GENERALIZATION FOR DISCRIMINATOR-GUIDED DIFFUSION MODELS VIA STRONG DUALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In the past few years, score-based generative models (SGMs) and diffusion models have proven to be efficient methods for learning distributions and have been of great practical significance. However, only a few lines of work are attempting to understand the theoretical guarantees of such models, and only one recent work Oko et al. (2023) focuses on the generalization abilities. In this work, we extend the study of generalization in SGMs and look to answer how model complexity emerges as a key player in the success of these models. For example, in other deep generative models, such as Generative Adversarial Networks (GANs), it has been revealed that the complexity of the discriminator set plays a crucial role in generalization. We prove that when diffusion models are further refined by discriminators (Kim et al., 2022a), the Integral Probability Metric (IPM) can be exactly represented through strong duality. Our findings advocate for discriminator refinement of deep generative models and, more specifically, unveil the generalization effect of using regularized discriminators in this setting. This result validates existing work on discriminator refinement to a great deal of generality. Therefore, our work provides theoretical validation for existing practices, provides a notion of regularization for SGMs, and contributes to the understanding of efficient distributional learning at large.

1 INTRODUCTION

In the past few years, score-based generative models (SGMs) (Hyvärinen & Dayan, 2005; Vincent, 2011; Song & Ermon, 2019) have emerged as a popular method to efficiently learn distributions for several different domains such as images (Chung & Ye, 2022; Batzolis et al., 2021; Ruiz et al., 2023), text (Popov et al., 2021; Kim et al., 2022b), graph (Fan et al., 2023; Zhang et al., 2023) and audio data (Serrà et al., 2022; Pascual et al., 2023; Richter et al., 2023; Wu, 2023; Qiang et al., 2023). Some examples of this include the Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020), along with the Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020a) that have demonstrated success at a large scale such as DALL-E 2 (Ramesh et al., 2022).

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Nowozin et al., 2016) are a similar line of work to learning distributions from data using a discriminator. It has been shown that learning this discriminator for GANs admits a (weakly) dual problem (Husain et al., 2019), which corresponds to learning an encoder, coinciding with the objective derived in Wasserstein Autoencoders (WAE). However, the original motivation for discriminators in GANs is to guide a generative model on regions of under-performance with the use of a critic, i.e., the discriminator. Recently, there have been attempts to use discriminators to improve and refine SGMs in the same sense, such as in (Kim et al., 2022a), which explicitly learns a density ratio and corrects the diffusion model. It was shown that this combination of discriminator and score-based model achieves a remarkable performance improvement.

Despite the practical success, the theoretical understanding of these generative models are not very well understood at large. In the case of SGMs, the majority of results do not consider discriminator intervention and prove convergence guarantees against observed data under various assumptions (Chen et al., 2022; Lee et al., 2022; 2023; Li et al., 2023; De Bortoli, 2022). Recently, Oko et al. (2023) considered the generalization performance and showed that SGMs with neural networks can achieve minimax optimal rates under smoothness assumptions of the true density.

Given the practical significance and considerable empirical benefits of discriminator-guided SGMs, convergence and generalization in such settings remain mystified. In this work, we target this exact problem. In particular, we generalize the algorithm presented in (Kim et al., 2022a) where one learns a discriminator by minimizing a proper composite loss (derived from an f-divergence), which recovers Kim et al. (2022a) in the setting of Jensen-Shannon (JS) divergence.

In order to prove these results, we revisit the duality structures existing in the GAN set-up, which take the following form:

$$\sup_{h \in \mathcal{H}} \mathsf{R}(h) = \inf_{\mu} \mathsf{L}(\mu), \tag{1}$$

where $\mathcal{H}$ is the set of discriminators and μ is taken over all probability measures. We derive conditions under which we can construct the optimal μ^* and find that it corresponds *exactly* to the refined generative model constructed in discriminator-guided diffusion models (Kim et al., 2022a). This is rather striking, given that this link has not been explored previously.

Using this link, we prove an identity that quantifies the Integral Probability Metric (IPM) between the refined diffusion model and data in terms of the gap between the original diffusion model and data, along with an additional quantity that closes the discrepancy based on the choice of discriminator set:

$$\mathrm{IPM}(\hat{P}, \mu_{\mathcal{H}}) = \mathsf{D}(\hat{P}, \mu) - \mathsf{I}_{\mathcal{H}}(\mu, \mu_{\mathcal{H}}), \tag{2}$$

where $\hat{P}$ is the data, μ the original model and $\mu_{\mathcal{H}}$ the refined model. The first term $\mathsf{D}(\hat{P}, \mu)$ can be bounded by many existing lines of work such as in (Chen et al., 2022; Lee et al., 2022; 2023; Li et al., 2023; De Bortoli, 2022), if we consider different assumptions. Thus, our work builds and extends directly upon existing work. We furthermore decompose the term $\mathsf{I}_{\mathcal{H}}$, and show that it increases when $\mathcal{H}$ has more discriminative abilities.

Finally, we use identity to prove a generalization bound that reveals additional insights into how refining a diffusion model can close the generalization gap if discriminators $\mathcal{H}$ are well regularized. Our findings, therefore, advocate for discriminator refinement for deep generative models and validate existing work on discriminator refinement to a great deal of generality. In summary, our technical contributions come in three parts: 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)** A characterization of strong duality for the f-divergence that reveals the optimality of refining generative models. This identity has interest outside the scope of this paper, as illustrated by the connections to variational inference.
▷ **(Theorem 3)** An application to diffusion models, which shows how a general binary classifier played with a proper composite loss can be used to construct a refined diffusion model that admits theoretical convergence. We derive the choice of f such that the refined diffusion corresponds to the framework of (Kim et al., 2022a).
▷ **(Theorem 6)** A study of generalization for refined diffusion models that unveils the importance of using discriminators to close the gap in generalization, along with guidance for the specific choices. This result parallels the discriminator-generalization trade-off other generative models, such as GANs, enjoy.

2 Related Work

We split our related work into two sections; first, we focus on results focused on duality and discriminator studies in Generative Adversarial Networks (GANs), and then we turn to convergence and theory results for score-based diffusion models.

GANs were developed in (Goodfellow et al., 2015) where a binary classification task was used to improve GANs, corresponding to Jensen-Shannon divergence minimization. This result was then generalized to f-divergences in (Nowozin et al., 2016; Nock et al., 2017). (Liu et al., 2017) conducted the first convergence guarantees, showing that the learned distribution is indistinguishable from the data distribution under the chosen set of discriminators. (Liu & Chaudhuri, 2018; Zhang et al., 2017; Husain, 2020) then showed that the generalization abilities of GANs are related to the complexity of the discriminator set. In particular, using a discriminator set that is too large will yield poor generalization, whereas a set too small leads to under-discriminated models, hence a discrimination-generalization tradeoff.

Theoretical guarantees surrounding the convergence of score-based generative models (SGMs) largely consist of bounding the gap between the observed (finitely supported) data distribution and the distribution induced by the discretized diffusion model under various divergences between distributions. The majority of work assumes there exists a parametrized model that can approximate the true score function sufficiently well enough. (Song et al., 2020b) consider the Kullback-Leibler (KL) divergence and use Girsanov's Theorem to bound this quantity for the non-discretized diffusion model. (Lee et al., 2022) then consider the discretized diffusion model and under log-Sobolev inequality (LSI) and smoothness of the true score provide convergence guarantees. Chen et al. (2022) and Lee et al. (2023) prove convergence without the LSI aassumption. Another line of work assumes the score function and parametrized approximation is bounded at each point and proves convergence (De Bortoli et al., 2021), with improved results under the manifold assumption (De Bortoli, 2022).

None of these results, however, consider discriminator-intervened diffusion models such as in (Kim et al., 2022a). Our work shows that the Integral Probability Metric (IPM) distance between the data and refined diffusion model is equal to the discrepancy between the original diffusion model and data, which any of the above results can bound. Therefore, our result extends and builds upon existing convergence guarantees.

3 Preliminaries

Notation We use Ω to denote a compact Polish space and denote Σ as the standard Borel σ-algebra on Ω, $\mathbb{R}$ denotes the real numbers and $\mathbb{N}$ natural 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 random variable X, we use $\mathcal{L}(\mathsf{X})$ to denote the law of X. Let $\mathcal{N}(\mu, \Sigma)$ denote the d-dimensional Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ. For any proposition $\mathscr{I}$, the Iverson bracket is $\llbracket \mathscr{I} \rrbracket = 1$ if $\mathscr{I}$ is true and 0 otherwise. We say a set of functions $\mathcal{H}$ is convex if $\lambda h + (1-\lambda)h' \in \mathcal{H}$ for all $h, h' \in \mathcal{H}$ and $\lambda \in [0,1]$. 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 $\mathcal{H} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, we use $\overline{\mathrm{co}}\,(\mathcal{H})$ to denote the closed convex hull of $\mathcal{H}$ and use $\|\mathcal{H}\| = \sup_{h \in \mathcal{H}} \|h\|_\infty$ as the maximum bound for all functions in $\mathcal{H}$.

Diffusion Models We adapt notation from (Oko et al., 2023). For some time-stamp $T \in \mathbb{N}$, we define $(B_t)_{[0,T]}$ and $\beta_t : [0,T] \to \mathbb{R}$ to denote d-dimensional Brownian motion and a weighting function. The forward process is defined as

$$d\mathsf{X}_t = -\beta_t \mathsf{X}_t dt + \sqrt{2\beta_t} dB_t, \quad \mathsf{X}_0 \sim \hat{P}_N, \tag{3}$$

where $\hat{P}_n = \sum_{j=1}^n \delta_{x_j}$ is the empirical distribution over observed n samples $\{x_i\}_{i=1}^n$. This process is referred to as the Ornstein-Uhlenbeck (OU) process whose transition density corresponds to $\mathsf{X}_t \mid \mathsf{X}_0 \sim \mathcal{N}(m_t \mathsf{X}_0, \sigma_t^2 I_d)$ where $m_t = \exp\left(-\int_0^t \beta_s ds\right)$, and $\sigma_t^2 = 1 - \exp\left(-2\int_0^t \beta_s ds\right)$. Denoting by $P_t = \mathcal{L}(\mathsf{X}_t)$, we can construct the reverse SDE process under mild conditions on $\hat{P}_N$:

$$d\mathsf{Y}_t = \beta_{T-t}\left(d\mathsf{Y}_t + 2\nabla \log p_{T-t}(\mathsf{Y}_t)\right) dt + \sqrt{2\beta_{T-t}} dB_t, \quad d\mathsf{Y}_0 \sim P_T. \tag{4}$$

In practice, since we do not have access to the score function $\log p_{T-t}(Y_t)$, we approximate this with a neural network $s_\vartheta(\mathsf{X}, t)$ for $\mathsf{X} \in \Omega$ and $t \in [0,T]$ and ϑ is a learnable parameter. A Euler-Maruyama scheme is used to discretize the SDE into K different steps $(\tau_k)_{k=1}^K$ where $\tau_0 = 0$ and $\tau_K = T$ and the resulting process is

$$d\hat{\mathsf{Y}}_t = \beta_{T-t}\left(d\hat{\mathsf{Y}}_t + 2 s_\vartheta(\mathsf{X}, T-t)\right) dt + \sqrt{2\beta_{T-t}} dB_t. \tag{5}$$

Generative Adversarial Losses We begin this section by first defining two important divergences between distributions: the f-divergence and Integration Probability Metrics (IPMs). Let $f : \mathbb{R} \to (-\infty, \infty]$ be a convex lower semi-continuous function such that $f(1) = 0$. The f-divergence between two probability measures $\mu, \nu \in \mathscr{P}(\Omega)$ is defined as $\mathsf{I}_f(\mu : \nu) := \mathbb{E}_{X \sim \nu}\left[f\left(\frac{d\mu}{d\nu}(\mathsf{X})\right)\right]$, if $\mu \ll \nu$ (existence of Radon-Nikodym derivative) otherwise $\mathsf{I}_f(\mu : \nu) = +\infty$. The f function is often referred to as the *generator* of I_f. On the other hand, for a given set of functions $\mathcal{H} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, the IPM is defined as

$$d_{\mathcal{H}}(\mu, \nu) = \sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_{\mathsf{X} \sim \mu}[h(\mathsf{X})] - \mathbb{E}_{\mathsf{X} \sim \nu}[h(\mathsf{X})] \right\} \tag{6}$$

The general adversarial loss we consider is that adapted from the generative adversarial networks literature (Liu et al., 2017; Zhang et al., 2017). Let $f : \mathbb{R} \to (-\infty, \infty]$ be a convex lower semi-continuous function such that $f(1) = 0$ and let $\mathcal{H} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ denote a set of bounded and measurable functions. We define a distance between distributions $\mu, \nu \in \mathscr{P}(\Omega)$ to be of the form

$$\mathsf{D}_{f,\mathcal{H}}(\nu, \mu) := \sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_{\mathsf{X} \sim \nu}[h(\mathsf{X})] - \mathbb{E}_{\mathsf{X} \sim \mu}[f^\star \circ h(\mathsf{X})] \right\}, \tag{7}$$

where $f^\star(t) = \sup_{t' \in \operatorname{dom} f} (t \cdot t' - f(t'))$ is the Fenchel conjugate of f. Note that $\mathsf{D}_{f,\mathcal{H}}$ can be viewed as a weaker divergences compared to f-divergences and IPMs noting that $f^\star(t) \geq t$ and so

$$\mathsf{D}_{f,\mathcal{H}} \leq d_{\mathcal{H}} \text{ and } \mathsf{D}_{f,\mathcal{H}} \leq \mathsf{I}_f. \tag{8}$$

Moreover, if $\mathcal{H}$ is chosen large enough then $\mathsf{D}_{f,\mathcal{H}}$ coincides with I_f for any optimal $h^* \in \mathcal{H}$ in equation 7. The divergence $\mathsf{D}_{f,\mathcal{H}}$ has been of great interest in the GAN community most notably appearing as the f-GAN objective (Nowozin et al., 2016), along with other theoretical studies such as in Liu et al. (2017); Liu & Chaudhuri (2018); Husain et al. (2019); Husain (2020).

The divergence $\mathsf{D}_{f,\mathcal{H}}$ is tied to binary classification in Savage's theory of properness (Savage, 1971). Let $\mathcal{Y} = \{-1, 1\}$ and define a joint distribution $\mathbb{P}(\mathsf{X}, \mathsf{Y})$ such that $\mathsf{Y} \in \mathcal{Y}$ with $\mathbb{P}(\mathsf{X} \mid \mathsf{Y} = -1) = \mu(\mathsf{X})$, $\mathbb{P}(\mathsf{X} \mid \mathsf{Y} = +1) = \nu(\mathsf{X})$ and $\mathbb{P}(\mathsf{Y} = -1) = \mathbb{P}(\mathsf{Y} = 1) = 1/2$. We then define a loss function $\ell_f : \mathcal{Y} \times [0, 1] \to \mathbb{R}$, with an invertible link $\Psi : (0, 1) \to \mathbb{R}$ set to be $\Psi(z) := f'(z/(1 - z))$:

$$\ell_f(-1, z) = -f'\left(\frac{\Psi^{-1}(z)}{1 - \Psi^{-1}(z)}\right) \qquad \ell_f(+1, z) = f^\star\left(f'\left(\frac{\Psi^{-1}(z)}{1 - \Psi^{-1}(z)}\right)\right). \tag{9}$$

We then have the following connection

$$\inf_{h \in \mathcal{H}} \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathbb{P}} \left[\ell_f\left(\mathsf{Y}, h(\mathsf{X})\right)\right] = -\frac{1}{2} \mathsf{D}_{f,\mathcal{H}}(\mu, \nu). \tag{10}$$

Losses defined in equation 9 are not any kind of losses: they are proper composite (Nock et al., 2017), *i.e.* they elicit Bayes prediction as an optimal predictor, composite meaning using Ψ to link real-valued prediction to class probabilities. In words, if $\mathcal{H}$ is chosen large enough then $\mathsf{D}_{f,\mathcal{H}}$ is also proportional to the loss of Bayes rule, encoded in some $h^* \in \mathcal{H}$ in equation 10.

4 Strong Duality for Adversarial Losses

In this section, we revisit the primal-dual relationship in GANs and prove a stronger result in the context of denoising diffusion models. Let $\nu \in \mathscr{P}(\Omega)$ denote the data distribution we are interested in learning. First we recall the duality result Liu & Chaudhuri (2018); Husain et al. (2019) where if $\mathcal{H}$ is convex and closed under additive constants then

$$\mathsf{D}_{f,\mathcal{H}}(\nu, \mu) = \inf_{\overline{\mu} \in \mathscr{P}(\Omega)} \left(d_{\mathcal{H}}(\nu, \overline{\mu}) + \mathsf{I}_f(\overline{\mu} : \mu) \right). \tag{11}$$

Under various settings, the optimization over $\overline{\mu}$ was shown to coincide with the optimization over an encoder function in Husain et al. (2019) where the $d_{\mathcal{H}}$ is a reconstruction loss, and I_f is a regularizer however there is no characterization of $\overline{\mu}$ beyond this. We now present the first result in this direction.

Theorem 1 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a strictly convex lower semi-continuous and differentiable function with $f(1) = 0$ and let $\mathcal{H}$ be convex and closed under addition. Denote by $h^* \in \mathcal{H}$ as the optimal solution of the primal problem in equation 11. Assuming $f'^{-1}(t) \geq 0$ for all $t \in \operatorname{dom}(f^\star)$, consider a distribution $\mu^{\mathcal{H}}$ whose Radon-Nikodym derivative with respect to μ is*

$$\frac{d\mu^{\mathcal{H}}}{d\mu}(\cdot) = f'^{-1}\left(h^*(\cdot)\right). \tag{12}$$

Then $\mu^{\mathcal{H}}$ is an optimal solution to equation 11.

When the distribution μ admits a Lebesgue density, we can write $\mu^{\mathcal{H}} = \mu \cdot f'^{-1}(h^*)$. Moreover, in order to sample from $\mu^{\mathcal{H}}$, one can derive the score function, which by virtue of the above Theorem is

$$\nabla \log \mu^{\mathcal{H}} = \nabla \log \mu + \nabla \log \left(f'^{-1} \left(h^*(\cdot) \right) \right). \tag{13}$$

Since in diffusion models, we approximate the score function $\nabla \log \mu$, the extra term $\nabla \log \left(f'^{-1} \left(h^*(\cdot) \right) \right)$ can be considered the *refinement*. Note that this refinement term must first require finding h^* in the optimization of $\mathsf{D}_{f,\mathcal{H}}$, which is equivalent to a binary classification problem. We thus refer to this refined distribution $\mu^{\mathcal{H}}$ as the *discriminator-guided* diffusion model.

While we aim to apply this Theorem to discriminator-guided diffusion models, it boasts a more considerable generality beyond this application. For example, in a somewhat trivialized setting where $\mathcal{H}$ is selected to be minimal, this Theorem recovers results of Variational Inference, which are of independent interest.

Example 1 (Variational Inference) *In the setting of $f(t) = t \log t$, the duality in equation 11 recovers the Evidence Lower Bound (ELBO) that commonly appears in Variational Inference (VI) (Knoblauch et al., 2019; Husain & Knoblauch, 2022). To see this, we have a data-dependent loss function $L : \Theta \to \mathbb{R}$ where Θ is the parameter space of models, then we construct $\mathcal{H} = \{-L + b : b \in \mathbb{R}\}$ as the minimal set satisfying convexity and closure under addition. Noting that $f^\star(t) = \exp(t-1)$ for the choice of KL-divergence, the duality then becomes*

$$\sup_{b \in \mathbb{R}} \left(\mathbb{E}_\nu[-L] + b - \mathbb{E}_\mu[\exp\left(-L + b - 1\right)] \right) = \inf_{\overline{\mu}} \left(\mathbb{E}_\nu[-L] - \mathbb{E}_{\overline{\mu}}[-L] + \mathrm{KL}(\overline{\mu}, \nu) \right) \tag{14}$$

$$\implies \sup_{b \in \mathbb{R}} \left(b - \mathbb{E}_\mu[\exp\left(-L + b - 1\right)] \right) = \inf_{\overline{\mu}} \left(\mathbb{E}_{\overline{\mu}}[L] + \mathrm{KL}(\overline{\mu}, \nu) \right) \tag{15}$$

$$\implies \log \mathbb{E}_\mu \left[\exp\left(-L\right) \right] = \inf_{\overline{\mu}} \left(\mathbb{E}_{\overline{\mu}}[L] + \mathrm{KL}(\overline{\mu}, \nu) \right), \tag{16}$$

which is precisely the ELBO as presented in VI where $\overline{\mu}$ is referred to as the posterior and μ is the prior distribution. When applying Theorem 1, we note that $h^* = -L - \log\left(\mathbb{E}_\mu \left[\exp\left(-L-1\right) \right] \right)$ (see Appendix for more details) then the optimal distribution $\mu^{\mathcal{H}}$ as per Theorem 1 is

$$\begin{aligned} \mu^{\mathcal{H}} &= \exp\left(-L - \log\left(\mathbb{E}_\mu \left[\exp\left(-L-1\right) \right] \right) - 1 \right) \cdot \mu & (17) \\ &= \mu \cdot \exp\left(-L\right) / \mathbb{E}_\mu[-L], & (18) \end{aligned}$$

which is exactly the generalized Bayesian posterior.

In the above, the optimal h^* amounts to solving for a constant b due to the choice of $\mathcal{H}$. If we select $\mathcal{H}$ to be a set of parametrized discriminators then with $f(t) = t \log t$, the refined distribution $\mu^{\mathcal{H}}$ simplifies to

$$\mu^{\mathcal{H}} = \mu \cdot \exp(h^*) / \mathbb{E}_\mu[h^*], \tag{19}$$

which corresponds to an exponential family whose base measure is μ and sufficient statistic h^* (the cumulant is defined from the denominator). It should be noted that a related work (Cranko & Nock, 2019) boosts densities via discriminators using the form of equation 19. We now move onto the case where picking f corresponds to h^* being the cross-entropy minimized discriminator, deriving exactly the framework of Kim et al. (2022a).

Example 2 (Refined Generative Models (Kim et al., 2022a)) *In the setting of $f(t) = t \log t - (t+1)\log(t+1) + 2\log 2$, if we define $\mathcal{H} = \{\log \eta_\theta : \eta_\theta : \Omega \to [0,1], \theta \in \Theta\}$, where η_θ is a parametrized model giving a softmax score such a neural network. The duality then becomes*

$$\mathsf{D}_{f,\mathcal{H}}(\nu, \mu) = 2\log 2 - \inf_{\theta \in \Theta} \left(\mathbb{E}_\nu[-\log\left(\eta_\theta\right)] + \mathbb{E}_\mu[-\log\left(1 - \eta_\theta\right)] \right), \tag{20}$$

which corresponds to binary cross-entropy minimization. Considering that $f'^{-1}(t) = \frac{\exp t}{1 - \exp t}$, the refined distribution $\mu^{\mathcal{H}}$ simplifies to

$$\frac{d\mu^{\mathcal{H}}}{d\mu} = \frac{\eta_{\theta^*}}{1 - \eta_{\theta^*}}, \tag{21}$$

where θ^* is the optimal parameter from equation 20.

It is rather striking that the work of Kim et al. (2022a) finds exactly the distribution on which strong duality occurs. In the next section, we revisit this example with the notation of diffusion models, elucidating the link. We summarize the general finding in the form of identity in this Theorem.

Theorem 2 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a strictly convex lower semi-continuous and differentiable function with $f(1) = 0$. Denote by $h^* \in \mathcal{H}$ as the optimal solution to equation 11, $\mu^{\mathcal{H}}$ defined in equation 12 . If $f'^{-1}(t) \geq 0$ for all $t \in \mathrm{dom}(f^\star)$ then we have*

$$d_{\mathcal{H}}\left(\nu, \mu^{\mathcal{H}}\right) = \mathsf{D}_{f,\mathcal{H}}\left(\nu, \mu\right) - \mathsf{I}_f\left(\mu^{\mathcal{H}} : \mu\right). \tag{22}$$

5 Convergence and Generalization Bounds for Diffusion

In the context of diffusion models, the distribution $\nu = \hat{P}_0 = \sum_{i=1}^n \delta_{x_i}$ where $\{x_i\}_{i=1}^n$ is the i.i.d data $x_i \sim P$ for some population distribution $P \in \mathscr{P}(\Omega)$. As defined in Section 3, we denote by $\hat{P}_t = \mathcal{L}(\mathsf{X}_t)$ as the law of the SDE followed by a simple Ornstein-Uhlerbeck process. For a given function $s_\vartheta : \Omega \times \mathbb{R} \to \Omega$ that approximates the score function, and a discretization scheme $0 = \tau_0, \ldots, \tau_{K-1} = T$, we define the process with $k = 0, \ldots, K$ and $t \in [\tau_k, \tau_{k+1}]$:

$$\mathsf{Y}_t^{\mathcal{H},\vartheta} = d\mathsf{Y}_t^{\mathcal{H},\vartheta} + 2s_\vartheta\left(\mathsf{Y}_t^{\mathcal{H},\vartheta}, \tau_k\right) + 2\nabla\log\left(f'^{-1}\left(h^*_{\tau_{k+1}}\left(\mathsf{Y}_t^{\mathcal{H},\vartheta}\right)\right)\right) + \sqrt{2}dB_t, \tag{23}$$

$$\mathsf{Y}_0^{\mathcal{H},\vartheta} \sim \gamma_d \tag{24}$$

where γ_d is a prior distribution, typically taken to be a d-dimensional Gaussian distribution and h_t^* is the corresponding optimal discriminator from Theorem 2 when $\nu = \hat{P}_{T-t}$ and $\mu = \mathcal{L}(\mathsf{Y}_t^\vartheta)$. The function s_ϑ is a neural network approximating the score function. In the setting of $f(t) = t\log t - (t+1)\log(t+1) + 2\log 2$, this is the process used to generate samples from discriminator-guided denoised diffusion model (Kim et al., 2022a). We therefore denote the model distribution as $\mu_{T,\vartheta}^{\mathcal{H}} = \mathcal{L}\left(\mathsf{Y}_T^{\mathcal{H},\vartheta}\right)$. We are now ready to present Theorem 2 in the context of denoised diffusion models.

Theorem 3 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a strictly convex lower semi-continuous and differentiable function with $f(1) = 0$. If $f'^{-1}(t) \geq 0$ for all $t \in \mathrm{dom}(f^\star)$ then it holds*

$$d_{\mathcal{H}}\left(\hat{P}_0, \mu_{T,\vartheta}^{\mathcal{H}}\right) = \mathsf{D}_{f,\mathcal{H}}\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right) - \mathsf{I}_f\left(\mu_{T,\vartheta}^{\mathcal{H}} : \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right). \tag{25}$$

This identity tells us that the gap (in IPM) between the refined model $\mu_{T,\vartheta}^{\mathcal{H}}$ and data distribution $\hat{P}_0$ is precisely equal to the gap between the original SGM model without refinement $\mathsf{D}_{f,\mathcal{H}}\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right)$ minus the difference in refinement $\mathsf{I}_f\left(\mu_{T,\vartheta}^{\mathcal{H}} : \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right)$. We split the next part of this section into discussing these two terms.

First, note that $\mathsf{D}_{f,\mathcal{H}}\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right)$ can readily be upper bounded in various ways using recent results with competing assumptions. This is attributed to the fact that $\mathsf{D}_{f,\mathcal{H}}$ lower bounds both the f-divergence I_f and the IPM $d_{\mathcal{H}}$ (8). Therefore, we can inherit results for the reverse-KL divergence ($f(t) = -\log t$) (Lee et al., 2023) or Total Variation ($f(t) = |t-1|$ (Chen et al., 2022), both showing fast rates of convergence. Alternatively, we can upper bound via the IPMs and take $\mathsf{D}_{f,\mathcal{H}}\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right) \leq \sup_{h\in\mathcal{H}} \mathrm{Lip}(h) \cdot W_1\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right)$ where $\mathrm{Lip}(h)$ is the Lipschitz constant of h and W_1 is the 1-Wasserstein distance. In this setting, (Oko et al., 2023) shows that μ approximates ν at a minimax optimal rate for the 1-Wasserstein distance, however the term $\sup_{h\in\mathcal{H}} \mathrm{Lip}(h)$ reveals an interesting trade-off on the complexity of $\mathcal{H}$.

We now present an additional analysis by taking the assumptions of Chen et al. (2022), which are considered to be minimal:

Assumption 1 *For all $t \geq 0$, $\mathsf{X} \mapsto \nabla\log\hat{P}_t(\mathsf{X})$ is L-Lipschitz.*

Assumption 2 *The second moment of $\hat{P}_0$ is bounded: $m_2^2 := \mathbb{E}_{\mathsf{X}\sim\hat{P}_0}\left[\|\mathsf{X}\|^2\right] < \infty$.*

Assumption 3 *for all $k = 0, \dots, K$, there exists a constant $\varepsilon_\vartheta > 0$ such that*

$$\mathbb{E}_{\mathsf{X} \sim \hat{P}_{\tau_k}} \left[\left\| s_\vartheta(\mathsf{X}, \tau_k) - \nabla \log \hat{P}_{\tau_k}(\mathsf{X}) \right\|^2 \right] \leq \varepsilon_\vartheta^2 \tag{26}$$

Under these assumptions, we have the following bound.

Theorem 4 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a strictly convex lower semi-continuous and differentiable function with $f(1) = 0$. Denote by $\Delta_T : \Omega \to \mathscr{P}(\Omega)$ as the forward diffusion process after T steps and $\#$ the push-forward operator. Assume that $\|\mathcal{H}\| := \sup_{h \in \mathcal{H}} \|h\|_\infty < \infty$ and $\Omega = \mathbb{R}^d$. For $k = 1, \dots, K$, set $\tau_k = kT/K$ (with $\tau_0 = 0$) with $s = T/K$ as the stepsize, we have under Assumptions 1-3*

$$\mathsf{D}_{f,\mathcal{H}}\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right) \leq \|\mathcal{H}\| \cdot \left(1 - \exp\left(-\left(\varepsilon_\vartheta^2 + L^2 ds + L^2 m_2^2 s^2\right) T\right)\right) + \mathsf{I}_f\left(\Delta_{T\#}\hat{P}_0 : \gamma_d\right). \tag{27}$$

This analysis follows the standard Girsanov's change of measure result utilized in (Chen et al., 2022; Oko et al., 2023) where the first term corresponds to the error incurred by approximating the score function ε_ϑ and discretization error appearing with the key difference here being the term $\mathsf{I}_f\left(\Delta_{T\#}\hat{P}_0 : \gamma_d\right)$ which appears due to the general choice of f. In particular, when $f(t) = t \log t$ is chosen to correspond to the KL divergence, then this term decays exponentially with T by the exponential convergence of Uhlerbeck-Ornstein processes studied in (Bakry et al., 2014, Theorem 5.2.1). Moreover, if f is chosen such that there exists some strictly increasing function $\psi_f : \mathbb{R} \to \mathbb{R}$ with $\mathsf{I}_f \leq \psi_f(\mathrm{KL})$ then we can guarantee convergence trivially by connecting the exponential convergence in KL divergence. In fact, we show in Lemma 4 that when $f(t) = t \log t - (t+1)\log(t+1) + 2\log 2$ then we directly have $\mathsf{I}_f \leq \mathrm{KL}$ giving us the required convergence. We show that we can relate f-divergences to the KL divergence.

Lemma 1 *Suppose $f : \mathbb{R} \to (-\infty, \infty]$ is a strictly convex differentiable lower semi-continuous function with $f(1) = 0$ and denote by $\Delta_T : \Omega \to \mathscr{P}(\Omega)$ as the forward diffusion process after T steps. Let $r_T = d(\Delta_{T\#}\hat{P}_0)/d\gamma_d$ be the Radon-Nikodym derivative then we have $\mathsf{I}_f\left(\Delta_{T\#}\hat{P}_0 : \gamma_d\right) \leq \sup_{\mathsf{X} \in \Omega} |f'(r_T(\mathsf{X}))| \cdot \sqrt{\mathrm{KL}\left(\Delta_{\mathrm{T}\#}\hat{\mathrm{P}}_0 \colon \gamma_{\mathrm{d}}\right)}$.*

In this result, we can thus see that convergence of the forward diffusion process in I_f is as fast as the KL divergence except for an extra term that depends on the density ratio r_T. Indeed, we assume that as T gets large, then $r_T \to 1$. We now move onto the second refinement term from Theorem 3: $\mathsf{I}_f\left(\mu_{T,\vartheta}^{\mathcal{H}} : \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right)$. Note that this quantity is subtracted from the total gap and is positive. In order to better understand this gain, we reparametrize it into a 1-dimensional integral: suppose the classifier decomposes with the natural composite link: $h^* = f'(\eta_{h^*}/(1 - \eta_{h^*}))$ where $\eta_{h^*} : \Omega \to [0, 1]$ is the class probability estimate.

Theorem 5 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a strictly convex lower semi-continuous and differentiable function with $f(1) = 0$. Denote by $h^* \in \mathcal{H}$ as the optimal discriminator and $\mu_{T,\vartheta}^{\mathcal{H}}$ such that $d\mu_{T,\vartheta}^{\mathcal{H}}/d\mu = f'^{-1}(h^*)$ be the optimal solutions to equation 11 . If $f'^{-1}(t) \geq 0$ for all $t \in \mathrm{dom}(f^\star)$ then we have*

$$\mathsf{I}_f\left(\mu_{T,\vartheta}^{\mathcal{H}} : \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)\right) = \int_0^1 f\left(\frac{t}{1-t}\right) d\rho_{h^*}(t) \tag{28}$$

where $\mu_T^\vartheta = \mathcal{L}\left(\mathsf{Y}_T^\vartheta\right)$ and η_{h^*} is such that $h^* = f'(\eta_{h^*}/(1 - \eta_{h^*}))$ and $\rho_{h^*} := \eta_{h^*\#}\mu_T^\vartheta$.

In order to understand this quantity, note that ρ_{h^*} corresponds to the regions in $[0, 1]$ where the classifier predicts μ_T^ϑ to be. Let us consider the case of binary-cross entropy corresponding to (Kim et al., 2022a) where $f(t) = t \log t - (t+1)\log(t+1) + 2\log 2$. The expression then simplifies to

$$\int_0^1 \left(\frac{t}{1-t} \log t + \log(1-t) + 2\log 2\right) d\rho_{h^*}(t). \tag{29}$$

Recalling that $\eta_{h^*}(\mathsf{X})$ corresponds to the class probability estimation of a point $\mathsf{X} \in \Omega$ belonging to the class $\hat{P}_0$ (as opposed to μ^{ϑ}_T), we can say that if $\mathcal{H}$ is rich enough to classify between $\hat{P}_0$ and μ^{ϑ}_T then η_{h^*} is close to 0 on the support of μ^{ϑ}_T and thus $\rho_{h^*}(t)$ is concentrated around 0, thereby increasing the above expression to $f(0) = 2\log 2$, the maximal possible value for $\mathsf{I}_f\left(\mu^{\mathcal{H}}_{T,\vartheta} : \mathcal{L}\left(\mathsf{Y}^{\vartheta}_T\right)\right)$. On the other hand, if the best classifier η_{h^*} is unable to classify the two classes, then $\eta_{h^*} \to 1/2$ (a random classifier) and thus the above expression tends to $f(1) = 0$. Therefore, the gain in refinement can be understood by the discriminative abilities of $\mathcal{H}$, and a richer choice for $\mathcal{H}$ will lead to more improvement. We will see in what follows how this forms a trade-off with generalization.

In order to proceed, we first define an important quantity that often counters the discriminative abilities of $\mathcal{H}$: the Rademacher complexity. Let R denote the uniform distribution over $\{-1, +1\}$ then for a class of functions $\mathcal{H} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ and distribution P, the *Rademacher* complexity (Bartlett & Mendelson, 2002) is

$$\mathscr{R}_n(\mathcal{H}) := \mathbb{E}_{\zeta \sim R^n, \mathsf{X} \sim P^n}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \zeta_i h(\mathsf{X}_i)\right], \tag{30}$$

where R^n and P^n are distributions over n-tuples. The Rademacher complexity is a cornerstone in many generalization studies, specifically dominating supervised learning; however, it has also appeared in unsupervised learning domains and generative models such as in (Zhang et al., 2017). We refer the reader to (Liang, 2016) for a more comprehensive analysis of this quantity for different choices of model complexity. We now link the Rademacher complexity of our discriminator set used for refinement $\mathcal{H}$ to the gap between the refined diffusion model and data distribution.

Theorem 6 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a strictly convex lower semi-continuous and differentiable function with $f(1) = 0$. Suppose $f'^{-1}(t) \geq 0$ for all $t \in \operatorname{dom}(f^\star)$ and let P denote the population distribution of $\hat{P}_0$ such that each $x_i \sim P$ i.i.d. It then holds that*

$$d_{\mathcal{H}}\left(P, \mu^{\mathcal{H}}_{T,\vartheta}\right) \leq \mathsf{D}_{f,\mathcal{H}}\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}^{\vartheta}_T\right)\right) - \mathsf{I}_f\left(\mu^{\mathcal{H}}_{T,\vartheta} : \mathcal{L}\left(\mathsf{Y}^{\vartheta}_T\right)\right) + \mathscr{R}_n(\mathcal{H}) + 2\|\mathcal{H}\| \cdot \sqrt{\frac{1}{2n}\ln\left(\frac{1}{\delta}\right)}, \tag{31}$$

with probability at least $1 - \delta$.

In light of the above results, it is clear that the choice of $\mathcal{H}$ is important. On the one hand, we should pick $\mathcal{H}$ to be as large as possible since the IPM on the left-hand side will be more discriminative as a discrepancy between P and $\mu^{\mathcal{H}}_{T,\vartheta}$. Additionally, it is seen from Theorem 5, convergence and overfitting to $\hat{P}_0$ is encouraged with more discriminative choices of $\mathcal{H}$. On the other hand, we see that we pay a price of the Rademacher complexity $\mathcal{R}_n(\mathcal{H})$, which along with Theorem 3, suggests a well-regularized choice of $\mathcal{H}$. This discussion parallels the discrimination-generalization trade-off previously studied in the context of GANs, such as in Zhang et al. (2017); Husain (2020). Note that for the left-hand side to be discriminative enough, we only require a choice of $\mathcal{H}$ that is dense in the space $\mathscr{F}(\Omega, \mathbb{R})$ while also being able to control $\mathscr{R}_n(\mathcal{H})$. An example choice of $\mathcal{H}$ is the set of 1-Lipschitz functions where $\mathcal{R}_n(\mathcal{H}) = O(n^{1/d})$.

In addition to the above, the term $\mathsf{D}_{f,\mathcal{H}}\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}^{\vartheta}_T\right)\right)$ is indeed the optimal (eventually Bayes) risk classification loss under the proper composite loss derived by f as mentioned in the preliminaries. Thus, the generalization bound itself resembles that of classical learning theory where the empirical classification risk appears in the upper bound (Bartlett & Mendelson, 2002).

6 Conclusion

We study discriminator-guided diffusion models with a prescribed discriminator set $\mathcal{H}$. In order to prove our result, we revisit the primal-dual link in GANs and extend results in that direction, characterizing the exact distribution under which strong duality holds. This result allows us to quantify exactly the Integral Probability Metric (IPM) between the data and discriminator-guided diffusion model distribution, leading us to characterize the generalization abilities of such models.

In particular, our results advocate using regularized discriminators and refining diffusion models to close the gap in generalization.

Some ways to extend this work include a tighter analysis for the term $\mathsf{D}_{f,\mathcal{H}}\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}_T^{\vartheta}\right)\right)$ which we bounded by either the IPM or f-divergence to utilize existing results. This term, however, is smaller and weaker than both these divergences; therefore, a tighter analysis can reveal a faster convergence rate. Additionally, since our framework recovers the binary cross entropy refinement from Kim et al. (2022a), we can derive loss functions beyond this case. Since our goal was to prove the link to strong duality, we leave the implementation of refined diffusion models with general proper composite losses as the subject for future work

REFERENCES

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

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.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. *Introduction to Statistical Learning Theory*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_8. URL `https://doi.org/10.1007/978-3-540-28650-9_8`.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:119–137, 1979.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2022.

Zac Cranko and Richard Nock. Boosted density estimation remastered. In *International Conference on Machine Learning*, pp. 1416–1425. PMLR, 2019.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

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

Wenqi Fan, Chengyi Liu, Yunqing Liu, Jiatong Li, Hang Li, Hui Liu, Jiliang Tang, and Qing Li. Generative diffusion models on graphs: Methods and applications. *arXiv preprint arXiv:2302.02591*, 2023.

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*, pp. 2672–2680, 2014.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hisham Husain. Distributional Robustness with IPMs and links to Regularization and GANs. *Advances in Neural Information Processing Systems*, 33, 2020.

Hisham Husain and Jeremias Knoblauch. Adversarial interpretation of bayesian inference. In *International Conference on Algorithmic Learning Theory*, pp. 553–572. PMLR, 2022.

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

Hisham Husain, Borja Balle, Zac Cranko, and Richard Nock. Local differential privacy for sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 3404–3413. PMLR, 2020.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Dongjun Kim, Yeongmin Kim, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022a.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022b.

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

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pp. 946–985. PMLR, 2023.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

Percy Liang. Cs229t/stat231: Statistical learning theory (winter 2016), 2016.

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

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

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.

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*, pp. 456–464, 2017.

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

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*, 2023.

Santiago Pascual, Gautam Bhattacharya, Chunghsin Yeh, Jordi Pons, and Joan Serrà. Full-band general audio synthesis with score-based diffusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pp. 8599–8608. PMLR, 2021.

Chunyu Qiang, Hao Li, Hao Ni, He Qu, Ruibo Fu, Tao Wang, Longbiao Wang, and Jianwu Dang. Minimally-supervised speech synthesis with conditional diffusion model and language model: A comparative study of semantic coding. *arXiv preprint arXiv:2307.15484*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, Tal Peer, and Timo Gerkmann. Speech signal improvement using causal generative diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2. IEEE, 2023.

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.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

L.-J. Savage. Elicitation of personal probabilities and expectations. pp. 783–801, 1971.

Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065*, 2022.

Alexander Soen, Hisham Husain, and Richard Nock. Data preprocessing to mitigate bias with boosted fair mollifiers. *arXiv preprint arXiv:2012.00188*, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Xianchao Wu. Duplex diffusion models improve speech-to-speech translation. *arXiv preprint arXiv:2305.12628*, 2023.

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

Mengchun Zhang, Maryam Qamar, Taegoo Kang, Yuna Jung, Chenshuang Zhang, Sung-Ho Bae, and Chaoning Zhang. A survey on graph diffusion models: Generative ai in science for molecule, protein and material. *arXiv preprint arXiv:2304.01565*, 2023.

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

A Appendix

A.1 Notation

We will be invoking general convex analysis on the space $\mathscr{F}(\Omega, \mathbb{R})$, in the same fashion as Liu & Chaudhuri (2018), 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 Rockafellar (1970). We introduce the conjugate specific to these spaces

Definition 1 (Rockafellar (1968)) *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 7 (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 $F^{\star\star} = F$.*

We also define the Frechet normal cone (or prenormal cone) of a prescribed set $\mathcal{H} \subseteq \mathscr{F}(\Omega, \mathbb{R})$ is

$$\mathsf{N}_{\mathcal{H}}(h) := \left\{ \mu' \in \mathscr{B}(\Omega) : \limsup_{\mathcal{H} \subseteq (h') \to h} \frac{\langle \mu', h' - h \rangle}{\|h' - h\|} \leq 0 \right\}, \tag{32}$$

where the $\langle \cdot, \cdot \rangle$ is the operator linking the two dual spaces. in this case of $\mathscr{F}(\Omega, \mathbb{R})$ and $\mathscr{B}(\Omega)$ is $\langle h, \mu \rangle = \mathbb{E}_{\mathsf{X} \sim \mu}[h(\mathsf{X})]$

A.2 Proof of Theorem 1

We begin by defining the objective

$$\mathscr{L}(h, Q) = \mathbb{E}_\nu[h] - \mathbb{E}_Q[h] + \mathsf{I}_f(Q : \mu). \tag{33}$$

Note that since $\mathscr{L}$ is upper semicontinuous concave in h and lower semicontinuous convex in Q, we have by Ky Fan's minimax Theorem (Fan, 1953) and (Liu & Chaudhuri, 2018, Lemma 27):

$$\sup_{h \in \mathcal{H}} \inf_{Q \in \mathscr{P}(\Omega)} \mathscr{L}(h, Q) = \inf_{Q \in \mathscr{P}(\Omega)} \sup_{h \in \mathcal{H}} \mathscr{L}(h, Q). \tag{34}$$

Lemma 2 *For any function $h \in \mathscr{F}(\Omega, \mathbb{R})$, we have*

$$\mathscr{L}(h, \mu_h) = \inf_{Q \in \mathscr{P}(\Omega)} \mathscr{L}(h, Q), \tag{35}$$

where μ_h is a probability measure such that the Radon-Nikodym derivative satisfies

$$\frac{d\mu_h}{d\mu} = f'^{-1}(h - \lambda_h), \tag{36}$$

where λ_h is a constant such that $\mathbb{E}_\mu[f'^{-1}(h - \lambda_h)] = 1$.

Proof First note Q must be absolutely continuous with respect to μ since it is a requirement of I_f being finite. We can therefore re-parametrize Q as $Q = r \cdot \mu$ where $\mathbb{E}_\mu[r] = 1$. We also require

$r \geq 0$ however for now, we will show our assumptions on f can alleviate this. The optimization can thus be rewritten as

$$\begin{aligned} & \mathbb{E}_\nu[h] - \mathbb{E}_Q[h] + \mathsf{I}_f(Q : \mu) && (37) \\ &= \mathbb{E}_\nu[h] - \mathbb{E}_\mu\left[r \cdot h\right] + \mathbb{E}_\mu[f(r)] + \lambda(\mathbb{E}_\mu[r] - 1) && (38) \\ &= \mathbb{E}_\nu[h] + \mathbb{E}_\mu\left[-r \cdot h + f(r) + \lambda(r-1)\right]. && (39) \end{aligned}$$

We differentiate the objective w.r.t r, and set the derivative to 0, which yields:

$$\begin{aligned} 0 &= -h^* + f'(r) + \lambda && (40) \\ \implies r &= f'^{-1}\left(h^* - \lambda\right). && (41) \end{aligned}$$

Differentiating with respect to λ yields $\mathbb{E}_\mu[r] = 1$ which can be satisfied if we set $\lambda = \overline{\lambda}$. Next, note that the assumption $f'^{-1}(t) \geq 0$ guarantees this solution satisfies $r \geq 0$. Thus, the optimal density satisfies

$$\frac{d\mu_h}{d\mu} = f'^{-1}(h - \lambda_h). \tag{42}$$

■

Note that

$$\begin{aligned} \mathscr{L}(h, \mu_h) &= \mathbb{E}_\nu[h] - \mathbb{E}_Q[h] + \mathsf{I}_f(Q : \mu) && (43) \\ &= \mathbb{E}_\nu[h] - \mathbb{E}_\mu\left[h \cdot \left(f'^{-1}(h - \lambda_h)\right)\right] + \mathbb{E}_\mu\left[f\left(f'^{-1}(h - \lambda_h)\right)\right] && (44) \\ &= \mathbb{E}_\nu[h - \lambda_h] - \mathbb{E}_\mu\left[(h - \lambda_h) \cdot \left(f'^{-1}(h - \lambda_h)\right)\right] + \mathbb{E}_\mu\left[f\left(f'^{-1}(h - \lambda_h)\right)\right] && (45) \\ &= \mathbb{E}_\nu\left[h - \lambda_h\right] - \mathbb{E}_\mu\left[f^\star \circ (h - \lambda_h)\right]. && (46) \end{aligned}$$

If we denote by $\mathsf{R}(h) = \mathbb{E}_\nu[h] - \mathbb{E}_\mu[f^\star \circ h]$, we are able to write

$$\begin{aligned} \sup_{h \in \mathcal{H}} \mathsf{R}(h - \lambda_h) &= \sup_{h \in \mathcal{H}} \mathscr{L}(h, \mu_h) && (47) \\ &= \sup_{h \in \mathcal{H}} \inf_{Q \in \mathscr{P}(\Omega)} \mathscr{L}(h, Q) && (48) \\ &= \sup_{h \in \mathcal{H}} \left(\mathbb{E}_\nu[h] - \mathbb{E}_\mu[f^\star \circ h]\right) && (49) \\ &= \sup_{h \in \mathcal{H}} \mathsf{R}(h). && (50) \end{aligned}$$

Hence the optimal solution h^* specified in the theorem statement can be decomposed into the form $h^* = \tilde{h} + \lambda_{\tilde{h}}$ for some $\tilde{h} \in \mathcal{H}$. Now we will show that $\mu_{\tilde{h}}$ is in the optimal solution to

$$\begin{aligned} \mathcal{R}(Q) &:= d_{\mathcal{H}}(\nu, Q) + \mathsf{I}_f(Q : \mu) && (51) \\ &= \sup_{h \in \mathcal{H}} \mathscr{L}(h, Q) && (52) \end{aligned}$$

In order to proceed, we will define an auxillary objective:

$$\mathsf{J}(h) = \mathbb{E}_\nu[h] - \mathbb{E}_{\mu_{\tilde{h}}}[h] + \delta_{\mathcal{H}}(h). \tag{53}$$

This objective corresponds to the inner IPM term inside $\mathcal{R}$. We require the following lemma to aid us in decomposing $\mathcal{R}(\mu_{\tilde{h}})$.

Lemma 3 *The function $h^* \in \mathcal{H}$ maximizes $\mathsf{J}(h)$.*

Proof Note that by definition, h^* maximizes the functional

$$\mathsf{J}_f(h) = \mathbb{E}_\nu[h] - \mathbb{E}_\nu[f^\star \circ h] + \delta_{\mathcal{H}}(h). \tag{54}$$

Thus, if we take the subgradient of $-\mathsf{J}_f$ and apply (Penot, 2012, Theorem 2.97), the following holds via optimality of h^*

$$0 \in -d\nu + (f^\star)'(h^*)d\mu + \mathsf{N}_{\mathcal{H}}(h^*), \tag{55}$$

where $\overline{h^*} := h^* + \overline{\lambda}$. We apply the same result to $-\mathsf{J}$ to yield a condition on the optimizer h:

$$\begin{aligned}
&h \text{ minimizes } -\mathsf{J}(h) && (56)\\
&\iff 0 \in -d\nu + d\mu_{\tilde{h}} + \mathsf{N}_{\mathcal{H}}(h) && (57)\\
&\iff 0 \in -d\nu + d\mu f'^{-1}(h^*) + \mathsf{N}_{\mathcal{H}}(h) && (58)\\
&\overset{(1)}{\iff} 0 \in -d\nu + (f^{\star})'(h^*)d\mu + \mathsf{N}_{\mathcal{H}}(h), && (59)
\end{aligned}$$

where (1) is due to the fact when f is strictly convex we have $(f^{\star})' = (f')^{-1}$. Finally, note that if we set $h = h^*$, the condition is met by optimality condition specified in equation 55. ■

Putting all the above together, we are able to write

$$\begin{aligned}
\inf_{Q\in\mathscr{P}(\Omega)} \mathcal{R}(Q) &\leq \mathcal{R}(\mu_{\tilde{h}}) && (60)\\
&= d_{\mathcal{H}}(\nu, \mu_{\tilde{h}}) + \mathsf{I}_f(\mu_{\tilde{h}} : \mu) && (61)\\
&= \sup_{h\in\mathscr{F}(\Omega)} \mathsf{J}(h) + \mathsf{I}_f(\mu_{\tilde{h}} : \mu) && (62)\\
&= \sup_{h\in\mathscr{F}(\Omega)} \mathsf{J}(h) + \mathbb{E}_{\mu}\left[f\left(f'^{-1}\left(h^*\right)\right)\right] && (63)\\
&\overset{(1)}{=} \mathsf{J}(h^*) + \mathbb{E}_{\mu}\left[f\left(f'^{-1}\left(h^*\right)\right)\right] && (64)\\
&= \mathbb{E}_{\nu}[h^*] - \mathbb{E}_{\mu}[f^{-1}(h^*)h^*] + \mathbb{E}_{\mu}\left[f\left(f'^{-1}\left(h^*\right)\right)\right] && (65)\\
&= \mathbb{E}_{\nu}[h^*] - \mathbb{E}_{\mu}\left[f^{-1}(h^*)h^* - f\left(f'^{-1}\left(h^*\right)\right)\right] && (66)\\
&\overset{(2)}{=} \mathbb{E}_{\nu}[h^*] - \mathbb{E}_{\mu}\left[f^{\star}\circ h^*\right] && (67)\\
&\leq \sup_{h'\in\mathcal{H}} \left(\mathbb{E}_{\nu}[h'] - \mathbb{E}_{\mu}\left[f^{\star}\left(h'\right)\right]\right) && (68)\\
&\overset{(3)}{=} \inf_{Q\in\mathscr{P}(\Omega)} \mathcal{R}(Q), && (69)
\end{aligned}$$

where (1) is via the optimality of h^* via Lemma 3, (2) is by definition of $f^{\star}$, and (3) is via primal-duality as specified in equation 11. Finally, note that $\mu_{\tilde{h}} = \mu^{\mathcal{H}}$ by the construction of $\tilde{h}$.

A.3 Proof of Theorem 2

Since $\mu_{\tilde{h}}$ achieves the strong duality, we have equality:

$$\begin{aligned}
&\mathsf{D}_{f,\mathcal{H}} = d_{\mathcal{H}}(\nu, \mu_{\tilde{h}}) + \mathsf{I}_f(\mu_{\tilde{h}} : \mu) && (70)\\
&\implies d_{\mathcal{H}}\left(\nu, \mu_{\tilde{h}}\right) = \mathsf{D}_{f,\mathcal{H}}\left(\nu, \mu\right) - \mathsf{I}_f\left(\mu_{\tilde{h}}, \mu\right) && (71)
\end{aligned}$$

A.4 Proof of Theorem 4

For any timestep $T > 0$, let $\Delta_T : \Omega \to \mathscr{P}(\Omega)$ denote the forward diffusion process dictated by the Ornstein-Uhlerbeck process and let $\overleftarrow{\Delta_T}$ be the reverse process. Furthermore, let $\overleftarrow{\mathsf{S}}_{\vartheta}$ be the process after following the DDPM algorithm at T steps. We then have

$$\begin{aligned}
\hat{P}_0 &= \overleftarrow{\Delta}_{T\#}\Delta_{T\#}\hat{P}_0 && (72)\\
\mathcal{L}(\mathsf{Y}^{\vartheta}_{T}) &= \overleftarrow{\mathsf{S}}_{\vartheta}\#\gamma_d. && (73)
\end{aligned}$$

We then have

$$\mathsf{D}_{f,\mathcal{H}}\left(\hat{P}_0, \mathcal{L}\left(\mathsf{Y}_T^{\vartheta}\right)\right) \tag{74}$$

$$= \mathsf{D}_{f,\mathcal{H}}\left(\overleftarrow{\Delta_{T\#}}\Delta_{T\#}\hat{P}_0, \overleftarrow{\mathsf{S}}_{\vartheta}\#\gamma_d\right) \tag{75}$$

$$= \inf_{\overline{\mu}\in\mathscr{P}(\Omega)}\left(d_{\mathcal{H}}\left(\overleftarrow{\Delta_{T\#}}\Delta_{T\#}\hat{P}_0, \overline{\mu}\right) + \mathsf{I}_f\left(\overline{\mu} : \overleftarrow{\mathsf{S}}_{\vartheta}\#\gamma_d\right)\right) \tag{76}$$

$$\leq d_{\mathcal{H}}\left(\overleftarrow{\Delta_{T\#}}\Delta_{T\#}\hat{P}_0, \overleftarrow{\mathsf{S}}_{\vartheta\#}\Delta_{T\#}\hat{P}_0\right) + \mathsf{I}_f\left(\overleftarrow{\mathsf{S}}_{\vartheta\#}\Delta_{T\#}\hat{P}_0 : \overleftarrow{\mathsf{S}}_{\vartheta}\#\gamma_d\right) \tag{77}$$

$$\overset{(1)}{\leq} \|\mathcal{H}\| \cdot \mathrm{TV}\left(\overleftarrow{\Delta_{T\#}}\Delta_{T\#}\hat{P}_0, \overleftarrow{\mathsf{S}}_{\vartheta\#}\Delta_{T\#}\hat{P}_0\right) + \mathsf{I}_f\left(\overleftarrow{\mathsf{S}}_{\vartheta\#}\Delta_{T\#}\hat{P}_0 : \overleftarrow{\mathsf{S}}_{\vartheta}\#\gamma_d\right) \tag{78}$$

$$\overset{(2)}{\leq} \|\mathcal{H}\| \cdot \mathrm{TV}\left(\overleftarrow{\Delta_{T\#}}\Delta_{T\#}\hat{P}_0, \overleftarrow{\mathsf{S}}_{\vartheta\#}\Delta_{T\#}\hat{P}_0\right) + \mathsf{I}_f\left(\Delta_{T\#}\hat{P}_0 : \gamma_d\right) \tag{79}$$

$$\overset{(3)}{\leq} \|\mathcal{H}\| \cdot \left(1 - \exp\left(-\,\mathrm{KL}\left(\overleftarrow{\Delta_{T\#}}\Delta_{T\#}\hat{P}_0 : \overleftarrow{\mathsf{S}}_{\vartheta\#}\Delta_{T\#}\hat{P}_0\right)\right)\right) + \mathsf{I}_f\left(\Delta_{T\#}\hat{P}_0 : \gamma_d\right) \tag{80}$$

$$\overset{(4)}{\leq} \|\mathcal{H}\| \cdot \left(1 - \exp\left(-\left(\varepsilon^2 + L^2 ds + L^2 m_2^2 s^2\right)T\right)\right) + \mathsf{I}_f\left(\Delta_{T\#}\hat{P}_0 : \gamma_d\right), \tag{81}$$

where (1) is due to the fact that $d_{\mathcal{H}} \leq \|\mathcal{H}\| \cdot \mathrm{TV}$, (2) is via the data processing inequality of f-divergences, (3) is by (Bretagnolle & Huber, 1979, Lemma 2.1), and (4) is via (Chen et al., 2022, Theorem 9) under the assumptions.

A.5 Proof of Lemma 1

Using the variational form of f-divergence, we know the witness is attained at $w^* = f'(r_T)$ using (Nguyen et al., 2010), let $\|w^*\| = \sup_{\mathsf{X}\in\Omega} |f'(r_T(\mathsf{X}))|$:

$$\mathsf{I}_f(\Delta_{T\#}\hat{P}_0 : \gamma_d) = \sup_{w:\Omega\to\mathrm{dom}(f^\star)}\left(\mathbb{E}_{\Delta_{T\#}\hat{P}_0}[w] - \mathbb{E}_{\gamma_d}[f^\star \circ w]\right) \tag{82}$$

$$= \mathbb{E}_{\Delta_{T\#}\hat{P}_0}[w^*] - \mathbb{E}_{\gamma_d}[f^\star \circ w^*] \tag{83}$$

$$\overset{(1)}{\leq} \mathbb{E}_{\Delta_{T\#}\hat{P}_0}[w^*] - \mathbb{E}_{\gamma_d}[w^*] \tag{84}$$

$$= \|w^*\| \cdot \left(\mathbb{E}_{\Delta_{T\#}\hat{P}_0}\left[\frac{w^*}{\|w^*\|}\right] - \mathbb{E}_{\gamma_d}\left[\frac{w^*}{\|w^*\|}\right]\right) \tag{85}$$

$$\leq \|w^*\| \cdot \sup_{h:\|h\|\leq 1}\left(\mathbb{E}_{\Delta_{T\#}\hat{P}_0}[h] - \mathbb{E}_{\gamma_d}[h]\right) \tag{86}$$

$$\overset{(2)}{\leq} \|w^*\| \cdot \sqrt{\mathrm{KL}\left(\Delta_{T\#}\hat{P}_0 : \gamma_d\right)}, \tag{87}$$

where (1) is due to the fact that $f^\star(t) = \sup_{t'}(t \cdot t' - f(t')) \geq t - f(1) = t$ and (2) is via Pinsker's inequality.

A.6 Proof of Theorem 3

We apply Theorem 2 to the diffusion setting which yields the result.

A.7 Proof of Theorem 5

For brevity, let $\mu := \mathcal{L}\left(\mathsf{Y}_T^{\vartheta}\right)$ and recalling $\rho_{h^*} = \eta_{h^*\#}\mu$, which defines a distribution over $[0, 1]$. We can then write

$$\begin{aligned}
\mathsf{I}_f &= \mathbb{E}_{\mathsf{X}\sim\mu}\left[f\left(f'^{-1}\left(h^*(\mathsf{X})\right)\right)\right] && (88)\\
&= \mathbb{E}_{\mathsf{X}\sim\mu}\left[f\left(f'^{-1}\left(f'\left(\frac{\eta_{h^*}(\mathsf{X})}{1-\eta_{h^*}(\mathsf{X})}\right)\right)\right)\right] && (89)\\
&= \mathbb{E}_{\mathsf{X}\sim\mu}\left[f\left(\frac{\eta_{h^*}(\mathsf{X})}{1-\eta_{h^*}(\mathsf{X})}\right)\right] && (90)\\
&= \mathbb{E}_{t\sim\rho_{h^*}}\left[f\left(\frac{t}{1-t}\right)\right] && (91)\\
&= \int_0^1 f\left(\frac{t}{1-t}\right) d\rho_{h^*}(t). && (92)
\end{aligned}$$

A.8 Proof of Theorem 6

Using a standard application of McDiarmind's inequality such as in (Bousquet et al., 2004, Lemma 5) and (Zhang et al., 2017, Theorem 3.1), we have

$$d_{\mathcal{H}}(P, \hat{P}_0) \le \mathscr{R}_n(\mathcal{H}) + 2\left\|\mathcal{H}\right\| \cdot \sqrt{\frac{1}{2n}\ln\left(\frac{1}{\delta}\right)}, \tag{93}$$

with probability $1-\delta$. Thus, we have

$$\begin{aligned}
d_{\mathcal{H}}\left(P, \mu_{T,\vartheta}^{\mathcal{H}}\right) &= \sup_{h\in\mathcal{H}}\left(\mathbb{E}_P[h] - \mathbb{E}_{\mu_{T,\vartheta}^{\mathcal{H}}}[h]\right) && (94)\\
&\le \sup_{h\in\mathcal{H}}\left(\mathbb{E}_P[h] - \mathbb{E}_{\hat{P}_0}[h] + \mathbb{E}_{\hat{P}_0}[h] - \mathbb{E}_{\mu_{T,\vartheta}^{\mathcal{H}}}[h]\right) && (95)\\
&\le \sup_{h\in\mathcal{H}}\left(\mathbb{E}_P[h] - \mathbb{E}_{\hat{P}_0}[h]\right) + \sup_{h\in\mathcal{H}}\left(\mathbb{E}_{\hat{P}_0}[h] - \mathbb{E}_{\mu_{T,\vartheta}^{\mathcal{H}}}[h]\right). && (96)
\end{aligned}$$

The first term can be bounded by equation 93 and the second term can be decomposed via Theorem 3.

A.9 Derivation for Example 1

We focus on the setting described for Variational Inference where we will show that

$$\sup_{b\in\mathbb{R}}\left(b - \mathbb{E}_\mu[\exp\left(-L+b-1\right)]\right) = \log\mathbb{E}_\mu\left[\exp\left(-L\right)\right]. \tag{97}$$

First we set $M = \mathbb{E}_\mu\left[\exp(-L-1)\right]$ then note that

$$\begin{aligned}
b - \mathbb{E}_\mu[\exp\left(-L+b-1\right)] &= b - e^b\mathbb{E}_\mu[\exp\left(-L-1\right)] && (98)\\
&= b - e^b M. && (99)
\end{aligned}$$

Differentiating this objective with respect to b yields the optimal b is $b^* = -\log M$.

A.10 Proof of Example 2

We first derive the conjugate of $f(t)$:

$$\begin{aligned}
f^\star(t) &= \sup_{t'}\left(t\cdot t' - f(t')\right) && (100)\\
&= \sup_{t'}\left(t\cdot t' - t'\log(t') + (t'+1)\log(t'+1) - 2\log 2\right) && (101)\\
&= \begin{cases}-2\log 2 - \log(1-\exp(t)) & \text{if } t<0\\ \infty & \text{if } t\ge 0.\end{cases} && (102)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathsf{D}_{f,\mathcal{H}}(\nu,\mu) &= \sup_{h\in\mathcal{H}} \left(\mathbb{E}_\nu[h] - \mathbb{E}_\mu[f^\star \circ h]\right) && (103)\\
&= \sup_{\theta\in\Theta} \left(\mathbb{E}_\nu[\log(\eta_\theta)] - \mathbb{E}_\mu[f^\star \circ (\log(\eta_\theta))]\right) && (104)\\
&= \sup_{\theta\in\Theta} \left(\mathbb{E}_\nu[\log(\eta_\theta)] + \mathbb{E}_\mu[\log(1-\eta_\theta)]\right) + 2\log 2 && (105)\\
&= -\inf_{\theta\in\Theta} \left(\mathbb{E}_\nu[-\log(\eta_\theta)] + \mathbb{E}_\mu[-\log(1-\eta_\theta)]\right) + 2\log 2. && (106)
\end{aligned}$$

A.11 Additional Results

Lemma 4 *For any $\mu,\nu \in \mathscr{P}(\Omega)$, if $f(t) = t\log t - (t+1)\log(t+1) + 2\log 2$, we have that $\mathsf{I}_f(\mu : \nu) \leq \mathrm{KL}(\mu : \nu)$.*

Proof Let $f_{\mathrm{KL}}(t) = t\log t$ and $f_{\mathrm{excess}}(t) = 2\log 2 - (t+1)\log(t+1)$ then we have

$$\begin{aligned}
\mathsf{I}_f(\mu : \nu) &= \mathbb{E}_\nu\left[f(d\mu/d\nu)\right] && (107)\\
&= \mathbb{E}_\nu\left[f_{\mathrm{KL}}(d\mu/d\nu) + f_{\mathrm{excess}}(d\mu/d\nu)\right] && (108)\\
&= \mathbb{E}_\nu\left[f_{\mathrm{KL}}(d\mu/d\nu)\right] + \mathbb{E}_\nu\left[f_{\mathrm{excess}}(d\mu/d\nu)\right] && (109)\\
&\overset{(1)}{\leq} \mathbb{E}_\nu\left[f_{\mathrm{KL}}(d\mu/d\nu)\right] + f_{\mathrm{excess}}\left(\mathbb{E}_\nu\left[(d\mu/d\nu)\right]\right) && (110)\\
&= \mathbb{E}_\nu\left[f_{\mathrm{KL}}(d\mu/d\nu)\right] + f_{\mathrm{excess}}\left(1\right) && (111)\\
&= \mathbb{E}_\nu\left[f_{\mathrm{KL}}(d\mu/d\nu)\right], && (112)
\end{aligned}$$

where (1) is via Jensen's inequality, noting that f_{excess} is concave. ■

The main duality result requires $\mathcal{H}$ to be closed under additive constants however we look to generalizing this, to get a better understanding of how this plays a role in discriminator guidance. We first consider a variant of Lemma 2 that relaxes the constraint of Q being a probability measure and even $f'^{-1}(t) \geq 0$.

Lemma 5 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a lower semi-continuous convex function with $f(1) = 0$. For a fixed $h \in \mathscr{F}(\Omega, \mathbb{R})$, we have*

$$\mu_h \in \operatorname*{arg\,inf}_{\mu\in\mathscr{B}(\Omega)} \mathscr{L}(h, Q) \implies \frac{d\mu_h}{d\mu} = f'^{-1}(h). \tag{113}$$

Proof Similar to the proof of Lemma 2, we can reparametrize Q with r: $Q = r\cdot\mu$ however we have no other restriction on r. Thus, we have:

$$\mathscr{L}(Q,\mu) = \mathbb{E}_\nu[h] - \mathbb{E}_\mu[r\cdot h] + \mathbb{E}_\mu[f(r)], \tag{114}$$

and differentiating with respect to r and setting the derivative to zero yields:

$$\begin{aligned}
0 &= -h + f'(r) && (115)\\
\implies r &= f'^{-1}(h). && (116)
\end{aligned}$$

■

Furthermore, we have that

$$\begin{aligned}
\sup_{h\in\mathcal{H}} \inf_{Q\in\mathscr{B}(\Omega)} \mathscr{L}(h,\mu) &= \sup_{h\in\mathcal{H}} \inf_{\mu\in\mathscr{B}(\Omega)} \left\{\mathbb{E}_\nu[h] - \mathbb{E}_Q[h] + \mathsf{I}_f(Q:\mu)\right\} && (117)\\
&= \sup_{h\in\mathcal{H}} \left\{\mathbb{E}_\nu[h] - \sup_{\mu\in\mathscr{B}(\Omega)} \left\{\mathbb{E}_Q[h] - \mathsf{I}_f(Q:\mu)\right\}\right\} && (118)\\
&\overset{(1)}{=} \sup_{h\in\mathcal{H}} \left\{\mathbb{E}_\nu[h] - \mathbb{E}_\mu[f^\star \circ h]\right\}, && (119)
\end{aligned}$$

where (1) is due to the fact that $E_\mu[f^\star \circ h]$ is the Legendre-Fenchel dual of $Q \mapsto \mathsf{I}_f(Q : \mu)$. In order to see this, note that $Q \mapsto \mathsf{I}_f(Q : \mu)$ is proper, convex and lower semicontinuous and thus by Theorem 7, it suffices to show $Q \mapsto \mathsf{I}_f(Q : \mu)$ is the Legendre-Fenchel dual of $h \mapsto \mathscr{K}(h) := \mathbb{E}_\mu[f^\star \circ h]$:

$$\mathscr{K}^\star(Q) = \sup_{h \in \mathscr{F}(\Omega, \mathbb{R})} \left\{ \mathbb{E}_Q[h] - \mathscr{K}(h) \right\} \tag{120}$$

$$= \sup_{h \in \mathscr{F}(\Omega, \mathbb{R})} \left\{ \mathbb{E}_Q[h] - \mathbb{E}_\mu[f^\star \circ h] \right\} \tag{121}$$

$$\overset{(1)}{=} \mathsf{I}_f(Q : \mu), \tag{122}$$

where (1) is due to the variational formulation of I_f (Nguyen et al., 2010). Note that we do not require $\mathcal{H}$ to be closed under additive constants, the only caveat is that the set $\mathscr{B}(\Omega)$ is not compact, meaning that we cannot apply a minimax theorem to get strong duality. However under the assumption of the existence of a compact set, we can get the result for $\mathcal{H}$ that are not additive.

Assumption 4 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a lower semi-continuous convex function with $f(1) = 0$. For any set of functions $\mathcal{H} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, suppose there exists a convex compact set $\mathsf{B} \subseteq \mathscr{B}(\Omega)$ such that $\mathscr{P}(\Omega) \subset \mathsf{B}$ and*

$$\left\{ \mu_h : \frac{d\mu_h}{d\mu} = f'^{-1}(h) \right\} \subset \mathsf{B}. \tag{123}$$

One way of having such an assumption satisfied is if each μ_h satisfies $\mu_h(\Omega) \leq 1$ then B can be the unit ball in $\mathscr{B}(\Omega)$ which is compact under the vague topology by the Banach-Alaoglu theorem. Note that if we pick $\mathcal{H}$ to be parametrized in the propoer composite loss framework (Savage, 1971):

$$\mathcal{H} = \left\{ f'\left(\frac{\eta_\theta}{1 - \eta_\theta} \right) : \theta \in \Theta \right\}, \tag{124}$$

where $\theta \mapsto \eta_\theta \in [0, 1]$ is an arbitrarily parametrized function such as a deep neural network then we have

$$\mu_h(\Omega) \leq 1 \iff \mathbb{E}_\mu \left[\frac{\eta_\theta}{1 - \eta_\theta} \right] \leq 1. \tag{125}$$

Since η_θ is the class probability estimate of a point to be *not* in the support of μ (since we discriminate between μ and ν), we can expect the parametrized models to mostly satsify this property. Under this assumption, we have a generalized duality.

Theorem 8 *Let $f : \mathbb{R} \to (-\infty, \infty]$ be a lower semi-continuous convex function with $f(1) = 0$. For any set of convex functions $\mathcal{H} \subseteq \mathscr{F}(\Omega, \mathbb{R})$, suppose there exists B from the above Assumption, then we have*

$$\sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_\nu[h] - \mathbb{E}_\mu[f^\star \circ h] \right\} = \inf_{Q \in \mathsf{B}} \left\{ d_{\mathcal{H}}(\nu, Q) + \mathsf{I}(Q : \mu) \right\}. \tag{126}$$

Proof Note that from equation 122 we have

$$\sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_\nu[h] - \mathbb{E}_\mu[f^\star \circ h] \right\} = \sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_\nu[h] - \sup_{Q \in \mathscr{B}(\Omega)} \left\{ \mathbb{E}_Q[h] - \mathsf{I}_f(Q : \mu) \right\} \right\} \tag{127}$$

$$= \sup_{h \in \mathcal{H}} \inf_{Q \in \mathscr{B}(\Omega)} \left\{ \mathbb{E}_\nu[h] - \mathbb{E}_Q[h] + \mathsf{I}_f(Q : \mu) \right\} \tag{128}$$

$$\overset{(1)}{=} \sup_{h \in \mathcal{H}} \inf_{Q \in \mathsf{B}} \left\{ \mathbb{E}_\nu[h] - \mathbb{E}_Q[h] + \mathsf{I}_f(Q : \mu) \right\} \tag{129}$$

$$\overset{(2)}{=} \inf_{Q \in \mathsf{B}} \sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_\nu[h] - \mathbb{E}_Q[h] + \mathsf{I}_f(Q : \mu) \right\} \tag{130}$$

$$= \inf_{Q \in \mathsf{B}} \left\{ d_{\mathcal{H}}(\nu, Q) + \mathsf{I}(Q : \mu) \right\}, \tag{131}$$

where (1) is due to the fact that B contains the optimal measure by Lemma 5 and (2) is due to the fact that since $\mathcal{H}$ and B are convex and B is compact, we are able to apply Ky Fan's minimax Theorem

(Fan, 1953) in the same way as (Liu & Chaudhuri, 2018, Lemma 27). ■

Putting this Theorem to use, we have that

$$\sup_{h\in\mathcal{H}} \{\mathbb{E}_\nu[h] - \mathbb{E}_\mu[f^\star \circ h]\} = \inf_{Q\in\mathsf{B}} \{d_{\mathcal{H}}(\nu, Q) + \mathsf{I}(Q : \mu)\} \leq \inf_{Q\in\mathscr{P}(\Omega)} \{d_{\mathcal{H}}(\nu, Q) + \mathsf{I}(Q : \mu)\}. \tag{132}$$

Note that this inequality becomes tight when $\mathcal{H}$ is large enough to be closed under addition however the optimal *refined* measure appearing in the optimization problem still satisfies the form

$$\frac{d\mu^{\mathcal{H}}}{d\mu} = f'^{-1}(h^*), \tag{133}$$

where h^* is the optimal function from the dual problem. The only difference here is that the measure $\mu^{\mathcal{H}}$ may not necessarily be a probability measure. Thus in practice, if $\mathcal{H}$ is not closed under addition, it is intuitive to compute $f'^{-1}(h^*)$ and normalize it as a heuristic.

We now consider deriving the full framework for $f(t) = t\log t$, the KL-divergence. In this case, note that f is still strictly convex, with $f'^{-1}(t) \geq 0$. Next, note that $f^\star(t) = \exp(t-1)$, thus the discriminator task is

$$\mathsf{D}_{f,\mathcal{H}} = \sup_{h\in\mathcal{H}} \left(\mathbb{E}_\nu\left[h\right] - \mathbb{E}_\mu[\exp(h-1)]\right), \tag{134}$$

and the final refined distribution will be:

$$\mu^{\mathcal{H}} = \mu \cdot \exp(h^*)/\mathbb{E}_\mu[h^*]. \tag{135}$$

We note that boosted density estimation algorithms have been developed using discriminators in combination with the expression in (Cranko & Nock, 2019; Husain et al., 2020; Soen et al., 2020).