 2003. Tikhonov regularization of large linear problems. *BIT Numerical Mathematics*, 43, 2 (2003), 263–283. (cited on page 1)

Charlier, J.; Singh, A.; Ormazabal, G.; State, R.; and Schulzrinne, H., 2019. Syngan: Towards generating synthetic network attacks using gans. *arXiv preprint arXiv:1908.09899*, (2019). (cited on page 102)

Cranko, Z.; Kornblith, S.; Shi, Z.; and Nock, R., 2018. Lipschitz networks and distributional robustness. *arXiv preprint arXiv:1809.01129*, (2018). (cited on page 6)

Csiszár, I., 1964. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8 (1964), 85–108. (cited on page 12)

Csiszár, I., 1967. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2 (1967), 229–318. (cited on page 12)

Diggle, P. J. and Gratton, R. J., 1984. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46, 2 (1984), 193–212. (cited on page 3)

Duchi, J.; Glynn, P.; and Namkoong, H., 2016. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, (2016). (cited on page 6)

Duchi, J. C.; Jordan, M. I.; and Wainwright, M. J., 2013. Local privacy and statistical minimax rates. In *FOCS*. (cited on page 6)

Dunford, N. and Schwartz, J. T., 1988. *Linear operators, part 1: general theory*, vol. 10. John Wiley & Sons. (cited on page 9)

Dvijotham, K.; Hayes, J.; Balle, B.; Kolter, Z.; Qin, C.; Gyorgy, A.; Xiao, K.; Gowal, S.; and Kohli, P., 2020. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*. (cited on page 103)

Farnia, F. and Tse, D., 2018. A convex duality framework for gans. In *Advances in Neural Information Processing Systems*, 5248–5258. (cited on pages 4 and 102)

Fedus, W.; Rosca, M.; Lakshminarayanan, B.; Dai, A. M.; Mohamed, S.; and Goodfellow, I., 2017. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446*, (2017). (cited on page 2)

Fichtenholz, G. and Kantorovitch, L., 1934. Sur les opérations linéaires dans l'espace des fonctions bornées. *Studia Mathematica*, 5, 1 (1934), 69–98. (cited on page 9)

Goodfellow, I., 2016. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, (2016). (cited on page 5)

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y., 2014a. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680. (cited on page 2)

Goodfellow, I. J.; Shlens, J.; and Szegedy, C., 2014b. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, (2014). (cited on page 5)

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A., 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13, 1 (2012), 723–773. (cited on page 13)

Guedj, B., 2019. A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*, (2019). (cited on page 5)

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, (2018). (cited on page 81)

Hayes, J. and Danezis, G., 2018. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, 43–49. IEEE. (cited on page 102)

Hazan, E.; Kakade, S.; Singh, K.; and Van Soest, A., 2019. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2681–2691. (cited on page 7)

Hildebrandt, T. H., 1934. On bounded linear functional operations. *Transactions of the American Mathematical Society*, 36, 4 (1934), 868–875. http://www.jstor.org/stable/1989829. (cited on page 9)

Ho, J. and Ermon, S., 2016. Generative adversarial imitation learning. In *Advances in neural information processing systems*, 4565–4573. (cited on page 7)

Jalal, A.; Ilyas, A.; Daskalakis, C.; and Dimakis, A. G., 2017. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, (2017). (cited on page 102)

Kingma, D. P. and Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, (2013). (cited on page 4)

Lee, H.; Han, S.; and Lee, J., 2017. Generative adversarial trainer: Defense to adversarial perturbations with gan. *arXiv preprint arXiv:1705.03387*, (2017). (cited on page 102)

Li, C.-L.; Chang, W.-C.; Cheng, Y.; Yang, Y.; and Póczos, B., 2017. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2203–2213. (cited on page 6)

Li, K. and Malik, J., 2018. On the implicit assumptions of gans. *arXiv preprint arXiv:1811.12402*, (2018). (cited on page 4)

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, (2017). (cited on page 5)

Masood, M. A. and Doshi-Velez, F., 2019. Diversity-inducing policy gradient: Using maximum mean discrepancy to find a set of diverse policies. *arXiv preprint arXiv:1906.00088*, (2019). (cited on page 1)

McAllester, D. A., 1999. Some pac-bayesian theorems. *Machine Learning*, 37, 3 (1999), 355–363. (cited on page 5)

Mohamed, S. and Lakshminarayanan, B., 2016. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, (2016). (cited on page 3)

Mroueh, Y.; Li, C.-L.; Sercu, T.; Raj, A.; and Cheng, Y., 2017. Sobolev gan. *arXiv preprint arXiv:1711.04894*, (2017). (cited on page 6)

Mroueh, Y. and Sercu, T., 2017. Fisher gan. In *Advances in Neural Information Processing Systems*, 2513–2523. (cited on page 6)

Müller, A., 1997. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29, 2 (1997), 429–443. (cited on page 13)

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I., 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56, 11 (2010), 5847–5861. (cited on page 12)

Nowozin, S.; Cseke, B.; and Tomioka, R., 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, 271–279. (cited on page 12)

Penot, J.-P., 2012. *Calculus without derivatives*, vol. 266. Springer Science & Business Media. (cited on pages 9 and 11)

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S., 2018. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4422–4431. (cited on page 102)

Reid, M.; Williamson, R.; et al., 2011. Information, divergence and risk for binary experiments. (2011). (cited on page 12)

Rockafellar, R., 1968. Integrals which are convex functionals. *Pacific journal of mathematics*, 24, 3 (1968), 525–539. (cited on page 10)

Rockafellar, R. T. and Wets, R. J.-B., 2009. *Variational analysis*, vol. 317. Springer Science & Business Media. (cited on page 13)

Ruderman, A.; Reid, M.; García-García, D.; and Petterson, J., 2012. Tighter variational representations of f-divergences via restriction to probability measures. *arXiv preprint arXiv:1206.4664*, (2012). (cited on page 12)

Russell, S. and Norvig, P., 2002. Artificial intelligence: a modern approach. (2002). (cited on page 2)

Samangouei, P.; Kabkab, M.; and Chellappa, R., 2018. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, (2018). (cited on page 102)

Scarf, H. E., 1957. A min-max solution of an inventory problem. Technical report, RAND CORP SANTA MONICA CALIF. (cited on page 5)

Shafieezadeh-Abadeh, S.; Kuhn, D.; and Esfahani, P. M., 2019. Regularization via mass transportation. *Journal of Machine Learning Research*, 20, 103 (2019), 1–68. (cited on page 6)

Song, Y.; Kim, T.; Nowozin, S.; Ermon, S.; and Kushman, N., 2017. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, (2017). (cited on page 102)

Song, Y.; Shu, R.; Kushman, N.; and Ermon, S., 2018. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, 8312–8323. (cited on page 102)

Sriperumbudur, B. K.; Fukumizu, K.; Gretton, A.; Schölkopf, B.; and Lanckriet, G. R., 2009. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, (2009). (cited on page 13)

Staib, M. and Jegelka, S., 2019. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, 9131–9141. (cited on page 6)

Strömberg, T., 1994. *A study of the operation of infimal convolution*. Ph.D. thesis, Luleå tekniska universitet. (cited on page 11)

Sugiyama, M.; Suzuki, T.; and Kanamori, T., 2012. *Density ratio estimation in machine learning*. Cambridge University Press. (cited on page 12)

Sutton, R. S. and Barto, A. G., 2018. *Reinforcement learning: An introduction*. MIT press. (cited on page 1)

Tolstikhin, I.; Bousquet, O.; Gelly, S.; and Schoelkopf, B., 2017. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, (2017). (cited on page 2)

Vapnik, V. and Chervonenkis, A., 1974. Theory of pattern recognition. (cited on page 5)

Vapnik, V. N., 1999. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10, 5 (1999), 988–999. (cited on page 5)

Villani, C., 2008. *Optimal transport: old and new*, vol. 338. Springer Science & Business Media. (cited on page 14)

Wang, H. and Yu, C.-N., 2019. A direct approach to robust deep learning using adversarial networks. *arXiv preprint arXiv:1905.09591*, (2019). (cited on page 102)

Watkins, C. J. and Dayan, P., 1992. Q-learning. *Machine learning*, 8, 3-4 (1992), 279–292. (cited on page 7)

Xiao, C.; Li, B.; Zhu, J.-Y.; He, W.; Liu, M.; and Song, D., 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, (2018). (cited on page 102)

Yoon, J.; Jordon, J.; and Schaar, M., 2018. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, 5689–5698. PMLR. (cited on page 3)

Zalinescu, C., 2002. *Convex analysis in general vector spaces*. World scientific. (cited on page 10)

Zhang, P.; Liu, Q.; Zhou, D.; Xu, T.; and He, X., 2017. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, (2017). (cited on page 4)

Zhao, H.; Le, T.; Montague, P.; De Vel, O.; Abraham, T.; and Phung, D., 2019. Perturbations are not enough: Generating adversarial examples with spatial distortions. *arXiv preprint arXiv:1910.01329*, (2019). (cited on page 102)

Zhao, Z.; Dua, D.; and Singh, S., 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, (2017). (cited on page 102)

Zhou, Z.; Liang, J.; Song, Y.; Yu, L.; Wang, H.; Zhang, W.; Yu, Y.; and Zhang, Z., 2019. Lipschitz generative adversarial nets. *arXiv preprint arXiv:1902.05687*, (2019). (cited on page 4)

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232. (cited on pages 3 and 102)

Zolotarev, V. M., 1984. Probability metrics. *Theory of Probability & Its Applications*, 28, 2 (1984), 278–302. (cited on page 13)