LEARNING WITH STOCHASTIC ORDERS

Anonymous authors
Paper under double-blind review

ABSTRACT

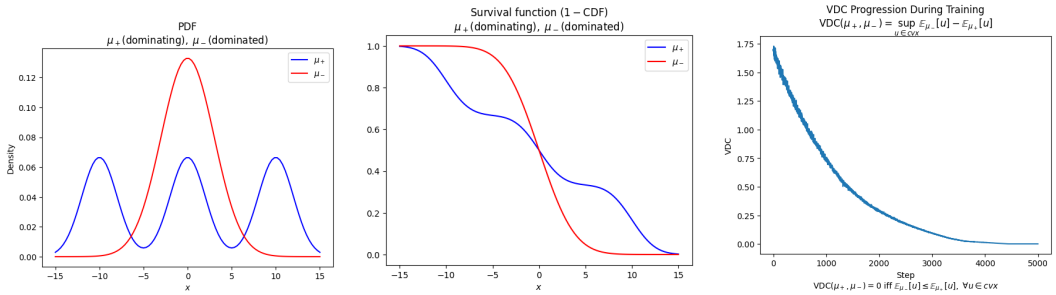
Learning high-dimensional distributions is often done with explicit likelihood modeling or implicit modeling via minimizing integral probability metrics (IPMs). In this paper, we expand this learning paradigm to stochastic orders, namely, the *convex* or *Choquet order* between probability measures. Towards this end, exploiting the relation between convex orders and optimal transport, we introduce the Choquet-Toland distance between probability measures, that can be used as a drop-in replacement for IPMs. We also introduce the *Variational Dominance Criterion* (VDC) to learn probability measures with dominance constraints, that encode the desired stochastic order between the learned measure and a known baseline. We analyze both quantities and show that they suffer from the curse of dimensionality and propose surrogates via input convex maxout networks (ICMNs), that enjoy parametric rates. We provide a min-max framework for learning with stochastic orders and validate it experimentally on synthetic and high-dimensional image generation, with promising results. Finally, our ICMNs class of convex functions and its derived Rademacher Complexity are of independent interest beyond their application in convex orders.

1 INTRODUCTION

Learning complex high-dimensional distributions with implicit generative models (Goodfellow et al., 2014; Mohamed & Lakshminarayanan, 2017; Arjovsky et al., 2017) via minimizing *integral probability metrics* (IPMs) (Müller, 1997a) has lead to the state of the art generation across many data modalities (Karras et al., 2019; De Cao & Kipf, 2018; Padhi et al., 2020). An IPM compares probability distributions with a witness function belonging to a function class $\mathcal{F}$, e.g., the class of Lipchitz functions, which makes the IPM correspond to the Wasserstein distance 1. While estimating the witness function in such large function classes suffer from the curse of dimensionality, restricting it to a class of neural networks leads to the so called neural net distance (Arora et al., 2017) that enjoys statistical parametric rates.

In probability theory, the question of comparing distributions is not limited to assessing only equality between two distributions. Stochastic orders were introduced to capture the notion of dominance between measures. Similar to IPMs, stochastic orders can be defined by looking at the integrals of measures over function classes $\mathcal{F}$ (Müller, 1997b). Namely, for $\mu_+, \mu_- \in \mathcal{P}_1(\mathbb{R}^d)$, μ_+ dominates μ_-, or $\mu_- \preceq \mu_+$, if for any function $f \in \mathcal{F}$, we have $\int_{\mathbb{R}^d} f(x)\, d\mu_-(x) \leq \int_{\mathbb{R}^d} f(x)\, d\mu_+(x)$ (See Figure 1a for an example). In the present work, we focus on the *Choquet* or *convex* order (Ekeland & Schachermayer, 2014) generated by the space of convex functions (See Sec. 2 for more details).

Previous work has focused on learning with stochastic orders in the one dimensional setting, as it has prominent applications in mathematical finance and distributional reinforcement learning (RL). The survival function gives a characterisation of the convex order in one dimension (See Figure 1b and Sec. 2 for more details). For instance, in portfolio optimization (Xue et al., 2020; Post et al., 2018; Dentcheva & Ruszczynski, 2003) the goal is to find the portfolio that maximizes a utility under dominance constraints of a known benchmark or baseline. A similar concept was introduced in distributional RL (Martin et al., 2020) for learning policies with dominance constraints on the distribution of the reward. While these works are limited to the univariate setting, our work is the first, to the best of our knowledge, that provides a computationally tractable characterization of stochastic orders that is scalable in high dimensions and is sample efficient.



(a) (b) (c)

Figure 1: VDC example in 1D. Figure 1a :μ_+ is mixture of 3 Gaussians , μ_- corresponds to a single mode of the mixture. μ_+ dominates μ_- in the convex order. Figure 1b : uni-variate characterization of the convex order with survival functions (See Sec. 2 for details). Figure 1c: Surrogate VDC computation with Input Convex Maxout Network and gradient descent. The surrogate VDC tends to zero at the end of the training and hence characterizes the convex dominance of μ_+ on μ_-.

The paper is organized as follows: we introduce in Sec. 3 the *Variational Dominance Criterion* (VDC); the VDC between measures μ_+ and μ_- takes value 0 if and only if μ_+ dominates μ_- in the convex order. The VDC suffers from the curse of dimension and can not be estimated efficiently from samples. To remediate this, we introduce in Sec. 4, a VDC surrogate via Input Convex Maxout Networks (ICMNs). ICMNs are new variants of Input Convex Neural Nets Amos et al. (2017) that we propose as proxy for convex functions and study their complexity. We show in Sec. 4 that the surrogate VDC has parametric rates and can be efficiently estimated from samples. The surrogate VDC can be computed using (stochastic) gradient descent on the parameters of the ICMN and can characterize convex dominance (See Figure 1c). We then show in Sec. 5 how to use the VDC and its surrogate to define a pseudo distance on the probability Space. Finally we propose in Sec. 6 penalizing generative models training losses with the surrogate VDC to learn implicit generative models that have better coverage and spread than known baselines. This leads to a min-max game similar to GANs. We validate our framework in Sec. 7 with experiments on portfolio optimization and image generation.

2 The Choquet or convex order

Denote by $\mathcal{P}(\mathbb{R}^d)$ the set of Borel probability measures on $\mathbb{R}^d$ and by $\mathcal{P}_1(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$ the subset of those which have finite first moment: $\mu \in \mathcal{P}_1(\mathbb{R})$ if and only if $\int_{\mathbb{R}^d} \|x\| \, d\mu(x) < +\infty$.

Comparing probability distributions Integral probability metrics (IPMs) are pseudo-distances between probability measures μ, ν defined as $d_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \mathbb{E}_\mu f - \mathbb{E}_\nu f$, for a given function class $\mathcal{F}$ which is symmetric with respect to sign flips. They are ubiquitous in optimal transport and generative modeling to compare distributions: if $\mathcal{F}$ is the set of functions with Lipschitz, then the resulting IPM is the 1-Wasserstein distance; if $\mathcal{F}$ is the unit ball of an RKHS, the IPM is its maximum mean discrepancy. Clearly, $d_{\mathcal{F}}(\mu, \nu) = 0$ if and only $\mathbb{E}_\mu f = \mathbb{E}_\nu f$, and when $\mathcal{F}$ is large enough, this is equivalent to $\mu = \nu$.

The Choquet or convex order When the class $\mathcal{F}$ is not symmetric with respect to sign flips, comparing the expectations $\mathbb{E}_\mu f$ and $\mathbb{E}_\nu f$ for $f \in \mathcal{F}$ does not yield a pseudo-distance. In the case where $\mathcal{F}$ is the set of convex functions, the convex order naturally arises instead:

Definition 1 (Choquet order, Ekeland & Schachermayer (2014), Def. 4)**.** *For $\mu_-, \mu_+ \in \mathcal{P}_1(\mathbb{R}^d)$, we say that $\mu_- \preceq \mu_+$ if for any convex function $f : \mathbb{R}^d \to \mathbb{R}$, we have*

$$\int_{\mathbb{R}^d} f(x) \, d\mu_-(x) \leq \int_{\mathbb{R}^d} f(x) \, d\mu_+(x).$$

$\mu_- \preceq \mu_+$ is classically denoted as "μ_- is a balayée of μ_+", or "μ_+ dominates μ_-". It turns out that $\preceq$ is a partial order on $\mathcal{P}_1(\mathbb{R}^d)$, meaning that reflexivity ($\mu \preceq \mu$), antisymmetry (if $\mu \preceq \nu$ and $\nu \preceq \mu$,

then $\mu = \nu$), and transitivity (if $\mu_1 \preceq \mu_2$ and $\mu_2 \preceq \mu_3$, then $\mu_1 \preceq \mu_3$) hold. As an example, if μ_-, μ_+ are Gaussians $\mu_- = N(0, \Sigma_-)$, $\mu_+ = N(0, \Sigma_+)$, then $\mu_- \preceq \mu_+$ if and only if $\Sigma_- \preceq \Sigma_+$ in the positive-semidefinite order (Müller, 2001). Also, since linear functions are convex, $\mu_- \preceq \mu_+$ implies that both measures have the same expectation: $\mathbb{E}_{x\sim\mu_-} x = \mathbb{E}_{x\sim\mu_-} x$.

In the univariate case, $\mu_- \preceq \mu_+$ implies that $\mathrm{supp}(\mu_-) \subseteq \mathrm{supp}(\mu_+)$ and that $\mathrm{Var}(\mu_-) \leq \mathrm{Var}(\mu_+)$, and we have that $\mu_- \preceq \mu_+$ holds if and only if for all $x \in \mathbb{R}$, $\int_x^{+\infty} \bar{F}_{\mu_-}(t)\, dt \leq \int_x^{+\infty} \bar{F}_{\mu_+}(t)\, dt$, where $\bar{F}_\cdot$ is the survival function (one minus the cumulative distribution function). Note that this characterization can be checked efficiently if one has access to samples of μ_- and μ_+.

In the high-dimensional case, there exists an alternative characterization of the convex order:

Proposition 1 (Ekeland & Schachermayer (2014), Thm. 10)**.** *If $\mu_-, \mu_+ \in \mathcal{P}_1(\mathbb{R}^d)$, we have $\mu_- \preceq \mu_+$ if and only if there exists a Markov kernel R (i.e. $\forall x \in \mathbb{R}^d$, $\int_{\mathbb{R}^d} y\, dR_x(y) = x$) such that $\mu_+ = \int_{\mathbb{R}^d} R_x\, d\mu_-$.*

Equivalently, there exists a coupling (X, Y) such that $\mathrm{Law}(X) = \mu_-$, $\mathrm{Law}(Y) = \mu_+$ and $X = \mathbb{E}(Y|X)$ almost surely. Intuitively, this means that μ_+ is more *spread out* than μ_-. Remark that this characterization is difficult to check, especially in high dimensions.

3 The Variational Dominance Criterion

In this section, we present a quantitative way to deal with convex orders. Given a bounded open convex subset $\Omega \subseteq \mathbb{R}^d$ and a compact set $K \subseteq \mathbb{R}^d$, let $\mathcal{A} = \{u : \Omega \to \mathbb{R},\ u \text{ convex and } \nabla u \in K \text{ almost everywhere}\}$. We define the *Variational Dominance Criterion* (VDC) between probability measures μ_+ and μ_- supported on Ω in analogously to IPMs, replacing $\mathcal{F}$ by $\mathcal{A}$:

$$\mathrm{VDC}_{\mathcal{A}}(\mu_+||\mu_-) := \sup_{u\in\mathcal{A}} \int_\Omega u\, d(\mu_- - \mu_+). \tag{1}$$

Remark that when $0 \in K$, $\mathrm{VDC}_{\mathcal{A}}(\mu_+||\mu_-) \geq 0$ because the zero function belongs to the set $\mathcal{A}$. We reemphasize that since $\mathcal{A}$ is not symmetric with respect to sign flips as $f \in \mathcal{A}$ does not imply $-f \in \mathcal{A}$, the properties of the VDC are very different from those of IPMs. Most importantly, the following proposition, shown in App. A, links the VDC to the Choquet order.

Proposition 2. *Let K compact such that the origin belongs to the interior of K. If $\mu_+, \mu_- \in \mathcal{P}(\Omega)$, $\mathrm{VDC}_{\mathcal{A}}(\mu_+||\mu_-) := \sup_{u\in\mathcal{A}} \int_\Omega u\, d(\mu_- - \mu_+) = 0$ if and only if $\int_\Omega u\, d(\mu_- - \mu_+) \leq 0$ for any convex function on Ω (i.e. $\mu_- \preceq \mu_+$ according to the Choquet order).*

That is, Proposition 2 states that the VDC between μ_+ and μ_- takes value 0 if and only if μ_+ dominates μ_-. Combining this with the interpretation of Proposition 1, we see that intuitively, *the quantity $\mathrm{VDC}_{\mathcal{A}}(\mu_+||\mu_-)$ is small when μ_+ is more spread out than μ_-, and large otherwise*. Hence, if we want to enforce or induce a Choquet ordering between two measures in an optimization problem, we can include the VDC (or rather, its surrogate introduced in Sec. 4) as a penalization term in the objective. Before this, we explore the connections between VDC and optimal transport, and study some statistical properties of the VDC.

3.1 The VDC and optimal transport

Toland duality provides a way to interpret the VDC through the lens of optimal transport. In the following, $W_2(\mu, \nu)$ denotes the 2-Wasserstein distance between μ and ν.

Theorem 1 (Toland duality, adapted from Thm. 1 of Carlier (2008))**.** *For any $\mu_+, \mu_- \in \mathcal{P}(\Omega)$, the VDC satisfies:*

$$\mathrm{VDC}_{\mathcal{A}}(\mu_+||\mu_-) = \sup_{\nu\in\mathcal{P}(K)} \left\{\tfrac{1}{2}W_2^2(\mu_+, \nu) - \tfrac{1}{2}W_2^2(\mu_-, \nu)\right\} - \tfrac{1}{2}\int_\Omega \|x\|^2\, d(\mu_+ - \mu_-)(x) \tag{2}$$

The optimal convex function u of VDC in (1) *and optimal ν in* (2) *satisfy $(\nabla u)_\# \mu_+ = (\nabla u)_\# \mu_- = \nu$.*

Note that under the assumption $0 \in K$, Theorem 1 implies that $\mathrm{VDC}_{\mathcal{A}}(\mu_+||\mu_-) = 0$ if and only if $W_2^2(\mu_+, \nu) - \frac{1}{2}\int_\Omega \|x\|^2\, d\mu_+ \leq W_2^2(\mu_-, \nu) - \frac{1}{2}\int_\Omega \|x\|^2\, d\mu_-$ for any $\nu \in \mathcal{P}(K)$. Under the

equivalence $\mathrm{VDC}_{\mathcal{A}}(\mu_+||\mu_-) = 0 \iff \mu_+ \succeq \mu_-$ shown by Proposition 2, this provides yet another characterization of the convex order for arbitrary dimension.

3.2 Statistical rates for VDC estimation

In this subsection, we prove an upper bound on the statistical rate of estimation of $\mathrm{VDC}_{\mathcal{A}}(\mu||\nu)$ using the estimator $\mathrm{VDC}_{\mathcal{A}}(\mu_n||\nu_n)$ based on the empirical distributions $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ built from i.i.d. samples $(x_i)_{i=1}^n$, $(y_i)_{i=1}^n$ from μ and ν, respectively.

Theorem 2. *Let $\Omega = [-1,1]^d$ and $K = \{x \in \mathbb{R}^d \,|\, \|x\|_2 \leq C\}$ for an arbitrary $C > 0$. With probability at least $1-\delta$,*

$$|\mathrm{VDC}_{\mathcal{A}}(\mu||\nu) - \mathrm{VDC}_{\mathcal{A}}(\mu_n||\nu_n)| \leq \sqrt{18C^2 d \log(\tfrac{\delta}{4})}(\tfrac{2}{\sqrt{n}}) + 8Kn^{-\frac{2}{d}},$$

where K depends on C and d.

The dependency on $n^{-\frac{2}{d}}$ is indicative of the curse of dimension: we need a number of samples n exponential in d to control the estimation error. While Theorem 2 only shows an upper bound on the difference between the VDC and its estimator, in Subsec. 5.1 we study a related setting where a $\tilde{\Omega}(n^{-\frac{2}{d}})$ lower bound is available. Hence, we hypothesize that VDC estimation is in fact cursed by dimension in general.

4 A VDC surrogate via input convex maxout networks

Given the link between the VDC and the convex order, one is inclined to use the VDC as a quantitative proxy to induce convex order domination in optimization problems. Estimating the VDC implies solving an optimization problem over convex functions. In practice, we only have access to the empirical versions μ_n, ν_n of the probability measures μ, ν; we could compute the VDC between the empirical measures by solving a linear program similar to the ones used in non-parametric convex regression (Hildreth, 1954). However, the statistical rates for the VDC estimation from samples are cursed by dimension (Subsec. 3.2), which means that we would need a number of samples exponential in the dimension to get a good estimate. Our approach is to focus on a surrogate problem instead:

$$\sup_{u \in \hat{\mathcal{A}}} \int_\Omega u \, d(\mu_- - \mu_+),$$

where $\hat{\mathcal{A}}$ is a class of neural network functions included in $\mathcal{A}$ over which we can optimize efficiently. In constructing $\hat{\mathcal{A}}$, we want to hardcode the constraints *u convex* and *$\nabla u \in K$ almost everywhere* into the neural network architectures. A possible approach would be to use the input convex neural networks (ICNNs) introduced by Amos et al. (2017), which have been used as a surrogate of convex functions for generative modeling with normalizing flows (Huang et al., 2021) in optimal transport Korotin et al. (2021a;b); Huang et al. (2021); Makkuva et al. (2020) and large scale Wasserstein flows Alvarez-Melis et al. (2021); Bunne et al. (2021); Mokrov et al. (2021).

However, we found in early experimentation that a superior alternative is to use input convex maxout networks (ICMNs), which are maxout networks (Goodfellow et al., 2013) that are convex with respect to inputs. Maxout networks and ICMNs are defined as follows:

Definition 2 (Maxout networks)**.** *For a depth $L \geq 2$, let $\mathcal{M} = (m_1, \ldots, m_L)$ be a vector of positive integers such that $m_1 = d$. Let $F_{L,\mathcal{M},k}$ be the space of k-maxout networks of depth L and widths $\mathcal{M}$, which contains functions of the form*

$$f(x) = \tfrac{1}{\sqrt{m_L}} \textstyle\sum_{i=1}^{m_L} a_i \max_{j\in[k]} \langle w_{i,j}^{(L-1)}, (x^{(L-1)}, 1)\rangle, \qquad a_i \in \mathbb{R},\, w_{i,j}^{(L-1)} \in \mathbb{R}^{m_{L-1}+1} \quad (3)$$

where for any $2 \leq \ell \leq L-1$, and any $1 \leq i \leq m_\ell$, the i-th component of $x^{(\ell)} = (x_1^{(\ell)}, \ldots, x_{m_\ell}^{(\ell)})$ is computed recursively as:

$$x_i^{(\ell)} = \tfrac{1}{\sqrt{m_\ell}} \max_{j\in[k]} \langle w_{i,j}^{(\ell-1)}, (x^{(\ell-1)}, 1)\rangle, \qquad w_{i,j}^{(\ell)} \in \mathbb{R}^{m_\ell+1}, \quad (4)$$

with $x^{(1)} = x$.

Definition 3 (Input convex maxout networks or ICMNs)**.** *A maxout network f of the form* (3)-(4) *is an input convex maxout network if (i) for any $1 \leq i \leq M_L$, $a_i \geq 0$, and (ii) for any $2 < \ell \leq L-1$, $1 \leq i \leq m_{\ell+1}$, $1 \leq j \leq k$, the first m_ℓ components of $w_{i,j}^{(\ell)}$ are non-negative. We denote the space of ICMNs as $F_{L,\mathcal{M},k,+}$.*

In other words, a maxout network is an ICMN if all the non-bias weights beyond the first layer are constrained to be non-negative. This definition is analogous to the one of ICNNs in Amos et al. (2017), which are also defined as neural networks with positivity constraints on non-bias weights beyond the first layer. Proposition 5 in App. B shows that ICMNs are convex w.r.t to their inputs.

It remains to impose the condition $\nabla u \in K$ *almost everywhere*, which in practice is enforced by adding the norms of the weights as a regularization term to the loss function. For theoretical purposes, we define $F_{L,\mathcal{M},k}(1)$ (resp. $F_{L,\mathcal{M},k,+}(1)$) as the subset of $F_{L,\mathcal{M},k}$ (resp. $F_{L,\mathcal{M},k,+}$) such that for all $1 \leq \ell \leq L-1$, $1 \leq i \leq m_\ell$, $1 \leq j \leq k$, $\|w_{i,j}^{(\ell)}\|_2 \leq 1$, and $\|a\|_2 = \sum_{i=1}^{m_L} a_i^2 \leq 1$. The following proposition, proven in App. B, shows simple bounds on the values of the functions in $F_{L,\mathcal{M},k}(1)$ and their derivatives.

Proposition 3. *Let f be an ICMN that belongs to $F_{L,\mathcal{M},k}(1)$. For x almost everywhere in $\mathcal{B}_1(\mathbb{R}^d)$, $\|\nabla f(x)\| \leq 1$. Moreover, for all $x \in \mathcal{B}_1(\mathbb{R}^d)$, $|f(x)| \leq L$, and for $1 \leq \ell \leq L$, $\|x^{(\ell)}\| \leq \ell$.*

When $K = \mathcal{B}_1(\mathbb{R}^d)$, we have that the space of ICMNs $F_{L,\mathcal{M},k,+}(1)$ is included in $\mathcal{A}$. Hence, we define the surrogate VDC associated to $F_{L,\mathcal{M},k,+}(1)$ as:

$$\mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(\mu_+||\mu_-) = \sup_{u \in F_{L,\mathcal{M},k,+}(1)} \int_\Omega u \, d(\mu_- - \mu_+). \tag{5}$$

Theorem 3. *Suppose that for all $1 \leq \ell \leq L$, the widths m_ℓ satisfy $m_\ell \leq m$, and assume that μ, ν are supported on the ball of $\mathbb{R}^d$ of radius r. With probability at least $1-\delta$,*

$$\begin{aligned}&|\mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(\mu||\nu) - \mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(\mu_n||\nu_n)| \\ &\leq \sqrt{18r^2\log(\tfrac{\delta}{4})}\big(\tfrac{2}{\sqrt{n}}\big) + 512\sqrt{\tfrac{(L-1)km(m+1)}{n}}\Big(\sqrt{(L+1)\log(2) + \tfrac{\log(L^2+1)}{2}} + \tfrac{\sqrt{\pi}}{2}\Big),\end{aligned}$$

We see from Theorem 3 that $\mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}$ in contrast to $\mathrm{VDC}_{\mathcal{A}}$ has parametric rates and hence favorable properties to be estimated from samples. In the following section, wes see that VDC can be used to defined a pseudo-distance on the probability space.

5 From the convex order back to a pseudo-distance

We define the *Choquet-Toland distance* (CT distance) as the map $d_{\mathrm{CT},\mathcal{A}} : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \mathbb{R}$ given by

$$d_{\mathrm{CT},\mathcal{A}}(\mu_+, \mu_-) := \mathrm{VDC}_{\mathcal{A}}(\mu_+||\mu_-) + \mathrm{VDC}_{\mathcal{A}}(\mu_-||\mu_+).$$

That is, the CT distance between μ_+ and μ_- is simply the sum of Variational Dominance Criteria. Applying Theorem 1, we obtain that $d_{\mathrm{CT},\mathcal{A}}(\mu_+, \mu_-) = \frac{1}{2}(\sup_{\nu \in \mathcal{P}(K)} \{W_2^2(\mu_+, \nu) - W_2^2(\mu_-, \nu)\} + \sup_{\nu \in \mathcal{P}(K)} \{W_2^2(\mu_+, \nu) - W_2^2(\mu_-, \nu)\})$. The following result, shown in App. C, states that $d_{CT,K}$ is indeed a distance.

Theorem 4. *Suppose that the origin belongs to the interior of K. $d_{CT,\mathcal{A}}$ is a distance, i.e. it fulfills*

(i) $d_{CT,\mathcal{A}}(\mu_+, \mu_-) \geq 0$ for any $\mu_+, \mu_- \in \mathcal{P}(\Omega)$ (non-negativity).

(ii) $d_{CT,\mathcal{A}}(\mu_+, \mu_-) = 0$ if and only if $\mu_+ = \mu_-$ (identity of indiscernibles).

(iii) If $\mu_1, \mu_2, \mu_3 \in \mathcal{P}(\Omega)$, we have that $d_{CT,\mathcal{A}}(\mu_1, \mu_2) \leq d_{CT,\mathcal{A}}(\mu_1, \mu_3) + d_{CT,\mathcal{A}}(\mu_3, \mu_2)$.

As in (5), we define the surrogate CT distance as:

$$d_{\mathrm{CT},F_{L,\mathcal{M},k,+}(1)}(\mu_+, \mu_-) = \mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(\mu_+||\mu_-) + \mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(\mu_-||\mu_+). \tag{6}$$

5.1 STATISTICAL RATES FOR CT DISTANCE ESTIMATION

We show almost-tight upper and lower bounds on the expectation of the Choquet-Toland distance between a probability measure and its empirical version. Namely, $\mathbb{E}[d_{\mathrm{CT},\mathcal{A}}(\mu,\mu_n)] = O(n^{-\frac{2}{d}}), \Omega(n^{-\frac{2}{d}}/\log(n))$.

Theorem 5. *Let C_0, C_1 be universal constants independent of the dimension d. Let $\Omega = [-1,1]^d$, and $K = \{x \in \mathbb{R}^d \mid \|x\|_2 \leq C\}$. Let μ_n be the n-sample empirical measure corresponding to a probability measure μ over $[-1,1]^d$. When μ is the uniform probability measure and $n \geq C_1 \log(d)$, we have that*

$$\mathbb{E}[d_{CT,\mathcal{A}}(\mu,\mu_n)] \geq C_0 \sqrt{\frac{C}{d(1+C)\log(n)}} n^{-\frac{2}{d}}. \tag{7}$$

For any probability measure μ over $[-1,1]^d$,

$$\mathbb{E}[d_{CT,\mathcal{A}}(\mu,\mu_n)] \leq K n^{-\frac{2}{d}}, \tag{8}$$

where K is a constant depending on C and d (but not the measure μ).

Overlooking logarithmic factors, we can summarize Theorem 5 as $\mathbb{E}[d_{\mathrm{CT},\mathcal{A}}(\mu,\mu_n)] \asymp n^{-\frac{2}{d}}$. The estimation of the CT distance is cursed by dimension: one needs a sample size exponential with the dimension d for μ_n to be at a desired distance from μ. It is also interesting to contrast the rate $n^{-\frac{2}{d}}$ with the rates for similar distances. For example, for the r-Wasserstein distance we have $\mathbb{E}[W_r^r(\mu,\mu_n)] \asymp n^{-\frac{r}{d}}$ (Singh & Póczos (2018), Section 4). Given the link of the CT distance and VDC with the squared 2-Wasserstein distance (see Subsec. 3.1), the $n^{-\frac{2}{d}}$ rate is natural.

The proof of Theorem 5, which can be found in App. D, is based on upper-bounding and lower-bounding the metric entropy of the class of bounded Lipschitz convex functions with respect to an appropriate pseudo-norm. Then we use Dudley's integral bound and Sudakov minoration to upper and lower-bound the Rademacher complexity of this class, and we finally show upper and lower bounds of the CT distance by the Rademacher complexity. Bounds on the metric entropy of bounded Lipschitz convex functions have been computed and used before (Balazs et al., 2015; Guntuboyina & Sen, 2013), but in L^p and supremum norms, not in our pseudo-norm.

Next, we see that the surrogate CT distance defined in (6) does enjoy parametric estimation rates.

Theorem 6. *Suppose that for all $1 \leq \ell \leq L$, the widths m_ℓ satisfy $m_\ell \leq m$. We have that*

$$\mathbb{E}[d_{CT,F_{L,\mathcal{M},k,+}(1)}(\mu,\mu_n)] \leq 256\sqrt{\frac{(L-1)km(m+1)}{n}}\left(\sqrt{(L+1)\log(2) + \frac{\log(L^2+1)}{2}} + \frac{\sqrt{\pi}}{2}\right),$$

In short, we have that $\mathbb{E}[d_{\mathrm{CT},F_{L,\mathcal{M},k,+}(1)}(\mu,\mu_n)] = O(Lm\sqrt{\frac{k}{n}})$. Hence, if we take the number of samples n larger than k times the squared product of width and depth, we can make the surrogate CT distance small. Theorem 6, proven in App. D, is based on a Rademacher complexity bound for the space $F_{L,\mathcal{M},k}(1)$ of maxout networks which may be of independent interest; to our knowledge, existing Rademacher complexity bounds for maxout networks are restricted to depth-two networks (Balazs et al., 2015; Kontorovich, 2018).

Theorems 5 and 6 show the advantages of the surrogate CT distance over the CT distance are not only computational but also statistical; the CT distance is such a strong metric that moderate-size empirical versions of a distribution are always very far from it. Hence, it is not a good criterion to compare how close an empirical distribution is to a population distribution. In contrast, the surrogate CT distance between a distribution and its empirical version is small for samples of moderate size. An analogous observation for the Wasserstein distance versus the neural net distance was made by Arora et al. (2017).

If μ_n, ν_n are empirical versions of μ, ν, it is also interesting to bound $|d_{\mathrm{CT},F_{L,\mathcal{M},k,+}(1)}(\mu_n,\nu_n) - d_{\mathrm{CT},\mathcal{A}}(\mu,\nu)| \leq |d_{\mathrm{CT},F_{L,\mathcal{M},k,+}(1)}(\mu_n,\nu_n) - d_{\mathrm{CT},F_{L,\mathcal{M},k,+}(1)}(\mu,\nu)| + |d_{\mathrm{CT},F_{L,\mathcal{M},k,+}(1)}(\mu,\nu) - d_{\mathrm{CT},\mathcal{A}}(\mu,\nu)|$. The first term has a $O(\sqrt{k/n})$ bound following from Theorem 6, while the second term is upper-bounded by $2\sup_{f\in\mathcal{A}} \inf_{\tilde{f}\in F_{L,\mathcal{M},k}(1)} \|f - \tilde{f}\|_\infty$, which is (twice) the approximation error

of the class $\mathcal{A}$ by the class $F_{L,\mathcal{M},k}(1)$. Such bounds have only been derived in $L = 2$: Balazs et al. (2015) shows that $\sup_{f\in\mathcal{A}} \inf_{\tilde{f}\in F_{2,(d,1),k}(1)} \|f - \tilde{f}\|_\infty = O(dk^{-2/d})$. Hence, we need k exponential in d to make the second term small, and thus $n \gg k$ exponential in d to make the first term small.

6 Learning distributions with the surrogate VDC and CT distance

We provide a min-max framework to learn distributions with stochastic orders. As in the generative adversarial network (GAN, Goodfellow et al. (2014); Arjovsky et al. (2017)) framework, we parametrize probability measures implicitly as the pushforward $\mu = g_\#\mu_0$ of a base measure μ_0 by a generator function g in a parametric class $\mathcal{G}$ and we optimize over g. The loss functions involve a maximization over ICMNs corresponding to the computation of a surrogate VDC or CT distance (and possibly additional maximization problems), yielding a min-max problem analogous to GANs.

Enforcing dominance constraints with the surrogate VDC. In some applications, we want to optimize a loss $L : \mathcal{P}(\Omega) \to \mathbb{R}$ under the constraint that $\mu = g_\#\mu_0$ dominates a baseline measure ν. We can enforce, or at least, bias μ towards the dominance constraint by adding a penalization term proportional to the surrogate VDC between μ and ν, in application of Proposition 2.

A first instance of this approach appears in portfolio optimization (Xue et al., 2020; Post et al., 2018). Let $\xi = (\xi_1, \dots, \xi_p)$ be a random vector of return rates of p assets and let $Y_1 := G_1(\xi), Y_2 := G_2(\xi)$ be real-valued functions of ξ which represent the return rates of two different asset allocations or portfolios, e.g. $G_i(\xi) = \langle \omega_i, \xi\rangle$ with $\omega_i \in \mathbb{R}^p$. The goal is to find a portfolio G_2 that enhances a benchmark portfolio G_1 in a certain way. For a portfolio G with return rate $Y := G(\xi)$, we let $F_Y^{(1)}(x) = P_\xi(Y \le x)$ be the CDF of its return rate, and $F_Y^{(2)}(x) = \int_{-\infty}^x F_Y^{(1)}(x')\,dx'$. If Y_1, Y_2 are the return rates of G_1, G_2, we say that Y_2 dominates Y_1 in second order, or $Y_2 \succeq_2 Y_1$ if for all $x \in \mathbb{R}$, $F_{Y_2}^{(2)}(x) \le F_{Y_1}^{(2)}(x)$, which intuitively means that the return rates Y_2 are less spread out than those Y_1, i.e. the risk is smaller. Formally, the portfolio optimization problem can be written as:

$$\max_{G_2} \ \mathbb{E}[Y_2 := G_2(\xi)], \qquad \text{s.t. } Y_2 \succeq_2 Y_1 := G_1(\xi). \tag{9}$$

It turns out that $Y_2 \succeq_2 Y_1$ if and only if $\mathbb{E}[(\eta - Y_2)_+] \le \mathbb{E}[(\eta - Y_1)_+]$ for any $\eta \in \mathbb{R}$, or yet equivalently, if $\mathbb{E}[u(Y_2)] \ge \mathbb{E}[u(Y_1)]$ for all concave non-decreasing $u : \mathbb{R} \to \mathbb{R}$ (Dentcheva & Ruszczyński, 2004). Although different, note that the second order is intimately connected to the Choquet order for 1-dimensional distributions, and it can be handled with similar tools. Define $F_{L,\mathcal{M},k,-+}(1)$ as the subset of $F_{L,\mathcal{M},k,+}(1)$ such that the first m_1 components of the weights $w_{i,j}^{(1)}$ are non-positive for all $1 \le i \le m_2$, $1 \le j \le k$. If we set the input width $m_1 = 1$, we can encode the condition $Y_2 \succeq_2 Y_1$ as $\mathrm{VDC}_{F_{L,\mathcal{M},k,-+}(1)}(\nu||\mu)$, where $\nu = \mathcal{L}(Y_1)$ and $\mu = \mathcal{L}(Y_2)$ are the distributions of Y_1, Y_2, resp. Hence, with the appropriate Lagrange multiplier we convert problem (9) into a min-max problem between μ and the potential u of VDC

$$\min_{\mu:\mu=\mathcal{L}(\langle\xi,\omega_2\rangle)} \ -\int_{\mathbb{R}} x\,d\mu(x) + \lambda \mathrm{VDC}_{F_{L,\mathcal{M},k,-+}(1)}(\nu||\mu). \tag{10}$$

A second instance of this approach is in GAN training. Assuming that we have a baseline generator g_0 that can be obtained via regular training, we consider the problem:

$$\min_{g\in\mathcal{G}} \left\{ \max_{f\in\mathcal{F}} \{\mathbb{E}_{X\sim\nu_n}[f(X)] - \mathbb{E}_{Y\sim\mu_0}[f(g(Y))]\} + \lambda \mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(g_\#\mu_0||(g_0)_\#\mu_0) \right\}. \tag{11}$$

The first term in the objective function is the usual WGAN loss (Arjovsky et al., 2017), although it can be replaced by any other standard GAN loss. The second term, which is proportional to $\mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(g_\#\mu_0||(g_0)_\#\mu_0) = \max_{u\in F_{L,\mathcal{M},k,+}(1)}\{\mathbb{E}_{Y\sim\mu_0}[u(g_0(Y))] - \mathbb{E}_{Y\sim\mu_0}[u(g(Y))]\}$, enforces that $g_\#\mu_0 \succeq (g_0)_\#\mu_0$ in the Choquet order, and thus u acts as a second 'Choquet' critic. Tuning λ appropriately, the rationale is that we want a generator that optimizes the standard GAN loss, with the condition that the generated distribution dominates the baseline distribution. As stated by Proposition 1, dominance in the Choquet order translates to $g_\#\mu_0$ being more spread out than $(g_0)_\#\mu_0$, which should help avoid mode collapse and improve the diversity of generated samples. In practice, this min-max game is solved via Algorithm 1 given in App. F. For the Choquet critic, this amounts to an SGD step followed by a projection step to impose non-negativity of hidden to hidden weights.

Generative modeling with the surrogate CT distance. The surrogate Choquet-Toland distance is well-suited for generative modeling, as it can used in GANs in place of the usual discriminator. Namely, if ν is a target distribution, ν_n is its empirical distribution, and $\mathcal{D} = \{\mu = g_{\#}\mu_0 | g \in \mathcal{G}\}$ is a class of distributions that can be realized as the push-forward of a base measure $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$ by a function $g : \mathbb{R}^{d_0} \to \mathbb{R}^d$, the problem to solve is $g^* = \arg\min_{g\in\mathcal{G}} d_{\text{CT},F_{L,\mathcal{M},k,+}(1)}(g_{\#}\mu_0, \nu_n) = \arg\min_{g\in\mathcal{G}}\{\max_{u\in F_{L,\mathcal{M},k,+}(1)}\{\mathbb{E}_{X\sim\nu_n}[u(X)] - \mathbb{E}_{Y\sim\mu_0}[u(g(Y))]\} + \max_{u\in F_{L,\mathcal{M},k,+}(1)}\{\mathbb{E}_{Y\sim\mu_0}[u(g(Y))] - \mathbb{E}_{X\sim\nu_n}[u(X)]\}\}$. Algorithm 2 given in App. F summarizes learning with the surrogate CT distance.

7 Experiments

Portfolio optimization under dominance constraints In this experiment, we use the VDC to optimize an illustrative example from Xue et al. (2020) (Example 1) that follows the paradigm laid out in Sec. 6. In this example, $\xi \in \mathbb{R} \sim P$ is drawn uniformly from $[0, 1]$, we define the benchmark portfolio as:

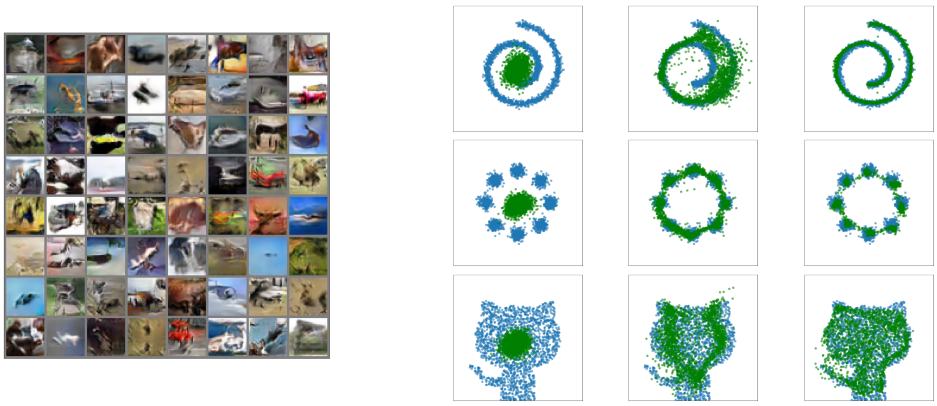
$$G_1(\xi) = \begin{cases} \frac{i}{20} & \xi \in [0.05 \times i, 0.05 \times (i+1)) \quad i = 0, \dots, 19 \\ 1 & \xi = 1 \end{cases}$$

and the optimization is over the parameterized portfolio $G_2(\xi) := G(\xi; z) = z\xi$.

The constrained optimization problem is thus specified as:

$$\min_z \; -\mathbb{E}_P[G(\xi; z)] \;\; \text{s.t.} \;\; G(\xi; z) \succeq_2 G_1(\xi), \; 1 \le z \le 2$$

As stated in Xue et al. (2020), this example has a known solution at $z = 2$ where $\mathbb{E}_P[G_2(\xi; 2)] = 1$ outperforms the benchmark $\mathbb{E}_P[G_1(\xi)] = 0.5$. We relax the constrained optimization by including it in the objective function, thus creating min-max game (10) introduced in Sec. 6. We parameterize $F_{L,\mathcal{M},k,-+}(1)$ with a 3-layer, fully-connected, decreasing ICMN with hidden dimension 32 and maxout kernel size of 4. After 5000 steps of stochastic gradient descent on z (learning rate 1e^{-3}) and the parameters of the ICMN (learning rate 1e^{-3}), using a batch size of 512 and $\lambda = 1$, we are able to attain accurate approximate values of the known solution: $z = 2$, $\frac{1}{512}\sum_{j=1}^{512}[G_z(\xi_j)] = 1.042$, and $\frac{1}{512}\sum_{j=1}^{512}[G_1(\xi_j)] = 0.496$.



(a) g^* CIFAR-10 samples (b) Initialisation (c) 5k iterations (d) 80k iterations

Figure 2: Training generative models by enforcing dominance with surrogate VDC on pre-trained CIFAR-10 WGAN-GP (*Left*) and with surrogate CT distance on 2D point clouds (*Right)*. Ground truth point cloud distributions (*blue*) consist of swiss roll (*Top*), circle of eight gaussians (*Middle*), and Github icon converted to a point cloud (*Bottom*).

Image generation with baseline model dominance Another application of learning with the VDC is in the high-dimensional setting of CIFAR-10 (Krizhevsky & Hinton, 2009) image generation. As detailed in Sec. 6, we start by training a baseline generator g_0 using the regularized Wasserstein-GAN paradigm (WGAN-GP) introduced in Arjovsky et al. (2017); Gulrajani et al. (2017), where

gradient regularization is computed for the discriminator with respect to interpolates between real and generated data. When training g^* and g_0, we used the same WGAN-GP hyperparameter configuration. We set λ in Equation (11) to 10 (see App. F and App. H for more details). Training runs for g^* and g_0

Table 1: FID scores for WGAN-GP and WGAN-GP with VDC surrogate for convex functions approximated by either ICNNs with softplus activations or ICMNs. ICMNs improve upon the baseline g_0 and outperform ICNNs with softplus. FID score for WGAN-GP + VDC includes mean values $\pm$ one standard deviation for 5 repeated runs with different random initialization seeds.

	FID
g_0: WGAN-GP	69.67
g^*: WGAN-GP + VDC CP-Flow ICNN	83.470 $\pm$ 3.732
g^*: WGAN-GP + VDC ICMN (Ours)	**67.317** $\pm$ 0.776

were performed in computational environments that contained 1 CPU and 1 A100 GPU. To evaluate performance of g^* vs. g_0, we rely on the Fréchet Inception Distance (FID) introduced in Heusel et al. (2017). Note that when training a WGAN-GP baseline from scratch we used hyperparameters that potentially differ from those used in state-of-the-art implementations. Additionally, for computing FIDs we use a `pytorch-lightning` implementation of the inception network, which is different from the widely used `Tensorflow` implementation (Salimans et al., 2016), resulting in potential discrepancies in our reported baseline FID and those in the literature. FID results are reported in Table 1, where we see improved image quality from g^* (as measured by lower FID) relative to the pre-trained baseline g_0. We therefore find that the VDC surrogate improves upon g_0 by providing g^* with larger support, preventing mode collapse. Samples generated from g^* are displayed in Figure 2a.

In order to highlight the representation power of ICMNs, we replace them in the VDC estimation by the ICNN implementation of Huang et al. (2021). Instead of maxout activation, Huang et al. (2021) uses a Softplus activation and instead of a projection step it also uses a Softplus operation to impose the non-negativity of hidden to hidden weights to enforce convexity. We see in Table 1 that VDC's estimation with ICMN outperforms Huang et al. (2021) ICNNs. While ICMNs have maximum of affine functions as a building block, ICNN's proof of universality in Huang et al. (2021) relies on approximating the latter, this could be one of the reason behind ICMN superiority.

Probing mode collapse To investigate how training with the surrogate VDC regularizer helps alleviate mode collapse in GAN training, we implemented GANs trained with the IPM objective alone and compared this to training with the surrogate VDC regularizer for a mixture of 8 Gaussians target distribution. In Figure 3 given in App. G we quantify mode collapse by looking at two scores: 1) the entropy of the discrete assignment of generated points to the means of the mixture 2) the negative log likelihood (NLL) of the Gaussian mixture. When training with the VDC regularizer to improve upon the collapsed generator g_0 (which is taken from step 55k from the unregularized GAN training), we see more stable training and better mode coverage as quantified by our scores.

2D point cloud generation with d_{CT} We apply learning with the CT distance in a 2D generative modeling setting. Both the generator and CT critic architectures are comprised of fully-connected neural networks with maxout non-linearities of kernel size 2. Progression of the generated samples can be found in the right-hand panel of Figure 2, where we see the trained generator accurately learn the ground truth distribution. All experiments were performed in a single-CPU compute environment.

8 Conclusion

In this paper, we introduced learning with stochastic order in high dimensions via surrogate Variational Dominance Criterion and Choquet-Toland distance. These surrogates leverage input convex maxout networks, a new variant of input convex neural networks. Our surrogates have parametric statistical rates and lead to new learning paradigms by incorporating dominance constraints that improve upon a baseline. Experiments on synthetic and real image generation yield promising results. Finally, our work, although theoretical in nature, can be subject to misuse, similar to any generative method.

REFERENCES

David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. Optimizing functionals on the space of probabilities with input convex neural networks, 2021.

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 06–11 Aug 2017.

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

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 224–232. JMLR, 2017.

Gabor Balazs, András György, and Csaba Szepesvari. Near-optimal max-affine estimators for convex regression. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pp. 56–64. PMLR, 2015.

Efim M. Bronshtein. ϵ-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17: 393–398, 1976.

Charlotte Bunne, Laetitia Meng-Papaxanthos, Andreas Krause, and Marco Cuturi. Jkonet: Proximal optimal transport modeling of population dynamics, 2021.

G. Carlier. Remarks on toland's duality, convexity constraint and optimal transport. *Pacific Journal of Optimization*, 4, 09 2008.

Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

Darinka Dentcheva and Andrzej Ruszczynski. Optimization with stochastic dominance constraints. *SIAM Journal on Optimization*, 14(2):548–566, 2003.

Darinka Dentcheva and Andrzej Ruszczyński. Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints. *Mathematical Programming*, 99:329–350, 01 2004. doi: 10.1007/s10107-003-0453-z.

R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Ivar Ekeland and Walter Schachermayer. Optimal transport and the geometry of $L^1(\mathbb{R}^d)$. *Proceedings of the American Mathematical Society*, 142, 10 2014.

William Falcon et al. Pytorch lightning. *GitHub. Note: https://github. com/PyTorchLightning/pytorch-lightning*, 3:6, 2019.

Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 1319–1327. PMLR, 2013.

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.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Adityanand Guntuboyina and Bodhisattva Sen. Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965, 2013.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Clifford Hildreth. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49(267):598–619, 1954.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Aryeh Kontorovich. Rademacher complexity of k-fold maxima of hyperplanes, 2018.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021a.

Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*, 2021b.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In Hal Daumé Iii And Singh (ed.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6672–6681. PMLR, 2020.

John D. Martin, Michal Lyskawinski, Xiaohu Li, and Brendan Englot. Stochastically dominant distributional reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. In *ICLR*, 2017.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows, 2021.

Youssef Mroueh and Tom Sercu. Fisher gan. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Alfred Müller. Stochastic orders generated by integrals: A unified study. *Advances in Applied Probability*, 29(2):414–428, 1997b.

Alfred Müller. Stochastic ordering of multivariate normal distributions. *Annals of the Institute of Statistical Mathematics*, 53:567–575, 02 2001.

Inkit Padhi, Pierre Dognin, Ke Bai, Cicero Nogueira dos Santos, Vijil Chenthamarakshan, Youssef Mroueh, and Payel Das. Learning implicit text generation via feature matching. *arXiv preprint arXiv:2005.03588*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Thierry Post, Selçuk Karabatı, and Stelios Arvanitis. Portfolio optimization based on stochastic dominance and empirical likelihood. *Journal of Econometrics*, 206(1):167–186, 2018. ISSN 0304-4076.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Shashank Singh and Barnabás Póczos. Minimax distribution estimation in wasserstein distance, 2018.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On integral probability metrics, φ-divergences and binary classification, 2009.

V. N. Sudakov. A remark on the criterion of continuity of gaussian sample function. In *Proceedings of the Second Japan-USSR Symposium on Probability Theory*, pp. 444–454, Berlin, Heidelberg, 1973. Springer Berlin Heidelberg.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Meng Xue, Yun Shi, and Hailin Sun. Portfolio optimization with relaxation of stochastic second order dominance constraints via conditional value at risk. *Journal of Industrial and Management Optimization*, 16(6):2581–2602, 2020.