# Moreau-Yosida $f$-divergences

Dávid Terjék [1]

## Abstract

Variational representations of $f$-divergences are central to many machine learning algorithms, with Lipschitz constrained variants recently gaining attention. Inspired by this, we define the Moreau-Yosida approximation of $f$-divergences with respect to the Wasserstein-1 metric. The corresponding variational formulas provide a generalization of a number of recent results, novel special cases of interest and a relaxation of the hard Lipschitz constraint. Additionally, we prove that the so-called tight variational representation of $f$-divergences can be to be taken over the quotient space of Lipschitz functions, and give a characterization of functions achieving the supremum in the variational representation. On the practical side, we propose an algorithm to calculate the tight convex conjugate of $f$-divergences compatible with automatic differentiation frameworks. As an application of our results, we propose the Moreau-Yosida $f$-GAN, providing an implementation of the variational formulas for the Kullback-Leibler, reverse Kullback-Leibler, $\chi^2$, reverse $\chi^2$, squared Hellinger, Jensen-Shannon, Jeffreys, triangular discrimination and total variation divergences as GANs trained on CIFAR-10, leading to competitive results and a simple solution to the problem of uniqueness of the optimal critic.

## 1. Introduction

Variational representations of divergences between probability measures are central to many machine learning algorithms, such as generative adversarial networks (Nowozin et al., 2016), mutual information estimation (Belghazi et al., 2018) and maximization (Hjelm et al., 2019), and energy-based models (Arbel et al., 2021). One class of such measures is the family of $f$-divergences (Csiszár, 1963; Ali & Silvey, 1966; Csiszár, 1967), generalizing the well-known Kullback-Leibler divergence from information theory. Another is the family of optimal transport distances (Villani, 2008), including the Wasserstein-1 metric. In general, variational representations are supremums of integral formulas taken over sets of functions, such as the Donsker-Varadhan formula (Donsker & Varadhan, 1976) for the Kullback-Leibler divergence or the Kantorovich-Rubinstein formula (Villani, 2008) for the Wasserstein-1 metric. Informally speaking, one can implement (Nowozin et al., 2016; Arjovsky et al., 2017) such a formula by constructing a real-valued neural network called the critic (or discriminator) taking samples from the two probability measures as inputs, which is then trained to maximize the integral formula in order to approximate the supremum, resulting in a learned proxy to the actual divergence of said probability measures. Implementing the Kantorovich-Rubinstein formula in such a way involves restricting the Lipschitz constant of the neural network (Gulrajani et al., 2017; Petzka et al., 2018; Miyato et al., 2018; Adler & Lunz, 2018; Terjék, 2020), which effectively stabilizes the approximation procedure. Recently, Lipschitz regularization has been incorporated (Farnia & Tse, 2018; Zhou et al., 2019; Ozair et al., 2019; Song & Ermon, 2020; Arbel et al., 2021; Birrell et al., 2020) into learning algorithms based on variational formulas of divergences other than the Wasserstein-1 metric, leading to the same empirical effect and a number of theoretical benefits.

Inspired by this, we study Lipschitz-constrained variational representations of $f$-divergences. We show that existing instances of such variants are special cases of the Moreau-Yosida approximation of $f$-divergences with respect to the Wasserstein-1 metric. To any divergence and pair of probability measures corresponds a set of optimal critics, which are exactly those functions which achieve the supremum in the variational representation. An optimal critic corresponding to $f$-divergences is not Lipschitz in general (not even continuous). Since any function represented by a neural network is Lipschitz, when a neural network is trained to approximate such a divergence, its ”target”, an optimal critic, will never be reached. We show that when the divergence is replaced by its Moreau-Yosida approximation, the corresponding optimal critics are all Lipschitz continuous with uniformly bounded Lipschitz constants, leading to a divergence which is easier to approximate in practice via neural networks. The approximation is parametrized by a

[1] Alfréd Rényi Institute of Mathematics, Budapest, Hungary. Correspondence to: Dávid Terjék <dterjek@renyi.hu>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

pair of real numbers, one of which controls the sharpness of the approximation and the Lipschitz constant of optimal critics. The other controls the behavior of the approximation such that a special case induces a hard Lipschitz constraint in the variational representation, and other values induce only a Lipschitz penalty term. While instances of the former already appeared in the literature, the latter is novel to our paper. A special case reduces to a novel, unconstrained variational representation of the Wasserstein-1 metric.

In order to prove these results, we first generalize the so-called tight variational representation of $f$-divergences to be taken over the space of Lipschitz functions or its quotient space, which is the subspace of functions vanishing at an arbitrary, fixed point. The latter leads to optimal critics being unique, having practical benefits. We additionally characterize the functions achieving the supremum in the variational representation. To apply the results, we propose an algorithm compatible with automatic differentiation frameworks to calculate the tight convex conjugate of $f$-divergences which in most cases does not admit a closed form, using Newton's method in the forward pass and implicit differentiation in the backward pass.

Finally, to demonstrate the usefulness of our results, we propose the Moreau-Yosida $f$-GAN, and implement it for the task of generative modeling on CIFAR-10. The experiments show that it is beneficial to use the Moreau-Yosida approximation as a proxy for $f$-divergences, the novel cases of which often outperform the ones with the hard Lipschitz constraint. On the other hand, the representation over the quotient space leads to a simple solution for the problem of uniqueness of the optimal critic.

To summarize, our contributions are

- a generalization of the tight variational representation of $f$-divergences between probability measures on compact metric spaces along with a characterization of functions achieving the supremum,
- a practical algorithm to calculate the tight convex conjugate of $f$-divergences compatible with automatic differentiation frameworks,
- variational formulas for the Moreau-Yosida approximation of $f$-divergences with respect to the Wasserstein-1 metric, including a relaxation of the hard Lipshcitz constraint and an unconstrained variational representation of the Wasserstein-1 metric, and
- the Moreau-Yosida $f$-GAN implementing the variational formulas for the Kullback-Leibler, reverse Kullback-Leibler, $\chi^2$, reverse $\chi^2$, squared Hellinger, Jensen-Shannon, Jeffreys, triangular discrimination and total variation divergences as GANs trained on CIFAR-10, leading to competitive performance.

## 2. Preliminaries

### 2.1. Notations

Denote the extended reals $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, the nonnegative reals $\mathbb{R}_+$, the extended nonnegative reals $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \infty$. The indicator of a set $A$ is denoted $i_A$ with $i_A(x) = 0$ if $x \in A$ and $i_A(x) = \infty$ otherwise. Absolute continuity and singularity of measures is denoted $\ll$ and $\perp$, the Radon-Nikodym derivative of a measure $\mu$ with respect to a nonnegative measure $\nu$ such that $\mu \ll \nu$ by $\frac{d\mu}{d\nu}$, the support of a measure $\mu$ by $\operatorname{supp}(\mu)$, a property to hold almost everywhere with respect to a measure $\mu$ by $\mu$-a.e. The relative interior of a subset $A$ of a vector space is denoted $\operatorname{relint} A$, which for subsets of $\mathbb{R}$ only differs from the interior for singletons whose relative interior is the singleton itself.

### 2.2. Convex analysis (Zalinescu, 2002)

Given a topological vector space $X$, denote its topological dual by $X^*$, i.e. the set of real-valued continuous linear maps on $X$, which is a topological vector space itself, and the canonical pairing by $\langle \cdot, \cdot \rangle : X \times X^* \to \mathbb{R}$, which is the continuous bilinear map $((x, x^*) \to \langle x, x^* \rangle = x^*(x))$. Given a function $f : X \to \overline{\mathbb{R}}$, the set $\operatorname{dom} f = \{x \in X : f(x) < \infty\}$ is the effective domain of $f$. A function $f$ is proper if $\operatorname{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in X$, otherwise it is improper. For a convex function $f : X \to \overline{\mathbb{R}}$, its convex conjugate is $f^* : X^* \to \overline{\mathbb{R}}$ defined by $f^*(x^*) = \sup_{x \in X}\{\langle x, x^* \rangle - f(x)\}$, and its subdifferential at $x \in X$ is the set $\partial f(x) = \{x^* \in X^* \mid \forall \hat{x} \in X : \langle \hat{x} - x, x^* \rangle \leq f(\hat{x}) - f(x)\}$. The biconjugate $f^{**}$ of $f$ is the conjugate of its conjugate $f^*$, i.e. $f^{**}(x) = \sup_{x^* \in X^*}\{\langle x, x^* \rangle - f^*(x^*)\}$, which is equivalent to $f$ if $f$ is proper, convex and lower semicontinuous. In that case, the supremum of the biconjugate representation is achieved precisely at elements of $\partial f(x)$. Conversely, the supremum in the conjugate representation of $f^*(x^*)$ is achieved at elements of $\partial f^*(x^*) = \{x \in X \mid \forall \hat{x}^* \in X^* : \langle x, \hat{x}^* - x^* \rangle \leq f^*(\hat{x}^*) - f^*(x^*)\}$.

### 2.3. $f$-divergences

Given a proper, convex and lower semicontinuous function[1] $\phi : \mathbb{R} \to \overline{\mathbb{R}}$, a measure $\mu$ and a nonnegative measure $\nu$ on a measurable space $X$, the $f$-divergence $D_\phi(\mu \| \nu)$ of $\mu$ from $\nu$ is defined (Csiszár, 1963; Ali & Silvey, 1966; Csiszár, 1967; Borwein & Lewis, 1993; Csiszár et al., 1999) as

$$\int \phi \circ \frac{d\mu_c}{d\nu} d\nu + \phi'(\infty)\mu_s^+(X) - \phi'(-\infty)\mu_s^-(X). \quad (1)$$

Here, $\mu_c \ll \nu, \mu_s \perp \nu$ are the absolutely continuous and singular parts of the Lebesgue decomposition of $\mu$ with

[1] Originally, $f$ is used in place of $\phi$ (hence the name), but we reserve the symbol $f$ for other functions.

respect to $\nu$, $\mu_s^+, \mu_s^- \geq 0$ is the Jordan decomposition of the singular part, and $\phi'(\pm\infty) = \lim_{x\to\pm\infty} \frac{\phi(x)}{x} \in \overline{\mathbb{R}}$. The well-known variational representation

$$D_\phi(\mu\|\nu) = \sup_{f:X\to\mathbb{R}} \left\{ \int f d\mu - \int \phi^* \circ f d\nu \right\} \quad (2)$$

can be obtained as the biconjugate of the mapping ($\mu \to D_\phi(\mu\|\nu)$). The so-called tight variational representation

$$D_\phi(\mu\|\nu) = \sup_{f:X\to\mathbb{R}} \left\{ \int f d\mu - \inf_{\sup f(X) - \phi'(\infty) \leq \gamma} \left\{ \int \phi_+^* \circ (f-\gamma) d\nu + \gamma \right\} \right\} \quad (3)$$

with $\phi_+ = \phi + i_{\mathbb{R}_+}$ was obtained in Agrawal & Horel (2020) as the biconjugate of the mapping ($\mu \to D_\phi(\mu\|\nu) + i_{P(X)}(\mu)$) (already considered in Ruderman et al. (2012)), and is valid for pairs of probability measures $\mu, \nu$.

### 2.4. Wasserstein-1 distance (Villani, 2008)

Given probability measures $\mu, \nu$ on a metric space $(X, d)$, the Wasserstein-1 distance of $\mu$ and $\nu$ is defined as

$$W_1(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int d(x_1,x_2) d\pi(x_1,x_2), \quad (4)$$

where $\Pi(\mu,\nu)$ is the set of probability measures supported on the product space $X \times X$ with marginals $\mu$ and $\nu$. It has a well-known variational representation called the Kantorovich-Rubinstein formula

$$W_1(\mu,\nu) = \sup_{\|f\|_L \leq 1} \left\{ \int f d\mu - \int f d\nu \right\}, \quad (5)$$

where

$$\|f\|_L = \sup_{x,y\in X, x\neq y} \left\{ \frac{|f(x)-f(y)|}{d(x,y)} \right\} \quad (6)$$

is the Lipschitz norm of $f$. The supremum is achieved by the so-called Kantorovich potentials $f : X \to \mathbb{R}$, unique $\mu, \nu$-a.e. up to an additive constant.

### 2.5. Moreau-Yosida approximation

Let $(X, d)$ be a metric space and $f : X \to \overline{\mathbb{R}}$ a proper function, and $0 < \lambda, \alpha \in \mathbb{R}$ constants. The Moreau-Yosida approximation of index $\lambda$ and order $\alpha$ of $f$ is defined (Jost & Li-Jost, 2008; Dal Maso, 1993) as

$$f_{\lambda,\alpha}(x) = \inf_{y\in X} \{f(y) + \lambda d(x,y)^\alpha\}. \quad (7)$$

It holds that $\overline{f}(x) = \sup_{\lambda>0} f_{\lambda,\alpha}(x) = \lim_{\lambda\to\infty} f_{\lambda,\alpha}(x)$, where $\overline{f}$ is the greatest lower semicontinuous function with $\overline{f} \leq f$.

## 3. Lipschitz representation of $f$-divergences

In this work, we consider the set $P(X)$ of probability measures on a compact metric space $(X, d)$, which is itself a compact metric space with the metric $W_1$, metrizing the weak convergence of measures. We prove that the tight variational representation of $D_\phi$ from Agrawal & Horel (2020) can be generalized in the sense that the supremum can be taken over the set $Lip(X, x_0)$ of Lipschitz continuous functions on $X$ that vanish at an arbitrary base point $x_0 \in X$. This is a strictly smaller set than the set of bounded and measurable functions over which the supremum was taken originally. To apply convex analytic techniques, we consider the duality between vector spaces of measures and Lipschitz functions. This aspect is detailed in Appendix 8.1. An important property of the choice of vector spaces is that the topology on the space of measures generalizes the usual weak convergence of probability measures (Hanin, 1999). Proofs and more precise statements of our propositions can be found in Appendix 8.2.

**Proposition 1.** *Given probability measures $\mu, \nu \in P(X)$ and a proper, convex and lower semicontinuous function $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ strictly convex at $1$ with $\phi(1) = 0$ and $1 \in \operatorname{relint} \operatorname{dom} \phi$, the $f$-divergence $D_\phi$ has the equivalent variational representation*

$$\begin{aligned} D_\phi(\mu\|\nu) &= \sup_{f\in Lip(X)} \left\{ \int f d\mu - D_\phi^*(f\|\nu) \right\} \\ &= \sup_{f\in Lip(X,x_0)} \left\{ \int f d\mu - D_\phi^*(f\|\nu) \right\}, \end{aligned} \quad (8)$$

*with the tight convex conjugate $D_\phi^*(\cdot\|\nu) : Lip(X) \to \mathbb{R}$ being*

$$\begin{aligned} D_\phi^*(f\|\nu) &= \sup_{\mu\in P(X)} \left\{ \int f d\mu - D_\phi(\mu\|\nu) \right\} \\ &= \min_{\sup f(X) - \phi'(\infty) \leq \gamma} \left\{ \int \phi_+^* \circ (f-\gamma) d\nu + \gamma \right\}. \end{aligned} \quad (9)$$

The conjugate $D_\phi^*(\cdot\|\nu)$ is a topical function (Mohebi, 2005), meaning that $D_\phi^*(f + C\|\nu) = D_\phi^*(f\|\nu) + C$ and $D_\phi^*(f_1\|\nu) \geq D_\phi^*(f_2\|\nu)$ both hold for $\forall C \in \mathbb{R}$ and $f_1 \geq f_2$. Based on the constant additivity property, the substitution $D_\phi^*(f\|\nu) = \int f d\nu + D_\phi^*\left(f - \int f d\nu \| \nu\right)$ leads to

$$\sup_{f\in Lip(X)} \left\{ \int f d\mu - \int f d\nu - D_\phi^*\left(f - \int f d\nu \| \nu\right) \right\},$$

reinterpreting the variational representation of $D_\phi(\mu\|\nu)$ as a penalized variant of maximum mean deviation. A closed form expression for $D_\phi^*(\cdot\|\nu)$ is available for the Kullback-Leibler divergence with $D_{KL}^*(f\|\nu) = \log \int e^f d\nu$.

We call functions $f_*$ for which $D_\phi(\mu\|\nu) = \int f_* d\mu - D_\phi^*(f_*\|\nu)$ holds, i.e. $f_* \in \partial D_\phi(\mu\|\nu)$, *Csiszár potentials* of

$\mu, \nu$. This is in analogy with Kantorovich potentials, which are similarly unique $\mu, \nu$-a.e. up to an additive constant. In the second variational representation in (8), the additive constant is unique since $f(x_0) = 0$ must hold. The following result is built on Borwein & Lewis (1993, Theorem 2.10).

**Proposition 2.** *Given probability measures $\mu, \nu \in P(X)$, a function $f_* \in Lip(X)$ is a Csiszár potential of $\mu, \nu$, i.e. $D_\phi(\mu\|\nu) = \int f_* d\mu - D^*_\phi(f_*\|\nu)$, if and only if there exists $C \in \mathbb{R}$ such that the conditions*

$$\sup f_*(X) + C \leq \phi'(\infty), \tag{10}$$

$$\frac{d\mu_c}{d\nu}(x) \in \partial\phi^*_+(f_*(x) + C) \ \nu\text{-a.e.} \tag{11}$$

*and*

$$\operatorname{supp}(\mu_s) \subset \{x \in X : f_*(x) + C = \phi'(\infty)\} \tag{12}$$

*hold. Such $f_*$ are unique $\mu, \nu$-a.e. up to an additive constant.*

If $\phi$ is of Legendre type (Borwein & Lewis, 1993), then $\phi_+$ and $\phi^*_+$ are both continuously differentiable on $\operatorname{int}\operatorname{dom}\phi_+$ and $\operatorname{int}\operatorname{dom}\phi^*_+$, respectively, while ${\phi^*_+}'$ is increasing, and invertible where its value is positive with its inverse given by the strictly increasing $\phi'_+$. With these, the second condition is equivalent to

$$f_*(x) + C = \phi'_+\left(\frac{d\mu_c}{d\nu}(x)\right) \ \mu_c\text{-a.e.} \tag{13}$$

Informally, this means that $f_*$ is the strictly increasing image of the likelihood ratio. One can then deduce from the Neyman-Pearson lemma (Reid & Williamson, 2011) that for the binary experiment of discriminating samples from $\mu$ and $\nu$, the statistical test $(x \to \chi_{[\tau,\infty]}(f_*(x)))$ is a most powerful test for any threshold $\tau \in \mathbb{R}$.

Conversely to the above proposition, given $\nu \in P(X)$ and $f \in Lip(X)$, the same conditions characterize the set of $\mu_* \in P(X)$ for which the supremum is achieved in the conjugate representation of $D^*_\phi(\cdot\|\nu)$, i.e. $\mu_* \in \partial D^*_\phi(f\|\nu)$. Denoting the optimal $\gamma$ in (9) by $\gamma_{\phi,\nu}(f)$, for any $\mu_* \in P(X)$ satisfying the conditions in Proposition 2 with $C = -\gamma_{\phi,\nu}(f)$ one has $\mu_* \in \partial D^*_\phi(f\|\nu)$. For the Kullback-Leibler divergence, this reduces to the softmax $\mu_* = \frac{1}{\int e^f d\nu} e^f \cdot \nu$. In case $X$ is a finite set, this leads to a family of prediction functions obtained as gradients of $D^*_\phi(f\|\nu)$ (Blondel et al., 2020).

**Algorithm 1** Calculate $\gamma_{\phi,\nu}(f)$ and $\nabla_f \gamma_{\phi,\nu}(f)$

```
Input: $f, \nu \in \mathbb{R}^n$, $\phi : \mathbb{R} \to \overline{\mathbb{R}}$, $0 < \epsilon, \tau \in \mathbb{R}$
if $\phi'(\infty) < \infty$ then
    $\gamma = \max(f) - \phi'(\infty) + \epsilon.$
else
    $\gamma = \langle \nu, f \rangle$
end if
repeat
    $s = \frac{-\langle \nu, (\phi^*_+)'(f-\gamma)\rangle + 1}{\langle \nu, (\phi^*_+)''(f-\gamma)\rangle}$
    $\gamma = \gamma - s$
until $|s| < \tau$
$\nabla_f \gamma = \frac{\nu \odot (\phi^*_+)''(f-\gamma)}{\langle \nu, (\phi^*_+)''(f-\gamma)\rangle}$
```

We propose an algorithm for the practical evaluation of $D^*_\phi(\cdot\|\nu)$ when no closed form expression is available in the case when the support of $\nu$ is finite[2] and $\phi$ is such that $\phi^*_+$ is twice differentiable on $\operatorname{int}\operatorname{dom}\phi^*_+$ with non-vanishing second derivative. Assuming that $f$ achieves its maximum on the support of $\nu$ and that $\gamma$ achieving the minimum is unique, finding $\gamma$ reduces to a finite dimensional problem, i.e. $f, \nu$ can be considered as elements of $\mathbb{R}^n$ with $n$ being the number of elements of the support of $\nu$. Based on Newton's method and the implicit function theorem, we propose Algorithm 1 to calculate $\gamma_{\phi,\nu}(f)$ and its gradient[3]. Then, the conjugate can be calculated as

$$D^*_\phi(f\|\nu) = \langle \nu, \phi^*_+(f - \gamma_{\phi,\nu}(f))\rangle + \gamma_{\phi,\nu}(f). \tag{14}$$

The derivation of the algorithm can be found in Appendix 8.3, along with the corresponding functions $\phi_+, \phi^*_+$ and their derivatives for the Kullback-Leibler, reverse Kullback-Leibler, $\chi^2$, reverse $\chi^2$, squared Hellinger, Jensen-Shannon, Jeffreys and triangular discrimination divergences. For the Kullback-Leibler divergence, one has the closed form $\gamma_{\phi,\nu}(f) = \log \int e^f d\nu$.

We found that exploiting the constant additivity property by calculating the conjugate as

$$D^*_\phi(f\|\nu) = D^*_\phi(f - \max(f)\|\nu) + \max(f) \tag{15}$$

is beneficial to avoid numerical instabilities. This can be seen as a generalization of the log-sum-exp trick.

## 4. Moreau-Yosida approximation of $f$-divergences

Since the mapping $(\mu \to D_\phi(\mu\|\nu))$ from the metric space $(P(X), W_1)$ to $\overline{\mathbb{R}}$ is proper and lower semicontinuous, it is an ideal candidate for Moreau-Yosida approximation, for which the infimum is always achieved since $(P(X), W_1)$ is compact if $(X, d)$ is. Given $0 < \lambda, \alpha \in \mathbb{R}$, the Moreau-Yosida approximation of index $\lambda$ and order $\alpha$ of $D_\phi(\cdot\|\nu)$ with respect to $W_1$ is therefore defined as

$$D_{\phi,\lambda,\alpha}(\mu\|\nu) = \min_{\xi \in P(X)} \{D_\phi(\xi\|\nu) + \lambda W_1(\mu, \xi)^\alpha\}. \tag{16}$$

[2] Such measures are dense in $(P(X), W_1)$.

[3] $\langle \cdot, \cdot \rangle$ and $\odot$ denote the standard dot product and the element-wise product in $\mathbb{R}^n$.

This is still a divergence in the sense that $D_{\phi,\lambda,\alpha}(\mu\|\nu) \geq 0$ with equality if and only if $\mu = \nu$. The original divergence can be recovered as $D_\phi(\mu\|\nu) = \sup_{\lambda>0} D_{\phi,\lambda,\alpha}(\mu\|\nu) = \lim_{\lambda\to\infty} D_{\phi,\lambda,\alpha}(\mu\|\nu)$ for any $\alpha > 0$. Moreover, for $\alpha \geq 1$, $D_{\phi,\lambda,\alpha}(\cdot\|\nu)$ is Lipschitz continuous with respect to $W_1$. If $\alpha = 1$, the Lipschitz constant is exactly $\lambda$. In some cases[4], variational representations are available.

**Proposition 3.** *Given probability measures $\mu, \nu \in P(X)$, $\lambda > 0$, $\alpha \geq 1$ and a proper, convex and lower semicontinuous function $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ strictly convex at $1$ with $\phi(1) = 0$ and $1 \in \operatorname{relint} \operatorname{dom} \phi$, the divergence $D_{\phi,\lambda,\alpha}(\mu\|\nu)$ has the equivalent variational representation*

$$\max_{f\in Lip(X,x_0),\|f\|_L\leq\lambda} \left\{ \int f d\mu - D^*_\phi(f\|\nu) \right\} \tag{17}$$

*if $\alpha = 1$, and*

$$\max_{f\in Lip(X,x_0)} \Big\{ \int f d\mu - D^*_\phi(f\|\nu) - (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} \lambda^{\frac{1}{1-\alpha}} \|f\|_L^{\frac{\alpha}{\alpha-1}} \Big\} \tag{18}$$

*if $\alpha > 1$.*

In the limit $\alpha \to 1$, (18) converges to (17) in the sense that $\lim_{\alpha\to1} (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} \lambda^{\frac{1}{1-\alpha}} \|f\|_L^{\frac{\alpha}{\alpha-1}} = 0$ if $\|f\|_L \leq \lambda$ and $\infty$ otherwise, providing an unconstrained relaxation of the hard constraint $\|f\|_L \leq \lambda$.

Choosing $\phi = i_{\{1\}}$ (so that $D_\phi(\cdot\|\nu) = i_{\{\nu\}}$ and $D^*_\phi(f\|\nu) = \int f d\nu$), one has $D_{\phi,\lambda,\alpha}(\mu\|\nu) = \lambda W_1(\mu,\nu)^\alpha$, leading to the following unconstrained variational representation of $W_1$.

**Proposition 4.** *Given $\mu, \nu \in P(X)$, $\lambda > 0$ and $\alpha > 1$, $W_1(\mu,\nu)$ has the equivalent unconstrained variational representation*

$$\left( \frac{1}{\lambda} \max_{f\in Lip(X,x_0)} \Big\{ \int f d\mu - \int f d\nu - (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} \lambda^{\frac{1}{1-\alpha}} \|f\|_L^{\frac{\alpha}{\alpha-1}} \Big\} \right)^{\frac{1}{\alpha}}. \tag{19}$$

*The maximum is achieved at $\alpha\lambda W_1(\mu,\nu)^{\alpha-1} f_*$, with $f_*$ being a Kantorovich potential of $\mu, \nu$.*

As stated, subgradients of the mapping $(\mu \to \lambda W_1(\mu,\nu)^\alpha)$ are nothing but the Kantorovich potentials $f_*$ achieving the supremum in the Kantorovich-Rubinstein formula, scaled by the coefficient $\alpha\lambda W_1(\mu,\nu)^{\alpha-1}$. This allows the characterization of subgradients of the mapping $(\mu \to D_{\phi,\lambda,\alpha}(\mu\|\nu))$.

**Proposition 5.** *Given probability measures $\mu, \nu \in P(X)$, $\lambda > 0$, $\alpha \geq 1$ and a proper, convex and lower semicontinuous function $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ strictly convex at $1$ with $\phi(1) = 0$ and $1 \in \operatorname{relint} \operatorname{dom} \phi$, let $\xi_* \in P(X)$ be a probability measure achieving the minimum in (16), i.e. for which $D_{\phi,\lambda,\alpha}(\mu\|\nu) = D_\phi(\xi_*\|\nu) + \lambda W_1(\mu,\xi_*)^\alpha$ holds. Then there exists an $f_* \in Lip(X)$ achieving the maximum in (17) if $\alpha = 1$ or (18) if $\alpha > 1$, which is a Csiszár potential of $\xi_*, \nu$ and $\alpha\lambda W_1(\mu,\xi_*)^{\alpha-1}$ times a Kantorovich potential of $\mu, \xi_*$ at the same time.*

These imply that for any $\tau \in \mathbb{R}$, the mapping $(x \to \chi_{[\tau,\infty]}(f_*(x)))$ is a most powerful test for discriminating samples from $\xi_*$ and $\nu$, and that $\|f_*\|_L = \alpha\lambda W_1(\mu,\xi_*)^{\alpha-1}$. Informally, since $\xi_*$ is close to $\mu$ in $W_1$, the above mapping can be seen as a Lipschitz regularized version of a most powerful test for discriminating $\mu$ and $\nu$.

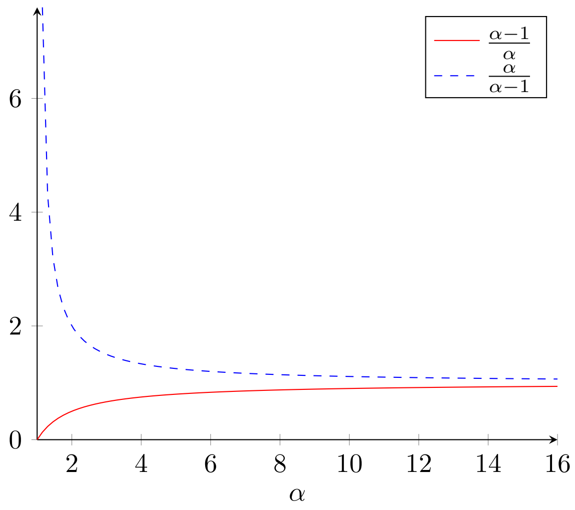

*Figure 1.* Multiplier and exponent of $\|f\|_L$

Consider the reparametrization $\lambda = \frac{1}{\alpha}\beta^{-\alpha}$, so that (18) reduces to

$$D_{\phi,\frac{1}{\alpha}\beta^{-\alpha},\alpha}(\mu\|\nu) = \max_{f\in Lip(X,x_0)} \Big\{ \int f d\mu - D^*_\phi(f\|\nu) - \frac{\alpha-1}{\alpha} (\beta\|f\|_L)^{\frac{\alpha}{\alpha-1}} \Big\}. \tag{20}$$

A plot of the respective values of the multiplier and the exponent for $\beta = 1$ and $\alpha \in [1, 16]$ are visualized in Figure 1. In the limit $\alpha \to \infty$, the multiplier and exponent both converge to 1. On the other hand, one has $\lim_{\alpha\to1} \frac{\alpha-1}{\alpha} \|f\|_L^{\frac{\alpha}{\alpha-1}} = 0$ if $\|f\|_L \leq 1$ and $\infty$ otherwise.

An interesting special case is the limit $\alpha \to \infty$, resulting in the minimum of $D_\phi(\xi\|\nu)$ with $\xi \in P(X)$ ranging over the

[4] Since the mapping $(\xi \to \lambda W_1(\mu,\xi)^\alpha)$ is neither convex nor concave if $0 < \alpha < 1$, we could not obtain a variational representation via Fenchel-Rockafellar duality in this case.

Wasserstein-1 ball of radius $\beta$ centered at $\mu$ as

$$\lim_{\alpha\to\infty} D_{\phi,\frac{1}{\alpha}\beta^{-\alpha},\alpha}(\mu\|\nu) = \min_{\xi\in P(X), W_1(\xi,\mu)\leq\beta}\{D_\phi(\xi\|\nu)\}$$
$$= \max_{f\in Lip(X,x_0)}\left\{\int f d\mu - D^*_\phi(f\|\nu) - \beta\|f\|_L\right\}. \quad (21)$$

This should be contrasted with (17) corresponding to the $\alpha = 1$ case, which also has a hard constraint, but in the dual formula.

Since the values of the above formulas for a given $f$ are invariant for constant translations $f + C$, the supremums can equivalently be taken over $Lip(X)$ instead of $Lip(X, x_0)$ in all cases.

## 5. Moreau-Yosida $f$-GAN

We propose the Moreau-Yosida $f$-GAN (MY$f$-GAN) as an implementation of the variational formula of the Moreau-Yosida regularization of $D_\phi$ with respect to $W_1$. The function $f$ in (17) or (18) is parametrized by a neural network called the critic, which is trained to maximize the formula inside the maximum, providing an approximation of the exact value of the divergence. One of the measures $\mu, \nu$ is represented by the dataset, and the other by a neural network called the generator. The generator transforms samples from a fixed noise distribution into ones resembling the data distribution, and is trained to minimize the divergence approximated by the critic.

Based on the reparametrized formula (20) with the substitution $D^*_\phi(f\|\nu) = \int f d\nu + D^*_\phi\left(f - \int f d\nu\|\nu\right)$, the two minimax games are the following. First let $\mu$ be the generated distribution and $\nu$ be the data, resulting in the forward (→) formulation

$$\min_{\theta_g\in\mathbb{R}^l}\max_{\theta_f\in\mathbb{R}^k} \mathbb{E}_{(\zeta_n,\nu_n)\sim(P_z,P_d)}\langle g_{\theta_g}\#\zeta_n, f_{\theta_f}\rangle - \langle\nu_n, f_{\theta_f}\rangle$$
$$- D^*_\phi(f_{\theta_f} - \langle\nu_n, f_{\theta_f}\rangle\|\nu_n)$$
$$- \frac{\alpha-1}{\alpha}(\beta\|f_{\theta_f}\|_{L,g_{\hat\theta_g\#}\zeta_n,\nu_n})^{\frac{\alpha}{\alpha-1}}. \quad (22)$$

Now let $\mu$ be the data and $\nu$ the generated distribution, leading to the reverse (←) formulation

$$\min_{\theta_g\in\mathbb{R}^l}\max_{\theta_f\in\mathbb{R}^k} \mathbb{E}_{(\mu_n,\zeta_n)\sim(P_d,P_z)}\langle\mu_n, f_{\theta_f}\rangle - \langle g_{\theta_g}\#\zeta_n, f_{\theta_f}\rangle$$
$$- D^*_\phi(f_{\theta_f} - \langle g_{\hat\theta_g\#}\zeta_n, f_{\theta_f}\rangle\|g_{\hat\theta_g\#}\zeta_n)$$
$$- \frac{\alpha-1}{\alpha}(\beta\|f_{\theta_f}\|_{L,\mu_n,g_{\hat\theta_g\#}\zeta_n})^{\frac{\alpha}{\alpha-1}}. \quad (23)$$

The notation of the minimax games is the following. The functions $f : X \times \mathbb{R}^k \to \mathbb{R}$ and $g : Z \times \mathbb{R}^l \to X$ are the critic and generator neural networks parametrized by weight vectors $\theta_f \in \mathbb{R}^k$ and $\theta_g \in \mathbb{R}^l$, and $f_{\theta_f}, g_{\theta_g}$ are shorthands for $f(\cdot, \theta_f), g(\cdot, \theta_g)$. The latent space is $Z = \mathbb{R}^m$. The sample space $X \subset \mathbb{R}^n$ is a compact subset of Euclidean space equipped with the restriction of the metric induced by the Euclidean norm, e.g. $X = [-1, 1]^{3*32*32}$ for CIFAR-10. $P_d \in P(X)$ denotes the data distribution and $P_z \in P(Z)$ the noise distribution, e.g. a standard normal. Empirical measures (corresponding to minibatches) are denoted $\mu_n \sim P$, meaning that $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_{\mu,i}}$ with $(x_{\mu,i}) \subset X$ being a realization of a sequence of $n$ independent and identical copies of the random variable corresponding to $P$. The empirical measure corresponding to the generated distribution is obtained as the pushforward $g_{\theta_g}\#\zeta_n$ of the latent empirical measure $\zeta_n$ (a minibatch of noise samples) through the generator $g_{\theta_g}$. Empirical means are denoted $\langle\mu_n, f\rangle = \frac{1}{n}\sum_{i=1}^n f(x_{\mu,i})$. The conjugate $D^*_\phi$ is calculated according to (14) using the stabilization trick (15). By $\hat\theta_g$ we denote a copy of $\theta_g$, meaning that $\theta_g$ is not optimized to minimize terms containing the copy, i.e. the loss function of the generator is $\pm\langle f_{\theta_f}, g_{\theta_g}\#\zeta_n\rangle$. The term $\|f_{\theta_f}\|_{L,\mu_n,\nu_n}$ denotes a possibly data-dependent estimate of $\|f_{\theta_f}\|_L$. The minimax games include the case $\lim_{\alpha\to\infty}\frac{\alpha-1}{\alpha} = \lim_{\alpha\to\infty}\frac{\alpha}{\alpha-1} = 1$.

*Lipschitz norm estimation.* Rademacher's theorem (Weaver, 2018) states that if $\|f\|_L < \infty$ for $f : \mathbb{R}^n \to \mathbb{R}$, then $\|(x \to \|\nabla f(x)\|_2)\|_\infty = \|f\|_L$ holds, i.e. that the supremum of the function mapping $x \in \mathbb{R}^n$ to the Euclidean norm of the gradient of $f$ at $x$ is equal to the Lipschitz norm of $f$. Based on this and the gradient penalty of Gulrajani et al. (2017), we propose for $\|f_{\theta_f}\|_{L,\mu_n,\nu_n}$ the estimator

$$\mathbb{E}_{\upsilon_n\sim\mathcal{U}[0,1)} \max_{x\in\text{supp}(\upsilon_n\mu_n+(1-\upsilon_n)\nu_n)} \|\nabla f_{\theta_f}(x)\|_2 \quad (24)$$

giving a lower bound to $\|f_{\theta_f}\|_L$. Here, $\mathcal{U}[0, 1)$ is the uniform distribution on $[0, 1)$ from which an empirical measure $\upsilon_n = \frac{1}{n}\sum_{i=1}^n \delta_{u_i}$ is drawn, and $u_n\mu_n + (1 - u)\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{u_i x_{\mu,i}+(1-u_i)x_{\nu,i}}$ denotes the corresponding interpolation of $\mu_n$ and $\nu_n$. This clearly biased estimator leaves room for improvement. Constructing an unbiased estimator would require assuming a distribution for the random variable representing the value of the gradient norm of the critic, which we leave for future work.

*Relaxation of hard Lipschitz constraint.* We implement the hard constraint case $\alpha = 1$ by replacing the last term in the minimax games with the one-sided gradient penalty (Gulrajani et al., 2017; Petzka et al., 2018) $\ell\mathbb{E}_{\upsilon_n\sim\mathcal{U}[0,1)}\langle\upsilon_n\mu_n + (1 - \upsilon_n)\nu_n, (\max\{0, \|\nabla f_{\theta_f}(\cdot)\|_2 - \beta^{-1}\})^2\rangle$ with the coefficient $0 < \ell \in \mathbb{R}$ controlling the strength of the penalty. This is a widely used method to enforce the hard constraint $\|f_{\theta_f}\|_L \leq \beta^{-1}$. We visualize the maximum, mean and minimum of minibatches of gradient norms of the critic during training in Figure 2 for $\alpha = 1$ with the gradient penalty and $\alpha = 1.05$ with the estimator detailed above. The $\alpha = 1$ case does not enforce the hard constraint, since only the mean

*Table 1.* MY$f$-GAN performance on CIFAR-10

| | | $\beta = 0$ | | $\alpha = 1.05, \beta = 1$ | | $\alpha = 2, \beta = 1$ | | $\alpha = \infty, \beta = 0.5 \to 0.2$ | |
|---|---|---|---|---|---|---|---|---|---|
| $D_\phi$ | | IS | FID | IS | FID | IS | FID | IS | FID |
| Kullback-Leibler | → | 7.16 | 34.12 | 8.26 | 13.22 | 8.33 | 14.83 | | |
| | ← | | | 8.20 | 13.85 | 8.09 | 13.42 | 8.20 | 12.51 |
| reverse Kullback-Leibler | → | | | 8.33 | **12.97** | 8.30 | 13.27 | | |
| | ← | | | 8.34 | 13.24 | 8.17 | 13.13 | 8.09 | 15.26 |
| $\chi^2$ | → | | | 8.18 | 14.17 | 8.26 | 13.36 | | |
| | ← | | | 8.37 | 13.36 | 8.23 | 12.95 | 8.27 | 13.46 |
| reverse $\chi^2$ | → | | | **8.47** | 13.89 | 8.26 | 14.59 | | |
| | ← | | | 8.24 | 14.04 | **8.45** | 12.28 | 8.11 | 14.17 |
| squared Hellinger | → | | | 8.03 | 16.41 | 8.07 | 16.06 | | |
| | ← | | | 8.25 | 15.89 | 8.25 | 13.93 | **8.52** | **12.18** |
| Jensen-Shannon | → | **7.51** | **30.17** | 8.30 | 14.49 | 8.34 | 12.71 | | |
| | ← | | | 8.04 | 16.04 | 8.37 | **11.57** | 8.27 | 12.58 |
| Jeffreys | → | | | 8.09 | 13.99 | 8.21 | 14.46 | | |
| | ← | | | 8.25 | 13.32 | 8.34 | 13.04 | | |
| triangular discrimination | → | 6.45 | 43.14 | 8.42 | 13.54 | 8.08 | 14.68 | | |
| | ← | | | 8.15 | 14.28 | 8.35 | 12.21 | 8.09 | 15.13 |
| total variation | → | 7.41 | 31.09 | 8.12 | 15.44 | 8.28 | 14.61 | | |
| | ← | | | 8.08 | 13.53 | 8.12 | 13.77 | 8.05 | 14.60 |
| trivial | | | | 8.07 | 15.97 | 8.04 | 14.75 | 6.67 | 36.48 |

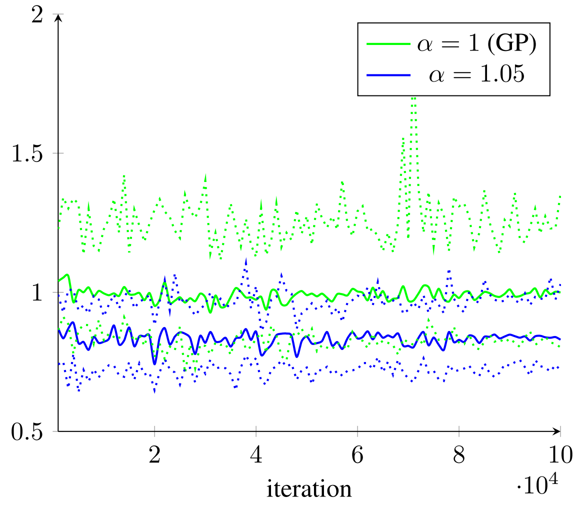

*Figure 2.* $\|\nabla f(X)\|_2$ with relaxed Lipschitz constraint

of the gradient norms is concentrated around $\beta^{-1} = 1$, and not their maximum. The $\alpha = 1.05$ case, being a relaxation of the hard constraint, empirically behaves very similarly to an ideal hard constraint implementation, in the sense that the maximum of the gradient norms is concentrated around $\beta^{-1} = 1$. This is no surprise in light of Proposition 5, since $\|f_*\|_L = \alpha\lambda W_1(\mu, \xi_*)^{\alpha-1} = \beta^{-\alpha} W_1(\mu, \xi_*)^{\alpha-1} = \beta^{-(1+\epsilon)} W_1(\mu, \xi_*)^{\epsilon}$ is very close to $\beta^{-1}$ in practice for small $\epsilon$, such as $\epsilon = 0.05$. We did not observe significant performance differences. This particular experiment used $\ell = 10$ and $\phi$ corresponding to the Kullback-Leibler divergence, but we observed identical behavior in other hyperparameter settings as well with a range of $\alpha$ close to 1. We argue that using the relaxation with some $\alpha = 1 + \epsilon$ is potentially beneficial for other applications requiring the satisfaction of a hard Lipschitz constraint.

*Choice of $f$-divergence.* Quantitative results in terms of Inception Score (IS) and Fréchet Inception Distance (FID) can be seen in Table 1. Missing values in the unregularized case ($\beta = 0$) indicate divergent training, showing that regularization ($\beta > 0$) not only improves performance, but leads to convergent training even in cases when it does not seem possible without regularization. The trivial case indicates $D_\phi(\cdot\|\nu) = i_{\{\nu\}}$, so that the forward and reverse formulations are identical. In this case, $D_{\phi,\frac{1}{\alpha}\beta^{-\alpha},\alpha}(\mu\|\nu)$ reduces to $\frac{1}{\alpha}\beta^{-\alpha}W_1(\mu\|\nu)^\alpha$. If $\alpha > 1$, this leads to an unconstrained formulation of the Wasserstein GAN corresponding to Proposition 4. The original, constrained Wasserstein GAN with gradient penalty led to an IS of 8.09 and an FID of 13.40 in our implementation. This is marginally better than the performance of the unconstrained variant as reported in Table 1. As shown in Figure 2, gradient penalty leads to a higher gradient norm than required by the hard constraint, which might lead to the observed marginal performance improvement. Indeed, increasing $\beta$ leads to better performance for the unconstrained variant, e.g. $\beta = 0.5$ with $\alpha = 2$ led to and IS of 8.14 and an FID of 13.33, which is in turn marginally better than the original, constrained variant. While it is hard to tell from these results which $f$-divergence is the best, it is definitely not the trivial one.

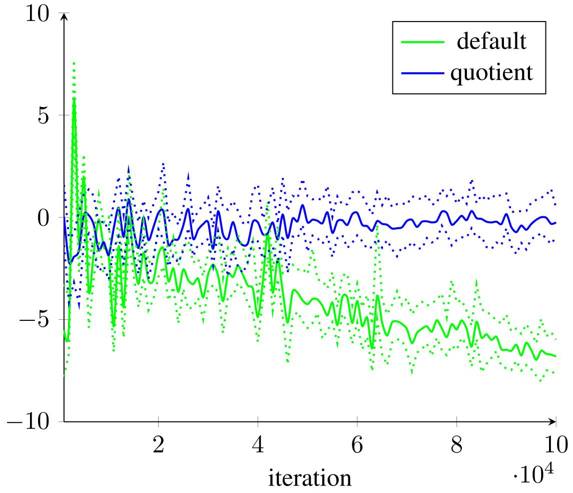

*Figure 3.* $f(X)$ for default and quotient critic

*Quotient critic.* To ensure that $f_{\theta_f} \in Lip(X, x_0)$, we simply modify the forward pass of the critic to return $f_{\theta_f}(x) - f_{\theta_f}(x_0)$ instead of $f_{\theta_f}(x)$. This induces negligible computational overhead since $f_{\theta_f}(x_0)$ can be calculated with a minibatch of size $1$, with the choice of $x_0$ being arbitrary, e.g. the zero vector in our implementation. We call this the quotient critic since $Lip(X, x_0)$ is isomorphic to the quotient space $\frac{Lip(X)}{\mathbb{R}}$. In Figure 3 we visualize the maximum, mean and minimum of the critic output over minibatches of generated samples during training. It is clear that the quotient critic solves the drifting of the output of the critic, which was found to hurt performance in some cases (Karras et al., 2018; Adler & Lunz, 2018). We observed only marginal performance improvement.

*Loss function of the generator.* The reason for picking the penalized mean deviation form of the variational formulas for this application is that in the reverse case, we found that using $-\langle g_{\theta_g}\#\zeta_n, f_{\theta_f}\rangle$ as the loss function of the generator leads to superior performance than using $-D^*_\phi(f_{\theta_f} \| g_{\theta_g}\#\zeta_n)$, which cripples performance in most cases. This suggests that gradients of the Csiszár potential $f_*$ might be of greater interest than the gradient of the conjugate $D^*_\phi(f_* \| \nu)$. The latter is a reweighting of the former, since the gradient of the conjugate is a probability distribution, such as the softmax for the Kullback-Leibler divergence.

*Optimal critic has bounded Lipschitz constant.* Notice that while the variational formula of $f$-divergences contains a supremum, the formula of their Moreau-Yosida approximations contains a maximum. This means that in the former case, even if the divergence is finite, the supremum might not be achieved by a Lipschitz function. The variational representation (8) only implies that a sequence of Lipschitz functions converges to a function achieving the supremum, but the limit is not necessarily Lipschitz continuous, in fact it might not even be continuous. On the other hand, for the Moreau-Yosida approximation, the maximum in (17) or (18) is always achieved by a Lipschitz function. Since any neural network is Lipschitz continuous, we argue that a trained critic can provide a better estimate of the Moreau-Yosida approximation, since its target $f_*$ is not only a Csiszár potential of $\xi_*, \nu$ but a scaled Kantorovich potential of $\mu, \xi_*$ as well, implying that it has a bounded Lipschitz constant.

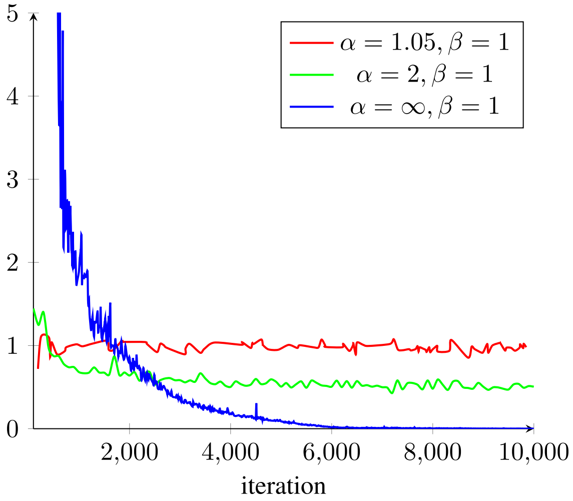

*Figure 4.* $\|f_{\theta_f}\|_{L,\mu_n,\nu_n}$ during training

*The $\alpha = 2$ and $\alpha = \infty$ cases.* Since $f_*$ is a Kantorovich potential scaled by the coefficient $\beta^{-\alpha} W_1(\mu, \xi_*)^{\alpha-1}$ and the Lipschitz norm of a Kantorovich potential is 1, the case $\alpha > 1$ can be seen as adaptive Lipschitz regularization, with $\|f_*\|_L$ decaying during training as $\mu$ and $\nu$ drift closer and $W_1(\mu, \xi_*)$ becomes smaller. We visualized $\|f_{\theta_f}\|_{L,\mu_n,\nu_n}$ in Figure 4 during training with $\alpha = 1.05, 2, \infty$ and $\beta = 1$. Ideally, the Lipschitz norm of the critic would vanish. This can be observed in the $\alpha = \infty$ case, which leads to finding a generated distribution with Wasserstein-1 distance of $\beta = 1$ from the data distribution, accordingly to (21). The best FID in Table 1 indicates that it can be beneficial to choose $\alpha = 2$ even though the Lipschitz norm does not vanish. While the case $\alpha = \infty$ (where we only consider the case $\leftarrow$) leads to low performance with high values of $\beta$ and unstable training with low values of $\beta$, we found that decaying $\beta$ e.g. from $0.5$ to $0.2$ led to the best IS as can be seen in Table 1[5]. The TRIVIAL case does not perform well in this setting, which is not surprising since the exact value of $D_{\phi,\frac{1}{\alpha}\beta^{-\alpha},\alpha}(\mu\|\nu)$ is $\infty$ if $\nu$ is not contained in the $W_1$ ball of radius $\beta$ centered at $\mu$, and $0$ otherwise.

[5] Numerical instabilities prevented us from evaluating the Jeffreys divergence in this setting.

Preliminary experiments showed that other values of $\alpha$ behave similarly to the ones we considered, which is why we restricted our attention to the representative values $1.05, 2$ and $\infty$. The implementation was done in TensorFlow, using the residual critic and generator architectures from Gulrajani et al. (2017). Training was done for $100000$ iterations, with $5$ gradient descent step per iteration for the critic, and $1$ for the generator. Additional results, details of the experimental setup and generated images can be found in Appendix 8.4, along with toy examples validating our approach for approximating $f$-divergences through the tight variational representations on categorical and Gaussian distributions. The original $f$-GAN losses (Nowozin et al., 2016) were particularly unstable in our implementation. Training the critic for $1$ instead of $5$ steps per iteration led to more stability, but even in this case only the $\chi^2$ divergence made it to $100000$ iterations without numerical errors, leading to an IS of $6.49$ and an FID of $40.64$. Source code to reproduce the experiments is available at `https://github.com/renyi-ai/moreau-yosida-f-divergences`.

## 6. Related work

In Farnia & Tse (2018), $D_{\phi,1,1}$ is defined, and a non-tight variational representation is given for symmetric choices of $D_\phi$. They also prove that $D_{\phi,1,1}$ between the data and generated distributions is a continuous function of the generator parameters, and provide a dual formula for the case $\alpha = 2$ using $W_2$ instead of $W_1$. A future direction is to prove analogous results for general $\alpha, \lambda$ and $W_p$. In Birrell et al. (2020), a generalization of $D_{\phi,1,1}$ is defined with arbitrary IPMs instead of $W_1$, but their assumptions on $\phi$ are more restrictive, and they explicitly define $D_\phi(\mu\|\nu)$ to be $\infty$ if $\mu \ll \nu$ does not hold. In Husain et al. (2019), the Lipschitz constrained version of the non-tight variational representation of $D_\phi$ is shown to be a lower bound to the Wasserstein autoencoder objective. In Laschos et al. (2019), it is proved that the supremum in the Donsker-Varadhan formula can equivalently be taken over Lipschitz continuous functions. In Song & Ermon (2020), based on the non-tight representation, another generalization of $f$-GANs and WGAN is proposed, with the importance weights $r$ analogous to the gradient of $D^*_\phi(f\|\nu)$ in our case. Connections to density ratio estimation and sample reweighting are discussed, which apply to our case as well. In Arbel et al. (2021), the Lipschitz constrained version of the Donsker-Varadhan formula is proposed as an objective function for energy-based models. For representation learning by mutual information maximization, Ozair et al. (2019) proposes the Lipschitz constrained version of the Donsker-Varadhan formula as a proxy for mutual information, which is shown to be empirically superior to the unconstrained formulation. In Zhou et al. (2019), it is shown that Lipschitz regularization improves the performance of GANs in general other than the Wasserstein GAN. The uniqueness of the optimal critic is investigated, and formulas are proposed for which uniqueness holds. We solve the uniqueness problem in another way, by implementing the quotient critic.

To summarize, the recognition of the primal formula being the Moreau-Yosida regularization of $D_\phi$ with respect to $W_1$ and the case $\alpha \neq 1$ are novel to our paper. This includes the unconstrained variational formula for $W_1$. Regarding $f$-divergences, the tight variational representation over the quotient space $Lip(X, x_0)$ and the characterization of Csiszár potentials are new as well. Additionally, we allow the same generality in terms of the choice of $\phi$ as Agrawal & Horel (2020). On the practical side, we proposed an algorithm to calculate the tight conjugate $D^*_\phi(f\|\nu)$ and its gradient. Experimentally, implementations are provided for GANs based on the tight variational representation not only of the Kullback-Leibler divergence, but the reverse Kullback-Leibler, $\chi^2$, reverse $\chi^2$, squared Hellinger, Jensen-Shannon, Jeffreys, triangular discrimination and total variation divergences as well.

## 7. Conclusions

In this paper, we studied the Moreau-Yosida regularization of $f$-divergences with respect to the Wasserstein-1 metric in a convex duality framework. We presented variational formulas and characterizations of optimal variables, generalizing a number of existing results and leading to novel special cases of interest, and proposed the MY$f$-GAN as an implementation of the formulas. Future directions include finding the variational formulas for Moreau-Yosida approximation with respect to all Wasserstein-$p$ metrics including the case $0 < \alpha < 1$, improving the estimation of the Lipschitz norm of the critic, making use of the fact that Csiszár-Kantorovich potentials can be seen as Lipschitz-regularized statistical tests, e.g. for sample reweighting, and scaling up to higher-dimensional datasets. Additionally, the results can potentially be applied to learning algorithms other than GANs, such as representation learning by mutual information maximization, energy-based models, generalized prediction functions and density ratio estimation.

## Acknowledgements

The author was supported by the Hungarian National Excellence Grant 2018-1.2.1-NKP-00008 and by the Hungarian Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program, and would like to thank the Artificial Intelligence Research Group at the Alfréd Rényi Institute of Mathematics, especially Dániel Varga and Diego González-Sánchez, for their generous help.

## References

Adler, J. and Lunz, S. Banach wasserstein GAN. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6755–6764, 2018.

Agrawal, R. and Horel, T. Optimal bounds between $f$-divergences and integral probability metrics. *CoRR*, abs/2006.05973, 2020.

Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B*, 28(1):131–142, 1966.

Arbel, M., Zhou, L., and Gretton, A. Generalized energy based models. In *International Conference on Learning Representations*, 2021.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 2017.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Hjelm, R. D., and Courville, A. C. Mutual information neural estimation. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 530–539. PMLR, 2018.

Birrell, J., Dupuis, P., Katsoulakis, M. A., Pantazis, Y., and Rey-Bellet, L. (f, $\Gamma$)-divergences: Interpolating between f-divergences and integral probability metrics. *CoRR*, abs/2011.05953, 2020. URL https://arxiv.org/abs/2011.05953.

Blondel, M., Martins, A. F. T., and Niculae, V. Learning with fenchel-young losses. *J. Mach. Learn. Res.*, 21: 35:1–35:69, 2020.

Borwein, J. M. and Lewis, A. S. Partially-finite programming in $l_1$ and the existence of maximum entropy estimates. *SIAM J. Optim.*, 3(2):248–267, 1993. doi: 10.1137/0803012.

Boyd, S. P. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2014. ISBN 978-0-521-83378-3. doi: 10.1017/CBO9780511804441.

Cobzaş, Ş., Miculescu, R., and Nicolae, A. *Lipschitz Functions*. Lecture Notes in Mathematics. Springer International Publishing, 2019. ISBN 9783030164881.

Csiszár, I. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8(1–2):85–108, 1963.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

Csiszár, I., Gamboa, F., and Gassiat, E. MEM pixel correlated solutions for generalized moment and interpolation problems. *IEEE Trans. Inf. Theory*, 45(7):2253–2270, 1999. doi: 10.1109/18.796367.

Dal Maso, G. *An Introduction to $\Gamma$-Convergence*. Birkhäuser Boston, Boston, MA, 1993. ISBN 978-1-4612-0327-8. doi: 10.1007/978-1-4612-0327-8.

Donsker, M. D. and Varadhan, S. R. S. Asymptotic evaluation of certain markov process expectations for large time—iii. *Communications on Pure and Applied Mathematics*, 29(4):389–461, 1976. doi: 10.1002/cpa.3160290405.

Farnia, F. and Tse, D. A convex duality framework for gans. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5254–5263, 2018.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5767–5777, 2017.

Hanin, L. G. An extension of the kantorovich norm. *Contemporary Mathematics*, 226:113–130, 1999.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Husain, H., Nock, R., and Williamson, R. C. A primal-dual link between gans and autoencoders. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 413–422, 2019.

Jost, J. and Li-Jost, X. *Calculus of Variations*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2008. ISBN 9780521057127.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Laschos, V., Obermayer, K., Shen, Y., and Stannat, W. A fenchel-moreau-rockafellar type theorem on the kantorovich-wasserstein space with applications in partially observable markov decision processes. *Journal of Mathematical Analysis and Applications*, 477(2):1133 – 1156, 2019. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2019.05.004.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Mohebi, H. *Topical Functions and their Properties in a Class of Ordered Banach Spaces*, pp. 343–361. Springer US, Boston, MA, 2005. ISBN 978-0-387-26771-5. doi: 10.1007/0-387-26771-9_12.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 271–279, 2016.

Ozair, S., Lynch, C., Bengio, Y., van den Oord, A., Levine, S., and Sermanet, P. Wasserstein dependency measure for representation learning. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 15578–15588, 2019.

Petzka, H., Fischer, A., and Lukovnikov, D. On the regularization of wasserstein gans. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Reid, M. D. and Williamson, R. C. Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.*, 12: 731–817, 2011.

Ruderman, A., Reid, M. D., García-García, D., and Petterson, J. Tighter variational representations of f-divergences via restriction to probability measures. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

Song, J. and Ermon, S. Bridging the gap between f-gans and wasserstein gans. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9078–9087. PMLR, 2020.

Tao, T. *Several Variable Differential Calculus*, pp. 127–161. Springer Singapore, Singapore, 2016. ISBN 978-981-10-1804-6. doi: 10.1007/978-981-10-1804-6_6.

Terjék, D. Adversarial lipschitz regularization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Villani, C. *Optimal transport – Old and new*, volume 338, pp. xxii+973. 01 2008. doi: 10.1007/978-3-540-71050-9.

Weaver, N. *Lipschitz Algebras*. WORLD SCIENTIFIC, 2nd edition, 2018. doi: 10.1142/9911.

Zalinescu, C. *Convex Analysis in General Vector Spaces*. World Scientific, 2002. ISBN 9789812380678.

Zhou, Z., Liang, J., Song, Y., Yu, L., Wang, H., Zhang, W., Yu, Y., and Zhang, Z. Lipschitz generative adversarial nets. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7584–7593. PMLR, 2019.