Proceedings of Machine Learning Research vol 132:1–20, 2022

Adversarial Interpretation of Bayesian Inference

Hisham Husain*
Amazon, Adelaide

Jeremias Knoblauch
The Department of Statistical Science, University College London, London

Abstract

We build on the optimization-centric view on Bayesian inference advocated by Knoblauch et al. (2019). Thinking about Bayesian and generalized Bayesian posteriors as the solutions to a regularized minimization problem allows us to answer an intriguing question: If minimization is the primal problem, then what is its dual? By deriving the Fenchel dual of the problem, we demonstrate that this dual corresponds to an adversarial game: In the dual space, the prior becomes the cost function for an adversary that seeks to perturb the likelihood [loss] function targeted by standard [generalized] Bayesian inference. This implies that Bayes-like procedures are adversarially robust—providing another firm theoretical foundation for their empirical performance. Our contributions are foundational, and apply to a wide-ranging set of Machine Learning methods. This includes standard Bayesian inference, generalized Bayesian and Gibbs posteriors (Bissiri et al., 2016), as well as a diverse set of other methods including Generalized Variational Inference (Knoblauch et al., 2019) and the Wasserstein Autoencoder (Tolstikhin et al., 2017).

Keywords: Bayesian Inference, Fenchel Duality, f-divergences, Integral Probability Metric

1. Introduction

Bayesian and generalized Bayesian methods are enjoying ever-growing popularity within Machine Learning. There are numerous theoretical and conceptual reasons for this, and the one quoted most often is that the incorporation of prior beliefs aids inference. More specifically, finding a probability distribution over 'good' models—rather than a single point in the model space—makes generalized Bayesian procedures typically more robust than methods based on point estimation. While Information and Decision Theory can explain this performance boost whenever the parameter of interest θ indexes a likelihood p_θ and the prior π encodes actual domain expertise (see e.g. Williams, 1980; Zellner, 1988; Bernardo, 2000; O'Hagan and Oakley, 2004; Goldstein, 2006; Berger, 2013; Buck and Meson, 2015), the advantages of Bayes-like procedures extend far beyond this idealized and narrow setting (see e.g. Fong and Holmes, 2019).

For example, Gibbs posteriors (also often called quasi-posteriors) do not even require likelihood functions and instead produce distributions over any parameter θ indexing arbitrary loss functions (see e.g. Grünwald, 2011, 2012; Hooker and Vidyashankar, 2014; Ghosh and Basu, 2016; Bissiri et al., 2016; Holmes and Walker, 2017; Jewson et al., 2018; Nakagawa and Hashimoto, 2019; Lyddon et al., 2019; Chérief-Abdellatif and Alquier, 2019; Knoblauch and Vomfell, 2020). Posteriors of this kind have an intimate relationship with PAC-Bayesian theory (Shawe-Taylor and Williamson, 1997; McAllester, 1999a,b; Catoni, 2007), but this theory can only explain the performance for a

. (*) Work done while at the Australian National University and Data61

© 2022 H. Husain* & J. Knoblauch.

sub-class of Gibbs posteriors constructed under relatively strict assumptions (see e.g. Germain et al., 2016; Guedj, 2019). Deviating even more extremely from Gibbs posterior distributions, the belief distributions in Reid et al. (2015); Knoblauch et al. (2019); Knoblauch (2019a); Alquier (2020) have dependencies on the prior that are non-multiplicative. For all these distributions—which we shall subsume under the name of *generalized Bayesian procedures* from here on out—the arguments in favour of standard Bayesian inference do not apply: Gibbs posteriors do not even depend on likelihoods, and the belief distributions of Reid et al. (2015); Knoblauch et al. (2019); Knoblauch (2019a); Alquier (2020) are *directly* formulated via optimization problems without even attempting to mimic the multiplicative nature of Bayes' rule. In spite of this substantial lack in theoretical arguments, the empirical performance of these approaches is significantly better than that of point estimates, with no clear answer in sight.

In this paper, we provide a new and universal explanation for the robustness of generalized Bayesian procedures. Instead of the update-rule interpretation of Bayesian inference, we draw on the optimization-centric viewpoint on Bayesian inference advocated for by Knoblauch et al. (2019). Doing so has a distinct advantage: This view—and by extensions the results we derive from it—encompass a wide range of Machine Learning methods, many of which are not even motivated as extensions to traditional Bayesian inference. For instance, in Example 7 we show that our results apply even to the Wasserstein Autoencoder (Tolstikhin et al., 2017). While the connections between traditional Bayesian inference and Wasserstein Autoencoders are tenuous at first glance, both turn out to share a substantial amount of structure once re-phrased as an optimization problem.

The second advantage of adopting an optimization-centric view on Bayesian inference is immediate: We can use ideas and proof techniques from optimization and apply them in the context of generalized Bayesian procedures. In fact, this general idea has precedence in a successful line of research using the tools of Γ-convergence (Braides et al., 2002; Dal Maso, 2012) to study the consistency and posterior concentration of Bayesian Inverse Problems and Variational Approximations to standard Bayesian posteriors (see e.g. Agapiou et al., 2012; Lu, 2017; Lu et al., 2017; Wang and Blei, 2018; Knoblauch, 2019a).

In the current paper, we use the optimization-centric formulation of Bayesian procedures to tap into a very different branch of optimization: Duality Theory. Doing so allows us to leverage the structural and constraint properties of the optimization problem to explain the robustness of generalized Bayesian procedures via *Fenchel* duality. While Fenchel duality has been extremely useful for the theoretical study of other machine learning methods such as Generative Adversarial Networks (see Farnia and Tse, 2018; Liu and Chaudhuri, 2018; Husain et al., 2019) and regularization (Husain, 2020), the current paper constitutes the first analysis of this kind for Bayesian methods. Our main finding is a fundamental connection between risk robustness and the variational optimization problem underlying Bayesian inference. This finding advances our understanding of Bayesian methods: Specifically, it provides a new, concise, and rigorous explanation why a large class of Bayesian and Bayes-like methods typically outperform point estimation methods. In summary, our technical contributions are

1. (Theorem 5) A duality theorem that shows the the link between the optimization-centric view of Bayesian inference and an adversarial robustness game.

2. (Theorem 7) A theorem detailing the robustness of Bayesian posteriors as solutions to the adversarial loss.

3. Derivation of adversarial games for a variety of different Bayesian inference schemes, including a novel formulation of the Wasserstein Autoencoders (WAE). This allows us to gain novel insight into the theoretical underpinnings behind WAEs.

2. Motivation

Standard Bayesian inference is typically formulated as an update derived from Bayes' rule (see Bayes, 1763; De Laplace, 1774): Given observations $\{x_i \in \mathcal{X}\}_{i=1}^n$ sampled from a probability measure $\mathbb{P}$ on $\mathcal{X}^n$, a likelihood model $p_\theta : \mathcal{X}^n \to \mathbb{R}_{\geq 0}$ indexed by a parameter $\theta \in \Theta$ and a prior belief $\pi(\theta) \in \mathcal{P}(\Theta)$ about good values of θ, the Bayesian posterior is given by

$$q_{\mathrm{B}}(\theta) = \frac{\pi(\theta) p_\theta(x_{1:n})}{\int_\Theta \pi(\theta) p_\theta(x_{1:n}) d\theta}. \tag{1}$$

Because the integral $\int_\Theta \pi(\theta) p_\theta(x_{1:n}) d\theta$ is generally intractable, any down-stream computation involving q_{B} will require some form of approximation. One of the most prominent approximation schemes is Variational Inference (VI), which is built on the optimization-centric view on Bayesian inference. More specifically, it uses the fact that one may rewrite q_{B} as the unique minimum to an optimization problem which for $\mathrm{KL}(q, \pi)$ denoting the forward Kullback-Leibler Divergence (Kullback and Leibler, 1951) between q and π is given as

$$q_{\mathrm{B}} = \inf_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{q(\theta)} \left[-\log p_\theta(x_{1:n}) \right] + \mathrm{KL}(q, \pi) \right\}. \tag{2}$$

This *optimization-centric perspective* of Bayesian inference is useful for understanding the role of priors and likelihoods. On the one hand, the KL term *regularizes* the problem by forcing the posterior belief to not deviate arbitrarily far from the prior. On the other hand, the negative log likelihood term acts as a *loss* function relating the parameter of interest θ to the observations $x_{1:n}$. In other words, Bayesian inference is a very specific regularized loss-minimization problem. Adopting this optimization-centric view on exact Bayesian inference, one can interpret Variational Inference (VI) as introducing *constraints*. In particular, rather than minimizing over $\mathcal{P}(\Theta)$, VI instead minimizes the same objective over the parameterized subset $\mathcal{Q} \subset \mathcal{P}(\Theta)$. Writing this explicitly, the variational posterior q_{VI} is

$$q_{\mathrm{VI}} = \inf_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{q(\theta)} \left[-\log p_\theta(x_{1:n}) \right] + \mathrm{KL}(q, \pi) \right\} \text{ s.t. } q \in \mathcal{Q}. \tag{3}$$

Whether one computes the exact Bayesian posteriors q_{B} or its approximation q_{VI}, both are conceptually limited: In particular, the invocation of Bayes' rule is only valid if the likelihood model p_θ and the prior belief π are available and correctly specified. Strictly speaking, this means that π has to encode *all* our prior knowledge about θ and that we need $p_{\theta_0} = \mathbb{P}$ for some $\theta_0 \in \Theta$, an assumption sometimes referred to as the *M-closed world* (Bernardo, 2000).

In practice of course, likelihoods and priors alike are misspecified—often severly so. In spite of substantial misspecification, Bayesian posteriors typically perform surprisingly well. The fact that belief distributions improve performance relative to point estimates even when the standard decision-theoretic arguments in favour of q_{B} are not applicable has inspired numerous generalized Bayesian procedures (see e.g. Grünwald, 2011, 2012; Berger et al., 1994; Hooker and Vidyashankar, 2014; Ghosh and Basu, 2016; Bissiri et al., 2016; Jewson et al., 2018; Knoblauch et al., 2018;

Futami et al., 2018; Miller and Dunson, 2019; Knoblauch, 2019a,b; Chérief-Abdellatif and Alquier, 2019; Nakagawa and Hashimoto, 2019; Knoblauch and Vomfell, 2020; Guedj, 2019; Alemi, 2019). While the conceptual motivations of these generalized Bayesian procedures vary greatly, their large majority can be expressed as solutions to a modified version of the optimization problem in eq. (2) (see Knoblauch et al., 2019). Adopting terminology of the same paper, we call the resulting distributions *Generalized Variational Inference* posteriors. To properly define this large class of posteriors, we first state the definition of statistical divergences used for our results.

Definition 1 (Divergences) *A divergence is a function* $\mathsf{D} : \mathcal{P}(\Theta) \times \mathcal{P}(\Theta) \to \mathbb{R}$*, so that* $D(\mu, \nu) \geq 0$*,* $D(\mu, \nu) = 0 \iff \nu = \mu$*, and which is proper convex lower semi-continuous in its first argument.*

Generalizing the structure underlying eq. (3) in three directions at once, Generalized Variational Inference posteriors are obtained as the minimizers of any D-regularized Π-constrained L-loss minimization problem. Denoting $\mathcal{F}_b(\Theta)$ as the set of bounded and measurable functions on Θ, we proceed by formally defining posteriors of this kind.

Definition 2 (Generalized Variational Inference) *For any* $\Pi \subseteq \mathcal{P}(\Theta)$*, any loss* $L \in \mathcal{F}_b(\Theta)$*, divergence* $\mathsf{D} : \mathcal{P}(\Theta) \times \mathcal{P}(\Theta) \to \mathbb{R}$*, the Generalized Variational Inference (GVI) objective with prior* $\pi \in \mathcal{P}(\Theta)$ *and* $\lambda > 0$ *is*

$$\mathscr{G}_{L,\mathsf{D},\Pi} := \inf_{q \in \Pi} \Big(\mathbb{E}_{q(\theta)} \left[L(\theta) \right] + \lambda \mathsf{D}(q, \pi) \Big) \tag{4}$$

and any minimizer of the above is referred to as a GVI posterior:

$$q_{L,\mathsf{D},\Pi} \in \arg\inf_{q \in \Pi} \Big(\mathbb{E}_{q(\theta)} \left[L(\theta) \right] + \lambda \mathsf{D}(q, \pi) \Big) \tag{5}$$

The special case of this problem when $\mathsf{D} = \mathrm{KL}$ and $\Pi = \mathcal{P}(\Theta)$ is well-studied: As eq. (2) shows, $q_{\mathrm{B}} = q_{L,\mathrm{KL},\mathcal{P}(\Theta)}$ whenever $L(\theta) = -\log p_\theta(x_{1:n})$. Similarly, if L is an arbitrary loss for which $\int_\Theta \pi(\theta) \exp\{-L(\theta)\} d\theta < \infty$, one recovers the Gibbs posterior—also called quasi-posterior (e.g. Ghosh and Basu, 2016) or Generalized posterior (Bissiri et al., 2016):

$$q_{L,\mathrm{KL},\mathcal{P}(\Theta)}(\theta) = \frac{\pi(\theta) \exp\{-\lambda^{-1} L(\theta)\}}{\int_\Theta \pi(\theta) \exp\{-\lambda^{-1} L(\theta)\} d\theta}. \tag{6}$$

Since $q_{L,\mathrm{KL},\mathcal{P}(\Theta)}$ requires computationally expensive sampling procedures to be made practically useful, one could alternatively approximate the posteriors $q_{L,\mathrm{KL},\mathcal{P}(\Theta)}$ with an element in a tractable and parameterized set of distributions $\mathcal{Q} \subset \mathcal{P}(\Theta)$. This recovers the standard VI posterior q_{VI} whenever $L(\theta) = -\log p_\theta(x_{1:n})$ and so-called Gibbs VI posteriors (see e.g. Alquier et al., 2016) for general losses $L(\theta) \neq -\log p_\theta(x_{1:n})$.

As a consequence, Definition 2 recovers the most well-known generalized Bayesian methods, both in their exact and approximate form. Beyond that however, it also encompasses a host of other belief distributions that have received only limited attention. These include the posteriors motivated in Reid et al. (2015); Knoblauch et al. (2019); Knoblauch (2019a); Alquier (2020), minimizers of non-standard PAC-Bayesian bounds such as those of Bégin et al. (2016); Alquier and Guedj

(2018); Ohnishi and Honorio (2020), and even the conditional distributions produced by Wasserstein Autoencoders (Tolstikhin et al., 2017).

Most importantly—and unlike the multiplicative update-rule of eq. (1)—Definition 2 constitutes an optimization-centric formulation of generalized Bayesian procedures. Crucially for the purposes of the current paper, this allows us to tap into duality theory. Duality is a corner stone of modern optimization theory and has previously been used for the derivation of new Bayesian methodology (e.g. Ganchev et al., 2010; Zhu et al., 2014; Dai et al., 2018). More in line with its use for our contribution, it is also a vital tool for building theoretical understanding of existing Machine Learning methods. For example, recent work have given insights for a number of areas including generative modelling (Liu and Chaudhuri, 2018; Husain et al., 2019), distributional robustness (Cranko et al., 2020; Husain, 2020), reinforcement learning (Husain et al., 2021) and optimal transport (Paty and Cuturi, 2020).

In the current paper, we add to this literature by deriving the Legendre-Fenchel dual of standard and generalized Bayesian procedures. This allows us to interpret Bayes-like methods in a new light. More precisely, they define an adversarial game: In the dual space, the prior belief becomes a cost function for an adversary seeking to change the loss function that is minimized.

3. Bayesian Inference as Adversarial Robustness

The constrained and regularized structure of the problem in Definition 2 makes it naturally amenable to an investigation into its *Legendre-Fenchel* dual. We pursue this endeavour in the remainder. First, Section 3.1 sets out notation and additional definitions. We then present the general results of our analysis in Section 3.2. Lastly, we elaborate on the implications of our results for two large classes of posteriors in Sections 3.3 and 3.4 by giving examples of the resulting dual problems as well as numerical demonstrations.

3.1. Preliminaries

Throughout, we will assume that the parameter space Θ as well as the data space $\mathcal{X}$ admit Polish topology. We will sometimes use $\mathscr{B}(\Theta)$ to denote the set of finitely-additive measures over Θ. Its topological dual space is the set of all bounded and measurable functions mapping from Θ to $\mathbb{R}$, which we denote by $\mathcal{F}_b(\Theta)$. Lastly, the set $\mathcal{P}(\Theta) \subset \mathscr{B}(\Theta)$ denotes the set of Borel probability measures on Θ.

To investigate the Legendre-Fenchel dual of the optimization problems in Definition 2, we also need to introduce the Legendre-Fenchel conjugate of the prior regularizers $\mathsf{D}(\cdot, \pi)$.

Definition 3 *For a given prior $\pi \in \mathcal{P}(\Theta)$, the Legendre-Fenchel conjugate of a regularizer $\mathsf{D}(\cdot, \pi) : \mathcal{P}(\Theta) \to \mathbb{R}$ is*

$$\mathsf{D}^{\star}_{\pi}(\ell) = \sup_{\mu \in \mathscr{B}(\Theta)} \left(\int_{\Theta} \ell d\mu - \mathsf{D}(\mu, \pi) \right), \tag{7}$$

for any $\ell \in \mathcal{F}_b(\Theta)$.

For convenience, we also define an auxiliary minimization problem which appears as part of the Legendre-Fenchel dual.

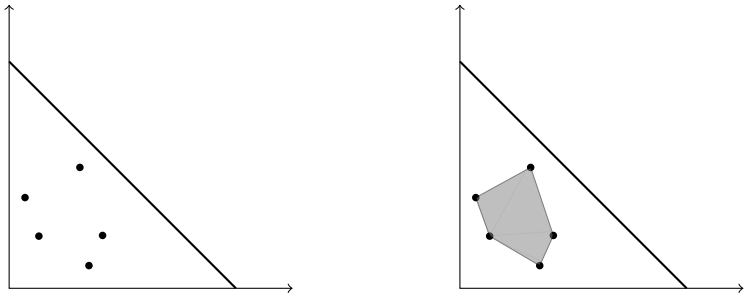

Figure 1: The left image illustrates a choice for Π which consists of five probability vectors over $\Theta = \{a, b, c\}$. The right illustrates $\overline{\text{co}}\,(\Pi)$ over this choice where one can see that the selection of probabilities increases vastly.

Definition 4 *For any set of probability distributions $\Pi \subseteq \mathcal{P}(\Theta)$, we define for any $L \in \mathcal{F}_b(\Theta)$*

$$\mathsf{E}_\Pi(L) = \inf_{q \in \Pi} \mathbb{E}_{q(\theta)}\left[L(\theta)\right] \tag{8}$$

In words, $\mathsf{E}_\Pi(L)$ denotes the smallest possible value achievable by integrating the loss L with an element from the class of probability distributions Π.

Lastly, we introduce the *closed convex hull* of a set Π of admissible solutions to the optimization problem in Definition 2. For a set of potential posteriors Π, $\overline{\text{co}}\,(\Pi)$ denotes the smallest *closed* and *convex* set containing Π. In particular, we will have

$$\lambda \cdot q + (1 - \lambda) \cdot q' \in \overline{\text{co}}\,(\Pi)\,, \tag{9}$$

for all $q, q' \in \Pi$ and $\lambda \in [0, 1]$. We illustrate this in Figure 1 for a discrete parameter space $\Theta = \{a, b, c\}$ with only three elements. In this setting, $\mathcal{P}(\Theta)$ is simply the set of vectors in $\mathbb{R}^3_{\geq 0}$ whose co-ordinates sum to 1 however can be viewed as elements in R^2 enclosed in a triangle with vertices $(0, 0)$, $(0, 1)$ and $(1, 0)$. Though the definition of the convex hull is best understood in the geometric sense, it also has a clear probabilistic counterpart in mixture models: For example, if Π is the set of normal distributions, then $\overline{\text{co}}\,(\Pi)$ is the set of all (finite and infinite) mixtures of normal distributions on Θ.

To clarify the setting studied throughout the remainder of the paper, we explain these definitions using a simple example that we will later re-use for numerical demonstrations.

Example 1 *Given a dataset $\{(X_i, Y_i)\}_{i=1}^n$ with $\mathcal{X} = \mathbb{R}^d$, one may consider the parameter space $\Theta = [0, 1]$ and the corresponding least squares loss*

$$L(\theta) = \sum_{i=1}^{n} \left(X_i^\top \theta - Y_i\right)^2. \tag{10}$$

Considering for some $m \in \mathbb{N}$ the discretely supported and uniform prior $\pi(\theta) = \frac{1}{m+1} \sum_{j=0}^{m} \delta_{(j/m)}(\theta)$ and the set of variational posteriors supported on the same atoms as

$$\Pi = \left\{ \sum_{j=0}^{m} w_j \delta_{(j/m)}(\theta) : w_j \geq 0, \sum_{j=0}^{m} w_j = 1 \right\},$$

it is clear both that $\pi \in \Pi$ *and that* $\overline{\text{co}}\,(\Pi) = \Pi$. *Considering as regularizer the* χ^2*-divergence given for any* $q \in \Pi$ *by*

$$\chi^2(q,\pi) = \mathbb{E}_{\pi(\theta)}\left[\left(\frac{\pi(\theta)}{q(\theta)} - 1\right)^2\right],$$

and taking $\lambda = 1$*, the corresponding Generalized Variational Inference posteriors are given by*

$$q_{L,\chi^2,\Pi} \in \arg\inf_{q\in\Pi}\left(\mathbb{E}_{q(\theta)}\left[L(\theta)\right] + \chi^2(q,\pi)\right).$$

Moreover, we have that

$$\mathsf{E}_\Pi(L) = \inf_{n\mathbb{R}} \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2)}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i^\top\theta - Y_i\right)^2\right], \tag{11}$$

The above corresponds to the square loss. If we employ a set of Gaussians as our model class $\Pi = \left\{\mathcal{N}(\mu,\sigma^2) : \mu \in \mathbb{R}^d, \sigma \in \mathbb{R}\right\}$ then we have

$$\mathsf{E}_\Pi(L) = \inf_{\mu\in\mathbb{R}^d,\sigma\in\mathbb{R}} \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2)}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i^\top\theta - Y_i\right)^2\right], \tag{12}$$

which amounts to fitting a normal distribution over the parameters and thus the value of $\mathbb{E}_\Pi[L]$ is precisely the minimally achievable loss using the family Π. It should be noted this is the GVI objective when there is no prior regularization: $\mathbb{E}_\Pi(L) = \mathscr{G}_{L,0,\Pi}$.

3.2. Main results

We present the main Theorem– a strong duality result for posteriors obtained as D-regularized and Π-constrained L-loss minimizers.

Theorem 5 (Strong Duality) *For any* $\Pi \subseteq \mathcal{P}(\Theta)$*,* $\pi \in \mathcal{P}(\Theta)$ *and loss* $L \in \mathcal{F}_b(\Theta)$*, the following holds for any* $\lambda > 0$

$$\mathscr{G}_{L,\mathsf{D},\overline{\text{co}}(\Pi)} = \sup_{\ell\in\mathcal{F}_b(\Theta)}\left(\mathsf{E}_\Pi(L+\ell) - \lambda\mathsf{D}_\pi^\star\left(\frac{\ell}{\lambda}\right)\right). \tag{13}$$

As this result shows, GVI procedures have a close correspondence to adversarial games. Specifically, the adversary in the game of Theorem 5 changes the original loss L via a perturbation ℓ so that the minimum achievable loss by Π given by $\mathbb{E}_\Pi[L+\ell]$, is as large as possible. Luckily, the adversary pays a price $\lambda D_\pi^\star(\ell/\lambda)$ for this, which stops it from making $\mathbb{E}_\Pi[L+\ell]$ infinitely large. Clearly, the exact choices for λ, π and D determine the precise cost of the adversarial perturbations ℓ. Note that $\Pi \subseteq \overline{\text{co}}\,(\Pi)$, which implies $\mathscr{G}_{L,\mathsf{D},\Pi} \geq \mathscr{G}_{L,\mathsf{D},\overline{\text{co}}(\Pi)}$. Combining this with Theorem 5 allows us to relate GVI (using only Π):

$$\mathscr{G}_{L,\mathsf{D},\Pi} \geq \sup_{\ell\in\mathcal{F}_b(\Theta)}\left(\mathsf{E}_\Pi(L+\ell) - \lambda\mathsf{D}_\pi^\star\left(\frac{\ell}{\lambda}\right)\right), \tag{14}$$

with equality if Π is closed and convex. The standard choice $\Pi = \mathcal{P}(\Theta)$ is closed and convex, immediately implying the following strong duality result that holds for full Bayesian inference as well as the the objects studied in (Reid et al., 2015; Knoblauch et al., 2019; Knoblauch, 2019a; Alquier, 2020).

Corollary 6 *If* $\Pi = \mathcal{P}(\Theta)$ *then for any* $\pi, \in \mathcal{P}(\Theta)$, $L \in \mathcal{F}_b(\Theta)$ *and* $\lambda > 0$ *it holds*

$$\mathscr{G}_{L,\mathsf{D},\Pi} = \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\mathsf{E}_\Pi(L + \ell) - \lambda \mathsf{D}_\pi^\star \left(\frac{\ell}{\lambda} \right) \right). \tag{15}$$

While Corollary 6 connects the two *values* of the objectives at the optimum, we can make this dual connection much firmer. In fact, the GVI primal and its adversarial dual share another—arguably even more important—connection: The GVI posterior minimizes the loss that results from the adversary's perturbation.

Theorem 7 (Adversarial Robustness of GVI) *Let* ℓ^* *denote a maximizer of the dual in* (13)*. If* Π *is convex and closed, it holds that*

$$q_{L,\mathsf{D},\Pi} \in \arg\inf_{q \in \Pi} \mathbb{E}_{q(\theta)} \left[L(\theta) + \ell^*(\theta) \right]. \tag{16}$$

This result is striking: It tells us that posteriors learned via the GVI objective of Definition 2 are adversarially robust. More specifically, they produce optimal beliefs in the presence of an adversary whose cost for perturbing the original loss L by ℓ is given by $\lambda \mathsf{D}_\pi^\star(\ell/\lambda)$.

3.3. Examples With f-Divergences

In this section we will explore our results in the popular setting when D is chosen as an f-divergence. We remark that Theorem 5 will be insightful and generalize results regarding log evidence however Theorem 7 will not give us much insight.

Example 2 (f-divergences) *The requirements of Definition 1 are satisfied by* f*-divergences. For a convex lower semicontinuous function* $f : \mathbb{R} \to (-\infty, \infty]$*, the corresponding* f*-divergence is* $\mathsf{D}_f(\mu, \nu) = \int_\Theta f(d\mu/d\nu) d\nu$ *if* $\nu \ll \mu$ *and* $\mathsf{D}_f(\mu, \nu) = \infty$ *otherwise. This includes the popular Kullback-Leibler (KL) divergence when* $f(t) = t \log t$ *and the* χ^2*-divergence if* $f(t) = (t-1)^2$*.*

At this point, the curious reader may wonder what the adversary's cost functions look like. To gain some intuition about this, we will study some examples with different choices of D. We begin with the standard choice corresponding to both Variational Inference and fully Bayesian Inference: The KL-divergence.

Example 3 (KL-divergence) *If* $\mathsf{D} = \mathrm{KL}$ *then the dual problem is*

$$\mathscr{G}_{L,\mathrm{KL},\mathcal{P}(\Theta)} = \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\mathsf{E}_\Pi(L + \ell) - \lambda \log \int_\Theta \exp\left(\frac{\ell(\theta)}{\lambda} \right) d\pi(\theta) \right). \tag{17}$$

As this example reveals, the KL-divergence penalizes the adversary for deviations ℓ from L proportionally to the prior belief π. This ultimately means that the prior becomes the adversary's cost function in the dual form. Jensen's inequality also gives a lower bound on the entire penalty. While coarse, this bound is perhaps more interpretable:

$$\lambda \log \int_\Theta \exp\left(\frac{\ell(\theta)}{\lambda} \right) d\pi(\theta) \geq \int_\Theta \ell(\theta) d\pi(\theta).$$

Specifically, it reveals that KL-regularization implies a perturbation penalty that is *stricter* and more costly than a linear one. The linear penalty (in ℓ) is a useful benchmark to compare against: Linear penalties punish the adversary by weighting its perturbation ℓ with the prior. Moreover, this holds for the case when D is chosen to be *any* f-divergence.

Example 4 (f-divergence) *For any lower semicontinuous convex function $f : \mathbb{R} \to (-\infty, \infty]$ with $f(1) = 0$, we have*

$$\mathscr{G}_{L,\mathsf{D}_f,\mathcal{P}(\Theta)} = \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\mathsf{E}_\Pi(L + \ell) - \inf_{b \in \mathbb{R}} \left[\int_\Theta f^\star(\ell(\theta) - b) d\pi(\theta) + b \right] \right), \tag{18}$$

where $f^\star(t) = \sup_{t' \in \mathrm{dom}(f)} (t \cdot t' - f(t'))$

Note for any f defined above, it holds that $f^\star(t) \geq t$ and so immediately we get

$$\inf_{b \in \mathbb{R}} \left[\int_\Theta f^\star(\ell(\theta) - b) d\pi(\theta) + b \right] \geq \int_\Theta \ell(\theta) d\pi(\theta) \tag{19}$$

As it turns out, the linear penalty term also re-surfaces quite naturally when other divergences are used in the GVI form. As the next two examples show, this includes the case of the χ^2-divergence.

Example 5 (χ^2-divergence) *If $\mathsf{D} = \lambda \cdot \chi^2$ for some $\lambda \geq 0$ then the dual problem is*

$$\mathscr{G}_{L,\chi^2,\mathcal{P}(\Theta)} = \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\mathsf{E}_\Pi(L + \ell) - \int_\Theta \ell(\theta) d\pi(\theta) - \frac{1}{4\lambda} \mathrm{Var}_\pi(\ell) \right). \tag{20}$$

We remark that this gives a recontextualization to the parameter λ, which is often added to enforce proximity to the prior however, here it is interpreted as allowing the loss perturbation ℓ to be as flexible for larger values.

3.4. Examples with Integral Probability Metrics

The choice of f-divergences can be regarded as a strong way of penalizing the posterior to deviate from the prior. This is due to the fact that the f-divergences requires absolute continuity, meaning that the posterior is forced to be supported wherever the prior is. To that end, we now consider a choice of divergence that is considered rather weaker, the Integral Probability Metrics (IPM).

Definition 8 (Integral Probability Metric) *For a set of functions $\mathcal{H} \subseteq \mathscr{F}_b(\Theta)$, the* IPM *between $\mu, \nu \in \mathcal{P}(\Theta)$ is*

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

IPMs have often been studied for theoretical interest in machine learning as they define metrics over probability spaces (Müller, 1997). One famous example is the 1-Wasserstein distance (Villani, 2008), which is typically sought as a remedy to strong penalizing effect of f-divergences. Another example of an IPM in machine learning is the kernel-based Maximum Mean Discrepancy, which can be easily computed. In particular, it is the choice of $\mathcal{H}$ that allows us to comment on the strength

of the IPM and has been shown that convergence rates of an IPM induced by $\mathcal{H}$ depends on the Rademacher complexity of $\mathcal{H}$ (Bartlett and Mendelson, 2002).

The downside of IPMs is that for a general class $\mathcal{H}$ cannot be easily computed, even if the densities of the distributions are known. Recently however, IPMs have become popularized due to Generative Adversarial Networks as deep neural networks have played the role of $\mathcal{H}$ with various kinds of parametrizations (Arbel et al., 2018; Li et al., 2017; Arjovsky et al., 2017; Mroueh et al., 2017; Mroueh and Sercu, 2017). They have also been used in the generalized variational inference framework, which as we exemplify with the Wasserstein Autoencoder. We first, apply our general result to the case when D is chosen to be an IPM.

Example 6 (Integral Probability Metric) *For a set of functions $\mathcal{H} \subseteq \mathcal{F}_b(\Theta)$, the Integral Probability Metric (IPM) generated by $\mathcal{H}$ is defined as $d_{\mathcal{H}}(q,\pi) = \sup_{h\in\mathcal{H}} (\mathbb{E}_q[h] - \mathbb{E}_\pi[h])$. If $\mathsf{D} = d_{\mathcal{H}}$ then the dual problem is*

$$\mathscr{G}_{L,\text{IPM},\mathcal{P}(\Theta)} = \sup_{\ell\in\lambda\cdot\mathcal{H}} \left(\mathsf{E}_\Pi(L+\ell) - \int_\Theta \ell(\theta) d\pi(\theta) \right). \tag{21}$$

Whether the primal form of GVI regularizes against the prior with a χ^2-divergence or an IPM, the adversary's cost function in the dual space additively decomposes into a linear penalty and an additional term. In the case of the χ^2-divergence, this additional term measures the variance of the perturbation (relative to the prior measure). In essence, this behaviour discourages perturbations whose fluctuations are large in regions of high prior mass. Furthermore, note that the dual of the IPM scheme also resembles that of the χ^2-divergence except that the IPM imposes a particularly *strong* penalty on the adversary: Any perturbation ℓ that is not in the set $\lambda \cdot \mathcal{H} = \{\lambda \cdot h : h \in \mathcal{H}\}$ incurs an infinitely large penalty. Conversely, this also means that it will be a relatively *weak* regularizer. In both cases, $\lambda \to \infty$ reduces the constraints on the penalty, however in the primal problem (GVI), a larger choice of λ suggests the regularization against the prior to be stronger and thus more constraint.

Example 7 (Wasserstein Autoencoder) *Consider the following instantiation: $\Theta = \mathcal{Z} \times \mathcal{X}$ where $\mathcal{Z}$ is referred to as a latent space. Let $G : \mathcal{Z} \to \mathcal{X}$ be a fixed mapping and suppose we have a cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then define the following (with $\theta = (z,x)$)*

$$L(\theta) = c(G(z), x) \tag{22}$$

$$\Pi = \{q \in \mathcal{P}(\Theta) : q(\mathcal{Z} \times A) = P_X(A), A \text{ measurable}\} \tag{23}$$

The Wasserstein Autoencoder objective (Tolstikhin et al., 2017) is given by

$$\inf_{G,q\in\Pi} \left(\int_\Theta c(G(z),x) dq(z,x) + \lambda \mathsf{D}(\mathsf{F}_\# q, \pi) \right), \tag{24}$$

where D is a divergence, $\mathsf{F} : \Theta \to \mathcal{Z}$ is a projection mapping defined as $F(z,x) = z$ and $\pi \in \mathcal{P}(\mathcal{Z})$ is a *prior* distribution over the latent space. We remark that it involves two infinums, one over the function G, which is referred to as the decoder and q, which is the encoder. In the above example, we show that for a fixed G, we can apply our duality result to the infinum problem over q.

The first term of the minimization over q is referred to as the *reconstruction* cost and the divergence term serves the role of smoothing the latent space. The choice of D is adhoc and chosen without well-posed motivation and left by (Tolstikhin et al., 2017) as the subject of theoretical investigation. For computational convenience, the MMD was chosen in (Tolstikhin et al., 2017) and furthermore in (Zhang et al., 2019) and (Patrini et al., 2018), the Wasserstein distance is chosen as the underlying D. Noticing that these prominent choices are both IPMs, we focus on this case where $\mathsf{D} = d_{\mathcal{H}}$ with $\mathcal{H} \subseteq \mathcal{F}_b(\mathcal{Z})$. To relate this to GVI, we require the divergence to be over Θ and not just $\mathcal{Z}$. This is easily fixed by simply embedding $\mathcal{H}$ into $\mathcal{F}_b(\Theta)$: $\tilde{\mathcal{H}} = \{f(x,z) = h(z) : h \in \mathcal{H}\}$. Moreover, we then define $\tilde{\mathsf{D}} = d_{\tilde{\mathcal{H}}}$ and $\tilde{\pi} = \pi \times \nu$ where $\nu \in \mathcal{P}(\mathcal{X})$ is an arbitrary probability measure. It then follows that the WAE objective is precisely $\mathscr{G}_{L,\tilde{\mathsf{D}},\Pi}$ with prior $\tilde{\pi}$. We now invoke our main result, noting that Π is closed and convex to derive the dual:

$$\mathscr{G}_{L_G,\tilde{\mathsf{D}},\Pi} = \sup_{\ell \in \mathcal{H}} \left(\mathsf{E}_{\Pi}\left[L_G + \ell\right] - \int_{\mathcal{Z}} \ell(z) d\pi(z) \right). \tag{25}$$

We use L as L_G to remind the dependence on G which makes the contribution to WAE in two-fold. First, it explicitly comments on the *robustness* of the encoder q as a solution to this adversarial problem given that IPMs are natural choices in this setting and therefore complimenting this choice of D as discussed above, despite theory only existing for when D is an f-divergence (Husain et al., 2019). Secondly, the minimization problem over G can now be interpreted as a min-max problem:

$$\inf_{G} \mathscr{G}_{L_G,\tilde{\mathsf{D}},\Pi} = \inf_{G} \sup_{\ell \in \mathcal{H}} \left(\mathsf{E}_{\Pi}\left[L_G + \ell\right] - \int_{\mathcal{Z}} \ell(z) d\pi(z) \right).$$

Therefore, the function G is minimizing the worst case reconstruction loss altered by an adversary budgeted in $\mathcal{H}$ with penalty based on the prior π. Thus the result provides a reinterpretation to *both* the generator and encoder training of WAE under this choice.

4. Conclusion

In this work, we exploit Fenchel duality to study the optimization-centric view of Bayesian inference and provide a foundational reinterpretation. Our findings provide commentary for existing methods such as f-divergence regularization but apply to Wasserstein Autoencoders (WAE) - a setting that is conceptually seperate from Bayesian inference yet belongs to the same family, attesting to the generality of our result.

Our work provides a number of avenues that are of particular practical interest. For example, our results suggest that IPM-regularization could provide more robust posterior inferences. While exploring applications of this insight goes beyond the scope of the current paper, we will study the methodological possibilities raised by our results in future work. our results theoretically show the advantages of using IPMs for robustifying posteriors in many settings whose choice would be studied for specialized domains. In a similar vein our dual form for the WAE hints at a new algorithm, our results suggest that IPM-regularization could provide more robust posterior inferences. While exploring applications of this insight goes beyond the scope of the current paper, we will study the methodological possibilities raised by our results in future work.

References

Sergios Agapiou, Stig Larsson, and Andrew M Stuart. Posterior consistency of the bayesian approach to linear ill-posed inverse problems. 2012.

Alexander A. Alemi. Variational predictive information bottleneck. In *Workshop on Information Theory, Advances in Neural Information Processing Systems*, 2019.

Pierre Alquier. Non-exponentially weighted aggregation: regret bounds for unbounded loss functions. *arXiv preprint arXiv:2009.03017*, 2020.

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

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.

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

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.

Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444, 2016.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

James O. Berger, Elías Moreno, Luis Raul Pericchi, M. Jesús Bayarri, José M. Bernardo, Juan A. Cano, Julián De la Horra, Jacinto Martín, David Ríos-Insúa, Dasgupta A. Betrò, Bruno, Paul Gustafson, Larry Wasserman, Joseph B. Kadane, Cid Srinivasan, Michael Lavine, Anthony O'Hagan, Wolfgang Polasek, Christian P. Robert, Constantinos Goutis, Fabrizio Ruggeri, Gabriella Salinetti, and Siva Sivaganesan. An overview of robust Bayesian analysis. *Test*, 3(1): 5–124, 1994.

José M. Bernardo. Bayesian theory. *Wiley Series in Probability and Statistics. 23 cm. 586 p.*, 2000.

Pier Giovanni Bissiri, Chris Holmes, and Stephen Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 1103–1130, 2016.

Andrea Braides et al. *Gamma-convergence for Beginners*, volume 22. Clarendon Press, 2002.

Caitlin E Buck and Bo Meson. On being a good Bayesian. *World Archaeology*, 47(4):567–584, 2015.

Olivier Catoni. Pac-bayesian supervised classification. *Lecture Notes-Monograph Series. IMS*, 2007.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. *arXiv preprint arXiv:1909.13339*, 2019.

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

Bo Dai, Hanjun Dai, Niao He, Weiyang Liu, Zhen Liu, Jianshu Chen, Lin Xiao, and Le Song. Coupled variational Bayes via optimization embedding. In *Advances in Neural Information Processing Systems 31*, pages 9713–9723, 2018.

Gianni Dal Maso. *An introduction to* Γ*-convergence*, volume 8. Springer Science & Business Media, 2012.

Pierre-Simon De Laplace. Mémoire sur la probabilité des causes par les événements. *Mém. de math. et phys. présentés à l'Acad. roy. des sci*, 6:621–656, 1774.

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

Farzan Farnia and David Tse. A convex duality framework for gans. In *Advances in Neural Information Processing Systems*, pages 5248–5258, 2018.

Edwin Fong and Chris Holmes. On the marginal likelihood and cross-validation. *arXiv preprint arXiv:1905.08737*, 2019.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 813–822. PMLR, 2018.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, 2010.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

Abhik Ghosh and Ayanendranath Basu. Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

Michael Goldstein. Subjective Bayesian analysis: principles and practice. *Bayesian Analysis*, 1(3): 403–420, 2006.

Peter Grünwald. Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420, 2011.

Peter Grünwald. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer, 2012.

Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Chris Holmes and Stephen Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

Giles Hooker and Anand N Vidyashankar. Bayesian model robustness via disparities. *Test*, 23(3): 556–584, 2014.

Hisham Husain. Distributional robustness with ipms and links to regularization and gans. *arXiv preprint arXiv:2006.04349*, 2020.

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.

Hisham Husain, Kamil Ciosek, and Ryota Tomioka. Regularized policies are reward robust. In *International Conference on Artificial Intelligence and Statistics*, pages 64–72. PMLR, 2021.

Jack Jewson, Jim Smith, and Chris Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

Jeremias Knoblauch. Frequentist consistency of generalized variational inference. *arXiv preprint arXiv:1912.04946*, 2019a.

Jeremias Knoblauch. Robust deep gaussian processes. *arXiv preprint arXiv:1904.02303*, 2019b.

Jeremias Knoblauch and Lara Vomfell. Robust bayesian inference for discrete outcomes with the total variation distance. *arXiv preprint arXiv:2010.13456*, 2020.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Doubly robust Bayesian inference for non-stationary streaming data using β-divergences. In *Advances in Neural Information Processing Systems*, pages 64–75, 2018.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*, 2019.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

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.

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

Yulong Lu. On the bernstein-von mises theorem for high dimensional nonlinear bayesian inverse problems. *arXiv preprint arXiv:1706.00289*, 2017.

Yulong Lu, Andrew Stuart, and Hendrik Weber. Gaussian approximations for probability measures on $\mathbb{R}^d$. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):1136–1165, 2017.

SP Lyddon, CC Holmes, and SG Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

David A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999a.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM, 1999b.

Jeffrey W. Miller and David B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

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

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

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

Tomoyuki Nakagawa and Shintaro Hashimoto. Robust Bayesian inference via γ-divergence. *Communications in Statistics-Theory and Methods*, pages 1–18, 2019.

Anthony O'Hagan and Jeremy E Oakley. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering & System Safety*, 85(1):239–248, 2004.

Juki Ohnishi and Jean Honorio. Novel change of measure inequalities with applications to pac-bayesian bounds and monte carlo estimation. *arXiv preprint arXiv:2002.10678*, 2020.

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.

François-Pierre Paty and Marco Cuturi. Regularized optimal transport is ground cost adversarial. In *International Conference on Machine Learning*, pages 7532–7542. PMLR, 2020.

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

Mark D Reid, Rafael M Frongillo, Robert C Williamson, and Nishant Mehta. Generalized mixability via entropic duality. In *Conference on Learning Theory*, pages 1501–1522, 2015.

R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

Ralph Rockafellar. Integrals which are convex functionals. *Pacific journal of mathematics*, 24(3): 525–539, 1968.

John Shawe-Taylor and Robert C Williamson. A PAC analysis of a Bayesian estimator. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 2–9, 1997.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

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

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, pages 1–15, 2018.

Peter M Williams. Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

Constantin Zalinescu. *Convex analysis in general vector spaces*. World scientific, 2002.

Arnold Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

Shunkang Zhang, Yuan Gao, Yuling Jiao, Jin Liu, Yang Wang, and Can Yang. Wasserstein-wasserstein auto-encoders. *arXiv preprint arXiv:1902.09323*, 2019.

Jun Zhu, Ning Chen, and Eric P Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1):1799–1847, 2014.

Appendix A. 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 $\mathcal{F}_b(\Theta)$, noting that $\mathcal{F}_b(\Theta)$ is a Hausdorff locally convex space (through the uniform norm). We use $\mathscr{B}(\Theta)$ to denote the denote the set of all bounded and finitely additive signed measures over Θ (with a given σ-algebra). For any set $D \subseteq \mathscr{B}(\Theta)$ and $h \in \mathcal{F}_b(\Theta)$, we use $\sigma_D(h) = \sup_{\nu \in D} \langle h, \nu \rangle$ and $\iota_D(\nu) = \infty \cdot [\![\nu \notin D]\!]$ to denote the *support* and *indicator* functions such as in Rockafellar (1970). We introduce the conjugate specific to these spaces

Definition 9 (Rockafellar (1968)) *For any proper convex function $F : \mathscr{B}(\Theta) \to (-\infty, \infty)$, we have for any $h \in \mathcal{F}_b(\Theta)$ we define*

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

and for any $\mu \in \mathscr{B}(\Theta)$ we define

$$F^{\star\star}(\mu) = \sup_{h \in \mathcal{F}_b(\Theta)} \left(\int_\Theta h d\mu - F^\star(h) \right).$$

Theorem 10 (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 $F^{\star\star} = F$.*

A.1. Proof of Theorem 5

Lemma 11 *For any $\Pi \subseteq \mathcal{P}(\Theta)$ and $L \in \mathcal{F}_b(\Theta)$, we have*

$$\mathbb{E}_{\overline{\text{co}}(\Pi)}[L] = \mathbb{E}_\Pi[L]. \tag{26}$$

Proof For any $n \in \mathbb{N}$, we denote $\Delta_n = \{\alpha \in [0,1]^n : \sum_{i=1}^n \alpha_i = 1\}$. We then have

$$\begin{aligned}
\mathbb{E}_{\overline{\text{co}}(\Pi)}[L] &= \inf_{q \in \overline{\text{co}}(\Pi)} \mathbb{E}_q[L] && (27)\\
&= \inf_{n \in \mathbb{N}: \alpha \in \Delta_n, q_i \in \Pi, \forall i = 1, \ldots, n} \mathbb{E}_{\sum_{i=1}^n \alpha_i q_i}[L] && (28)\\
&\overset{(1)}{=} \inf_{n \in \mathbb{N}: \alpha \in \Delta_n, q_i \in \Pi, \forall i = 1, \ldots, n} \sum_{i=1}^n \alpha_i \mathbb{E}_{q_i}[L] && (29)\\
&= \inf_{n \in \mathbb{N}: \alpha \in \Delta_n} \sum_{i=1}^n \alpha_i \inf_{q_i \in \Pi} \mathbb{E}_{q_i}[L] && (30)\\
&= \inf_{n \in \mathbb{N}: \alpha \in \Delta_n} \sum_{i=1}^n \alpha_i \mathbb{E}_\Pi[L] && (31)\\
&= \mathbb{E}_\Pi[L], && (32)
\end{aligned}$$

where (1) holds due to linearity of expectation. Moreover, since $\mathbb{E}_\Pi$ is the infinum of linear function, it follows that taking the closure will be attained and thus justifies $\overline{\text{co}}\,(\Pi)$. ■

We will employ Theorem 10 to derive the duality result and present in the form of an early Lemma for consistency in notation.

Lemma 12 *For any prior $\pi \in \mathcal{P}(\Theta)$, we have*

$$\mathsf{D}(q,\pi) = \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\mathbb{E}_{q(\theta)}[\ell(\theta)] - \mathsf{D}_\pi^\star(\ell) \right) \tag{33}$$

Proof Using Theorem 10, we have $\mathsf{D} = \mathsf{D}^{\star\star}$ since D is proper convex and lower semicontinuous by assumption and by applying Definition 9, we get the desired result. ■

We then need a Lemma that takes care of technical conditions

Lemma 13 *For any prior $\pi \in \mathcal{P}(\Theta)$, regularizer D and set $\Pi \subseteq \mathcal{P}(\Theta)$, define a function $F : \mathcal{P}(\Theta) \times \mathcal{F}_b(\Theta) \to \mathbb{R}$ as*

$$F(q,\ell) = \mathbb{E}_{q(\theta)}\left[L(\theta)\right] + \mathbb{E}_{q(\theta)}\left[\ell(\theta)\right] - \mathsf{D}_\pi^\star\left(\ell\right) + \iota_{\overline{\mathrm{co}}(\Pi)}(q). \tag{34}$$

It holds that

$$\inf_{q \in \mathcal{P}(\Theta)} \sup_{\ell \in \mathcal{F}_b(\Theta)} F(q,\ell) = \sup_{\ell \in \mathcal{F}_b(\Theta)} \inf_{q \in \mathcal{P}(\Theta)} F(q,\ell) \tag{35}$$

Proof First note that since $L, \ell \in \mathcal{F}_b(\Theta)$ and $\overline{\mathrm{co}}\left(\Pi\right)$ is closed and convex (by construction), it holds that the mapping $q \mapsto F(q,\ell)$ is convex and lower semicontinuous (Penot, 2012). Furthermore note that $\mathsf{D}_\pi^\star(\ell)$ is convex and lower semicontinuous for any choice of D and since $q \in \mathcal{P}(\Theta) \subset \mathscr{B}(\Theta)$, it follows that the mapping $\ell \mapsto F(q,\ell)$ is also convex and lower semicontinuous. Next, by endowing $\mathscr{B}(\Theta)$ with the topology via Banach-Alaoglu using strong duality between $\mathcal{F}_b(\Theta)$ and $\mathcal{P}(\Theta)$, it follows that $\mathcal{P}(\Theta)$ is compact (Liu and Chaudhuri, 2018, Lemma 27(b)). Finally, noting that all conditions for Ky Fan's minimax Theorem are satisfied (Fan, 1953, Theorem 2), the result follows. ■

We now proceed to prove the main result.

$$\begin{aligned}
&\inf_{q \in \overline{\mathrm{co}}(\Pi)} \left(\mathbb{E}_{q(\theta)}\left[L(\theta)\right] + \mathsf{D}(q,\pi) \right) \\
&\overset{(1)}{=} \inf_{q \in \overline{\mathrm{co}}(\Pi)} \left(\mathbb{E}_{q(\theta)}\left[L(\theta)\right] + \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\mathbb{E}_{q(\theta)}\left[\ell(\theta)\right] - \mathsf{D}_\pi^\star\left(\ell\right) \right) \right) \\
&= \inf_{q \in \mathcal{P}(\Theta)} \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\mathbb{E}_{q(\theta)}\left[L(\theta)\right] + \mathbb{E}_{q(\theta)}\left[\ell(\theta)\right] - \mathsf{D}_\pi^\star\left(\ell\right) + \iota_{\overline{\mathrm{co}}(\Pi)}(q) \right) \\
&\overset{(2)}{=} \sup_{\ell \in \mathcal{F}_b(\Theta)} \inf_{q \in \mathcal{P}(\Theta)} \left(\mathbb{E}_{q(\theta)}\left[L(\theta)\right] + \mathbb{E}_{q(\theta)}\left[\ell(\theta)\right] - \mathsf{D}_\pi^\star\left(\ell\right) + \iota_{\overline{\mathrm{co}}(\Pi)}(q) \right) \\
&= \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\inf_{q \in \mathcal{P}(\Theta)} \left(\mathbb{E}_{q(\theta)}\left[L(\theta) + \ell(\theta)\right] + \iota_{\overline{\mathrm{co}}(\Pi)}(q) \right) - \mathsf{D}_\pi^\star\left(\ell\right) \right) \\
&\overset{(3)}{=} \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\mathsf{E}_{\overline{\mathrm{co}}(\Pi)}(L + \ell) - \mathsf{D}_\pi^\star\left(\ell\right) \right) \\
&\overset{(4)}{=} \sup_{\ell \in \mathcal{F}_b(\Theta)} \left(\mathsf{E}_{\Pi}(L + \ell) - \mathsf{D}_\pi^\star\left(\ell\right) \right),
\end{aligned}$$

where (1) is due to Lemma 12, (2) is due to Lemma 13, (3) is by definition of E_Π and (4) holds due to Lemma 11. The proof concludes by noting that the dual of $\lambda\mathsf{D}(\cdot,\pi)$ is $\lambda\mathsf{D}_\pi^\star(\cdot/\lambda)$.

A.2. Proof of Theorem 7

First note that $\mathbb{E}_{q_{\mathsf{D},L,\Pi}(\theta)}\left[L(\theta)+\ell^*(\theta)\right] \geq \inf_{q\in\Pi}\mathbb{E}_{q(\theta)}\left[L(\theta)+\ell^*(\theta)\right]$ by definition. For the other direction, we have

$$
\begin{aligned}
&\inf_{q\in\Pi}\mathbb{E}_{q(\theta)}\left[L(\theta)+\ell^*(\theta)\right]-\mathbb{E}_{q_{\mathsf{D},L,\Pi}(\theta)}\left[L(\theta)+\ell^*(\theta)\right]\\
&=\left(\mathbb{E}_\Pi[L+\ell^*]-\lambda\mathsf{D}_\pi^\star\left(\frac{\ell^*}{\lambda}\right)\right)+\lambda\mathsf{D}_\pi^\star\left(\frac{\ell^*}{\lambda}\right)-\mathbb{E}_{q_{\mathsf{D},L,\Pi}(\theta)}\left[L(\theta)+\ell^*(\theta)\right]\\
&\overset{(1)}{=}\sup_{\ell\in\mathcal{F}_b(\Theta)}\left(\mathbb{E}_\Pi[L+\ell]-\lambda\mathsf{D}_\pi^\star\left(\frac{\ell}{\lambda}\right)\right)+\lambda\mathsf{D}_\pi^\star\left(\frac{\ell^*}{\lambda}\right)-\mathbb{E}_{q_{\mathsf{D},L,\Pi}(\theta)}\left[L(\theta)+\ell^*(\theta)\right]\\
&\overset{(2)}{=}\inf_{q\in\Pi}\left(\mathbb{E}q[L]+\lambda\mathsf{D}(q,\pi)\right)+\lambda\mathsf{D}_\pi^\star\left(\frac{\ell^*}{\lambda}\right)-\mathbb{E}_{q_{\mathsf{D},L,\Pi}(\theta)}\left[L(\theta)+\ell^*(\theta)\right]\\
&\overset{(3)}{=}\mathbb{E}q_{\mathsf{D},L,\Pi}[L]+\lambda\mathsf{D}\left(q_{\mathsf{D},L,\Pi},\pi\right)+\lambda\mathsf{D}_\pi^\star\left(\frac{\ell^*}{\lambda}\right)-\mathbb{E}_{q_{\mathsf{D},L,\Pi}(\theta)}\left[L(\theta)+\ell^*(\theta)\right]\\
&=\lambda\mathsf{D}\left(q_{\mathsf{D},L,\Pi},\pi\right)+\lambda\mathsf{D}_\pi^\star\left(\frac{\ell^*}{\lambda}\right)-\mathbb{E}_{q_{\mathsf{D},L,\Pi}(\theta)}\left[\ell^*(\theta)\right]\\
&\overset{(4)}{\geq}0,
\end{aligned}
$$

where (1) is due to the optimality of ℓ^*, (2) is via Theorem 5 noting that Π is closed and convex by assumption, (3) is due to optimality of $q_{\mathsf{D},L,\Pi}$ and (4) holds by applying the Fenchel-Young inequality on D.

A.3. Proof of Example 4

Note that when we pick D as an f-divergence, there is a standard result we can recall in the following lemma.

Lemma 14 *For any lower semicontinuous convex function $f:\mathbb{R}\to(-\infty,\infty]$ with $f(1)=0$ so that $\mathsf{D}_\pi=\mathsf{D}_f(\cdot,\pi)$, $\mu\in\mathcal{P}(\Theta)$ and $h\in\mathcal{F}_b(\Theta)$, it holds that*

$$
\mathsf{D}_\pi^\star(h)=\inf_{b\in\mathbb{R}}\left(\mathbb{E}_\pi[f^\star(h-b)]+b\right), \tag{36}
$$

where $f^\star(t)=\sup_{t'}\left(tt'-f(t')\right)$ is the convex conjugate.

A proof of this result can be found in Equation (22) of (Liu and Chaudhuri, 2018). Using this, we can now prove the example for the Kullback-Leibler and χ^2 divergences.

A.4. Proof of Example 3

Noting that $f^\star(t)=\exp(t-1)$, the inner b problem can easily be solved:

$$
\begin{aligned}
\inf_{b\in\mathbb{R}}\left(\mathbb{E}_\pi[\exp(h-b)]+b\right)&=\inf_{b\in\mathbb{R}}\left(\exp(-b)\cdot\mathbb{E}_\pi[\exp h]+b\right) &(37)\\
&=\log\mathbb{E}_\pi[\exp h] &(38)
\end{aligned}
$$

A.5. Proof of Example 5

In this case we have $f^\star(t) = t + \frac{t^2}{4}$ and in particular $(\lambda f)^\star(t) = t + \frac{t^2}{4\lambda}$. The infinum problem, similar to the KL case becomes easily tractable:

$$\inf_{b\in\mathbb{R}} \left(\mathbb{E}_\pi[h] + \frac{1}{4\lambda}\mathbb{E}_\pi\left[(h-b)^2\right] \right) = \mathbb{E}_\pi[h] + \frac{1}{4\lambda} \inf_{b\in\mathbb{R}} \mathbb{E}_\pi\left[(h-b)^2\right] \tag{39}$$

$$= \mathbb{E}_\pi[h] + \frac{1}{4\lambda} \operatorname{Var}_\pi[h] \tag{40}$$

A.6. Proof of Example 6

For this case, we invoke (Husain, 2020, Lemma 5) which in combination with our main result yields the desired result.