Learning with stochastic orders

Carles Domingo-Enrich[a], Yair Schiff[b] and Youssef Mroueh[c]
[a]New York University, [b]Cornell University, [c] IBM Research AI

ICLR May, 2023

Generative adversarial networks

- The goal of generative modeling is to be able to generate artificial samples from a distribution given a sample $(X_i)_{i=1}^n$ from it.
- **Generative adversarial networks** (GANs) (Goodfellow et al., 2014) are a popular generative modeling technique where two deep neural networks, the generator g and the discriminator f, are trained adversarially.

Generative adversarial networks

- The goal of generative modeling is to be able to generate artificial samples from a distribution given a sample $(X_i)_{i=1}^n$ from it.
- **Generative adversarial networks** (GANs) (Goodfellow et al., 2014) are a popular generative modeling technique where two deep neural networks, the generator g and the discriminator f, are trained adversarially.
- A common choice for the training loss (Arjovsky et al., 2017) is:

$$\min_{g\in\mathcal{G}}\left\{\max_{f\in\mathcal{F}}\{\mathbb{E}_{X\sim\nu_n}[f(X)]-\mathbb{E}_{Y\sim\mu_0}[f(g(Y))]\}\right\}, \qquad \text{where } \nu_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}. \quad (1)$$

- A usual failure mode of GANs is **mode collapse**: the generator fails to capture entire modes of the data distribution.

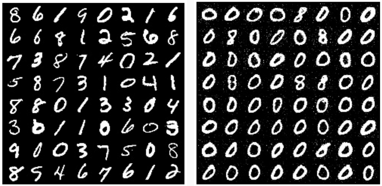


Figure 1: (*Left*) Samples from the MNIST dataset. (*Right*) GAN-generated samples suffering from mode collapse.

Generative adversarial networks

- The goal of generative modeling is to be able to generate artificial samples from a distribution given a sample $(X_i)_{i=1}^n$ from it.
- **Generative adversarial networks** (GANs) (Goodfellow et al., 2014) are a popular generative modeling technique where two deep neural networks, the generator g and the discriminator f, are trained adversarially.
- A common choice for the training loss (Arjovsky et al., 2017) is:

$$\min_{g\in\mathcal{G}} \left\{ \max_{f\in\mathcal{F}} \{\mathbb{E}_{X\sim\nu_n}[f(X)] - \mathbb{E}_{Y\sim\mu_0}[f(g(Y))]\} \right\}, \qquad \text{where } \nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}. \quad (1)$$

- A usual failure mode of GANs is **mode collapse**: the generator fails to capture entire modes of the data distribution.

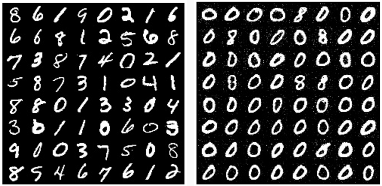


Figure 1: (*Left*) Samples from the MNIST dataset. (*Right*) GAN-generated samples suffering from mode collapse.

- **Question**: How can we modify the GAN objective to prevent mode collapse? Let's look at stochastic orders first!

Stochastic orders

- Can we compare probability measures beyond equality? $\Longrightarrow$ *stochastic orders*

Stochastic orders

- Can we compare probability measures beyond equality? $\implies$ *stochastic orders*

Definition (Convex or Choquet order, Ekeland and Schachermayer (2014))

For μ_-, μ_+ probability measures, we say that we say μ_+ dominates μ_- in the convex order, or $\mu_- \preceq_{cx} \mu_+$, if for any **convex function** $u : \mathbb{R}^d \to \mathbb{R}$, we have

$$\mathbb{E}_{x\sim\mu_-} u(x) \leq \mathbb{E}_{x\sim\mu_+} u(x).$$

Stochastic orders

- Can we compare probability measures beyond equality? $\implies$ *stochastic orders*

Definition (Convex or Choquet order, Ekeland and Schachermayer (2014))

For μ_-, μ_+ probability measures, we say that we say μ_+ dominates μ_- in the convex order, or $\mu_- \preceq_{cx} \mu_+$, if for any **convex function** $u : \mathbb{R}^d \to \mathbb{R}$, we have

$$\mathbb{E}_{x\sim\mu_-} u(x) \leq \mathbb{E}_{x\sim\mu_+} u(x).$$

- $\preceq_{cx}$ is a partial order, meaning that reflexivity, antisymmetry and transitivity hold.
- The space of convex functions is not the only choice to define orders (other *cones* can be considered).

Stochastic orders

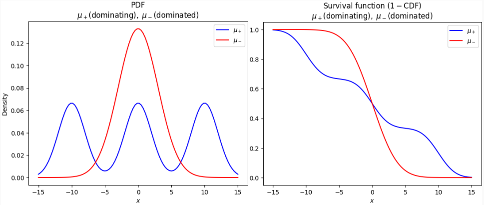
- Can we compare probability measures beyond equality? $\implies$ *stochastic orders*

Definition (Convex or Choquet order, Ekeland and Schachermayer (2014))

For μ_-, μ_+ probability measures, we say that we say μ_+ dominates μ_- in the convex order, or $\mu_- \preceq_{cx} \mu_+$, if for any **convex function** $u : \mathbb{R}^d \to \mathbb{R}$, we have

$$\mathbb{E}_{x\sim\mu_-} u(x) \leq \mathbb{E}_{x\sim\mu_+} u(x).$$

- $\preceq_{cx}$ is a partial order, meaning that reflexivity, antisymmetry and transitivity hold.
- The space of convex functions is not the only choice to define orders (other *cones* can be considered).
- The convex order in one dimension admits a characterization in terms of the integral of the CDF.



Stochastic orders in terms of Markov kernels

Proposition (Ekeland and Schachermayer (2014))

We have $\mu_- \preceq_{cx} \mu_+$ if and only if there exists a martingale Markov kernel R (i.e. $\int_{\mathbb{R}^d} y\, dR_x(y) = x, \forall x$) such that $\mu_- = \int_{\mathbb{R}^d} R_x \, d\mu_+$.

Stochastic orders in terms of Markov kernels

Proposition (Ekeland and Schachermayer (2014))

We have $\mu_- \preceq_{cx} \mu_+$ if and only if there exists a martingale Markov kernel R (i.e. $\int_{\mathbb{R}^d} y \, dR_x(y) = x, \forall x$) such that $\mu_- = \int_{\mathbb{R}^d} R_x \, d\mu_+$.

- This characterization is difficult to check, especially in high dimensions.
- Intuitively, this means that μ_- is more *spread out* than μ_+.

Variational Dominance Criterion (VDC)

Definition (Variational Dominance Criterion (VDC))

Given a bounded open convex subset $\Omega \subseteq \mathbb{R}^d$, a pair of Borel probability measures $\mu_+, \mu_- \in \mathcal{P}(\Omega)$, and a compact set $K \subseteq \mathbb{R}^d$ $(0 \in K)$, define:

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

where $\mathcal{A} = \{u : \Omega \to \mathbb{R}, \;$ u convex and $\nabla u \in K$ almost everywhere$\}$.

Remark that since $0 \in \mathcal{K}$, $\mathrm{VDC}_{\mathcal{A}}(\mu_+||\mu_-) \geq 0$ for all μ_+, μ_- because the zero function belongs to the set $\mathcal{A}$.

Variational Dominance Criterion (VDC)

Definition (Variational Dominance Criterion (VDC))

Given a bounded open convex subset $\Omega \subseteq \mathbb{R}^d$, a pair of Borel probability measures $\mu_+, \mu_- \in \mathcal{P}(\Omega)$, and a compact set $K \subseteq \mathbb{R}^d$ ($0 \in K$), define:

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

where $\mathcal{A} = \{u : \Omega \to \mathbb{R}, \quad u \text{ convex and } \nabla u \in K \text{ almost everywhere}\}$.

Remark that since $0 \in \mathcal{K}$, $\mathrm{VDC}_{\mathcal{A}}(\mu_+ || \mu_-) \geq 0$ for all μ_+, μ_- because the zero function belongs to the set $\mathcal{A}$.

Proposition

$$\mu_- \preceq_{cx} \mu_+ \quad \Longleftrightarrow \quad \mathrm{VDC}_{\mathcal{A}}(\mu_+ || \mu_-) = 0$$

- Intuition: $\mathrm{VDC}_{\mathcal{A}}(\mu_+ || \mu_-) = 0 \quad \Longleftrightarrow \quad \mathbb{E}_{x \sim \mu_-} u(x) \leq \mathbb{E}_{x \sim \mu_+} u(x)$ for all $u \in \mathcal{A}$ $\Longleftrightarrow \quad \mathbb{E}_{x \sim \mu_-} u(x) \leq \mathbb{E}_{x \sim \mu_+} u(x)$ for all u convex.

Variational Dominance Criterion (VDC)

Definition (Variational Dominance Criterion (VDC))

Given a bounded open convex subset $\Omega \subseteq \mathbb{R}^d$, a pair of Borel probability measures $\mu_+, \mu_- \in \mathcal{P}(\Omega)$, and a compact set $K \subseteq \mathbb{R}^d$ $(0 \in K)$, define:

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

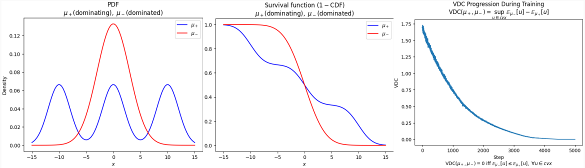
where $\mathcal{A} = \{u : \Omega \to \mathbb{R},\ \ u \text{ convex and } \nabla u \in K \text{ almost everywhere}\}$.

Remark that since $0 \in \mathcal{K}$, $\mathrm{VDC}_{\mathcal{A}}(\mu_+ || \mu_-) \geq 0$ for all μ_+, μ_- because the zero function belongs to the set $\mathcal{A}$.

Proposition

$$\mu_- \preceq_{cx} \mu_+ \iff \mathrm{VDC}_{\mathcal{A}}(\mu_+ || \mu_-) = 0$$

- Intuition: $\mathrm{VDC}_{\mathcal{A}}(\mu_+ || \mu_-) = 0 \iff \mathbb{E}_{x \sim \mu_-} u(x) \leq \mathbb{E}_{x \sim \mu_+} u(x)$ for all $u \in \mathcal{A}$ $\iff \mathbb{E}_{x \sim \mu_-} u(x) \leq \mathbb{E}_{x \sim \mu_+} u(x)$ for all u convex.
- Informally, the proposition implies that $\mathrm{VDC}_{\mathcal{A}}(\mu_+ || \mu_-)$ *is small when* μ_+ *is more spread out than* μ_-*, and large otherwise.*



Input Convex Maxout Networks

- **Problem**: Statistical rates of estimation of the VDC are cursed by dimension, i.e. $|\mathrm{VDC}_K(\mu_+||\mu_-) - \mathrm{VDC}_K(\mu_{+,n}||\mu_{-,n})| \lesssim Cn^{-2/d}$. The set of convex functions is too *large* (its Rademacher complexity scales like $n^{-2/d}$).

Input Convex Maxout Networks

- **Problem**: Statistical rates of estimation of the VDC are cursed by dimension, i.e. $|\mathrm{VDC}_K(\mu_+||\mu_-) - \mathrm{VDC}_K(\mu_{+,n}||\mu_{-,n})| \lesssim Cn^{-2/d}$. The set of convex functions is too *large* (its Rademacher complexity scales like $n^{-2/d}$).
- Idea: Approximate convex functions with bounded gradients using neural networks

Input Convex Maxout Networks

- **Problem**: Statistical rates of estimation of the VDC are cursed by dimension, i.e. $|\mathrm{VDC}_K(\mu_+||\mu_-) - \mathrm{VDC}_K(\mu_{+,n}||\mu_{-,n})| \lesssim Cn^{-2/d}$. The set of convex functions is too *large* (its Rademacher complexity scales like $n^{-2/d}$).
- Idea: Approximate convex functions with bounded gradients using neural networks
- Previous work: Input Convex Neural Networks (Amos et al., 2017). But we can do better in our setting!

Input Convex Maxout Networks

- **Problem**: Statistical rates of estimation of the VDC are cursed by dimension, i.e. $|\mathrm{VDC}_K(\mu_+||\mu_-) - \mathrm{VDC}_K(\mu_{+,n}||\mu_{-,n})| \lesssim Cn^{-2/d}$. The set of convex functions is too *large* (its Rademacher complexity scales like $n^{-2/d}$).
- Idea: Approximate convex functions with bounded gradients using neural networks
- Previous work: Input Convex Neural Networks (Amos et al., 2017). But we can do better in our setting!
- Idea: maximum of affine functions are good approximations of convex functions. Can we stack them in layers? Yes $\implies$ Input Convex Maxout Networks.

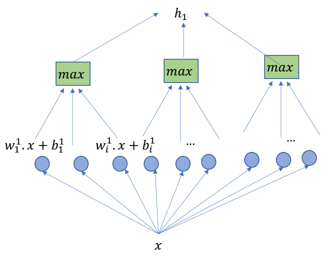


Figure 2: Shallow maxout network. ICMNs are maxout networks with convex increasing activations such that all weights beyond the first layer are non-negative.

Surrogate VDC via ICMNs

- $F_{L,\mathcal{M},k,+}(1)$: set of ICMNs with fixed architecture and bound on weights, such that $F_{L,\mathcal{M},k}(1) \subseteq \mathcal{A}$.

Surrogate VDC via ICMNs

- $F_{L,\mathcal{M},k,+}(1)$: set of ICMNs with fixed architecture and bound on weights, such that $F_{L,\mathcal{M},k}(1) \subseteq \mathcal{A}$.
- We replace $\mathcal{A} = \{u : \Omega \to \mathbb{R},\ \text{u convex and } \nabla u \in K \text{ a.e.}\}$ by $F_{L,\mathcal{M},k}(1)$, and obtain the surrogate VDC:

$$\mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(\mu_+ || \mu_-) = \sup_{u \in F_{L,\mathcal{M},k,+}(1)} \int_\Omega u \, d(\mu_- - \mu_+). \tag{2}$$

Surrogate VDC via ICMNs

- $F_{L,\mathcal{M},k,+}(1)$: set of ICMNs with fixed architecture and bound on weights, such that $F_{L,\mathcal{M},k}(1) \subseteq \mathcal{A}$.
- We replace $\mathcal{A} = \{u : \Omega \to \mathbb{R},\ \text{u convex and } \nabla u \in K \text{ a.e.}\}$ by $F_{L,\mathcal{M},k}(1)$, and obtain the surrogate VDC:

$$\mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(\mu_+||\mu_-) = \sup_{u \in F_{L,\mathcal{M},k,+}(1)} \int_\Omega u \, d(\mu_- - \mu_+). \quad (2)$$

- The surrogate VDC solves two problems at once:
  - It enjoys parametric estimation rates: $|\mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(\mu_+||\mu_-) - \mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(\mu_{+,n}||\mu_{-,n})| \lesssim Cn^{-1/2}$.
  - We can use gradient descent to solve the variational problem (2) (no guarantees, but it works in practice).

Training GANs with surrogate VDC penalization

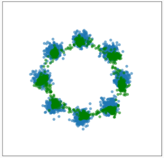
- We take a base generator g_0 trained using the baseline GAN training loss, and consider the problem:

$$\min_{g \in \mathcal{G}} \left\{ \max_{f \in \mathcal{F}} \{ \mathbb{E}_{X \sim \nu_n}[f(X)] - \mathbb{E}_{Y \sim \mu_{\mathbf{0}}}[f(g(Y))] \} + \lambda \mathrm{VDC}_{F_{L,\mathcal{M},k,+}(1)}(g_{\#}\mu_0 || (g_0)_{\#}\mu_0) \right\}. \tag{3}$$

  Here $g_{\#}\mu_0$ is the distribution of the generated samples $g(X)$, $X \sim \mu_0$.
- That is, we add the surrogate VDC between the learned and the baseline distribution: we want to bias $g_{\#}\mu_0$ to be more spread-out than $(g_0)_{\#}\mu_0$.

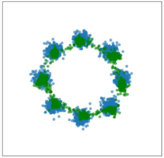
Mode collapse mitigation: mixture of Gaussians

- The target μ_r is a mixture of 8 gaussians in two dimensions
- g_0 is a mode collapsed generator
- g^* is trained with WGAN-GP penalized with the surrogate VDC.



Mode collapse mitigation: mixture of Gaussians

- The target μ_r is a mixture of 8 gaussians in two dimensions
- g_0 is a mode collapsed generator
- g^* is trained with WGAN-GP penalized with the surrogate VDC.



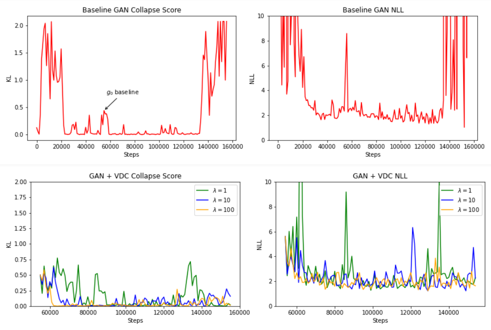
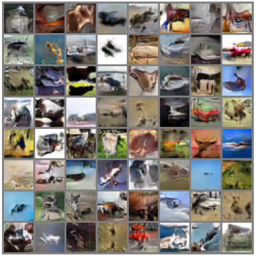


Figure 3: Probing mode collapse for GAN training. A converged generator needs to have a low negative likelihood and low mode collapse score. Collapse score: KL divergence between the discrete distribution obtained by assignment to closest centroid, and uniform distribution.

GAN experiments in high dimensions

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, on the CIFAR-10 dataset. 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 ± 3.732
g^*: WGAN-GP + VDC ICMN (Ours)	**67.317** $\pm$ 0.776



Other potential applications of our framework

- Portfolio optimization (Post et al., 2018; Xue et al., 2020): The goal is to find a portfolio G_2 that enhances a benchmark portfolio G_1 in a certain way: the return of G_2 must have high expectation, but its distribution must be less spread out than for G_1—*less risk*.
- Distributional reinforcement learning (Martin et al., 2020): We want to learn policies with dominance constraints on the distribution of the reward.

Thank you!

Contacts: cd2754@nyu.edu, yzs2@cornell.edu, mroueh@us.ibm.com

References i

Amos, B., Xu, L., and Kolter, J. Z. (2017). Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 146–155. PMLR.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Martin, J. D., Lyskawinski, M., Li, X., and Englot, B. (2020). Stochastically dominant distributional reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Post, T., Karabati, S., and Arvanitis, S. (2018). Portfolio optimization based on stochastic dominance and empirical likelihood. *Journal of Econometrics*, 206(1):167–186.

Xue, M., Shi, Y., and Sun, H. (2020). 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.