# Adversarial Divergences are Good Task Losses for Generative Modeling

**Gabriel Huang**[1] **Gauthier Gidel**[1] **Hugo Berard**[1] **Ahmed Touati**[1] **and Simon Lacoste-Julien**[1]

[1]Montreal Institute for Learning Algorithms, Université de Montréal

## Overview

### Summary

Generative modeling of high dimensional data, like images, is notoriously **difficult** and **ill-defined**. It is not obvious how to specify **relevant evaluation metrics** and **meaningful objectives** to optimize. In this work, we give arguments why **adversarial divergences** are **good objectives for generative modeling**, and perform experiments to better understand their properties.

### Contributions

- Unify structured prediction and generative adversarial networks using **statistical decision theory**. **Relate theoretical results** on structured losses with the notion of **weak** and **strong** divergences.
- Show that compared to traditional divergences, adversarial divergences are a **good objective** in terms of sample complexity, computation, ability to integrate prior knowledge, flexibility and ease of optimization.
- Show experimentally the importance of choosing a divergence that **reflects the final task**.

## Context and Motivation

### Problems with KL divergence

Maximimum Likelihood Estimation (MLE), or minimizing the Kullback-Leibler divergence $\mathbf{KL}(p||q_\theta) = \mathbf{E}_{\boldsymbol{x}\sim p}[\log \frac{p(\boldsymbol{x})}{q_\theta(\boldsymbol{x})}]$ have several drawbacks, including:

- No meaningful **training signal** when $p$ and $q_\theta$ are far away. Workarounds generally involve smoothing $q_\theta$, which makes it hard to learn sharp distributions.
- Requires evaluating $q_\theta(x)$, so **cannot be directly used with implicit models**.
- **Teacher-forcing** on autoregressive models.
- **Hard** to enforce properties that **characterize the final task**.

### Adversarial Divergences

We define **(neural) adversarial divergences** as

$$\boxed{\mathbf{Adv}\Delta(p||q_\theta) \hat{=} \sup_{\phi\in\Phi} \mathbf{E}_{(\boldsymbol{x},\boldsymbol{x}')\sim p\otimes q_\theta}[\Delta(f_\phi(\boldsymbol{x}), f_\phi(\boldsymbol{x}'))]}$$

where the choice of the discriminator neural network $f_\phi$ and function $\Delta$ determine properties of the adversarial divergence. For instance, the adversarial Jensen-Shannon from GANs writes

$$\boxed{\mathbf{AdvJS}(p||q_\theta) \hat{=} \sup_{\phi\in\Phi} \mathbf{E}_{\boldsymbol{x}\sim p}[\log f_\phi(\boldsymbol{x})] + \mathbf{E}_{\boldsymbol{x}'\sim q_\theta}[\log(1 - f_\phi(\boldsymbol{x}'))]}$$

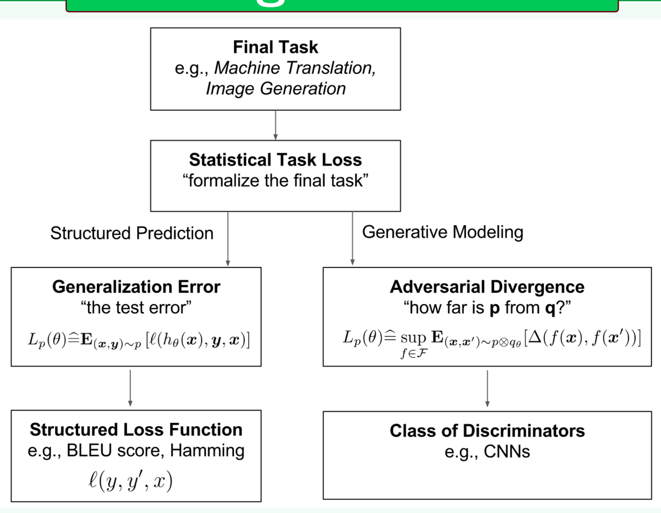
Other adversarial divergences: adversarial Wasserstein, MMD-GANs, ...

## Statistical Decision Theory Framework

### General Framework

- $\mathcal{P}$: set of possible states.
- $\mathcal{A}$: set of actions available.
- $L_p(a)$: cost of playing action $a \in \mathcal{A}$ when the current state is $p \in \mathcal{P}$.
- **Goal**: find $a \in \mathcal{A}$ minimizing the **(statistical) task loss** $L_p(a)$.

### Building a Task Loss




### MLE, Structured Prediction (SP) and GANs

| | $\mathcal{P}$ | $\mathcal{A}$ | $L_p(a)$ |
|---|---|---|---|
| MLE | $\{p(\boldsymbol{x})\}$ | $\{q_\theta\,;\,\theta\in\Theta\}$ | $\mathbf{E}_{\boldsymbol{x}\sim p}[-\log(q_\theta(\boldsymbol{x}))]$ |
| SP | $\{p(\boldsymbol{x},\boldsymbol{y})\}$ | $\{h_\theta\,;\,\theta\in\Theta\}$ | $\mathbf{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p}[\ell(h_\theta(\boldsymbol{x}),\boldsymbol{y},\boldsymbol{x})]$ |
| GAN | $\{p(\boldsymbol{x})\}$ | $\{q_\theta\,;\,\theta\in\Theta\}$ | $\sup_{f\in\mathcal{F}}\mathbf{E}_{(\boldsymbol{x},\boldsymbol{x}')\sim p\otimes q_\theta}[\Delta(f(\boldsymbol{x}), f(\boldsymbol{x}'))]$ |

where $\ell : \mathcal{Y}\times\mathcal{Y}\times\mathcal{X} \to \mathbb{R}$ is a structured loss function, while the class of discriminators $\mathcal{F}$ and $\Delta : \mathbb{R}^{d'}\times\mathbb{R}^{d'} \to \mathbb{R}$ determine properties of the adversarial divergence.

### Consequences

- Analogy between choice of structured loss $\ell$ and class of discriminators $\mathcal{F}$ in order to build a statistical task losses that **reflect the final task**.
- Insights from **theoretical structured prediction** (Osokin et al. [1]).

## Results by Osokin et al. [1]

### Intuition

- Strong losses such as the 0-1 loss are **hard to learn** because they do **not** give any **flexibility** on the prediction. We roughly need as many training examples as $|\mathcal{Y}|$, which is **exponential** in the dimension of $y$.
- Conversely, weaker losses like the Hamming loss have **more flexibility**; because they tell us how close a prediction is to the ground truth, **less example** are needed to generalize well.

### Theory to Back the Intuition

Formalize the intuition and compare the 0-1 loss to the Hamming loss,

$$\ell_{0-1}(\boldsymbol{y},\boldsymbol{y}') \hat{=} \mathbf{1}\{\boldsymbol{y}\neq\boldsymbol{y}'\}\,, \qquad \ell_{Ham}(\boldsymbol{y},\boldsymbol{y}') \hat{=} \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{\boldsymbol{y}_t\neq\boldsymbol{y}'_t\}$$

when $\boldsymbol{y}$ decomposes as $T = \log_2|\mathcal{Y}|$ binary variables $(y_t)_{1\le t\le T}$. They derive a **worst case** sample complexity to get an error $\epsilon > 0$ and obtain,

- For 0-1 loss: $O(|Y|/\epsilon^2)$ (**exponential**). $\Rightarrow$ **BAD!**
- For Hamming loss[a]: $O(\log_2|\mathcal{Y}|/\epsilon^2)$ (**polynomial**) $\Rightarrow$ **GOOD!**

[a]under certain constraints, see [1]

### Insights

**Flexible** statistical task losses, which can "smoothly" distinguish between good and bad models, are easier to optimize in the context of structured prediction, which can be related to the belief that **weaker adversarial divergences** are **easier to optimize** in generative modeling.

## Adversarial vs. Traditional Divergences

### Statistical and computational properties

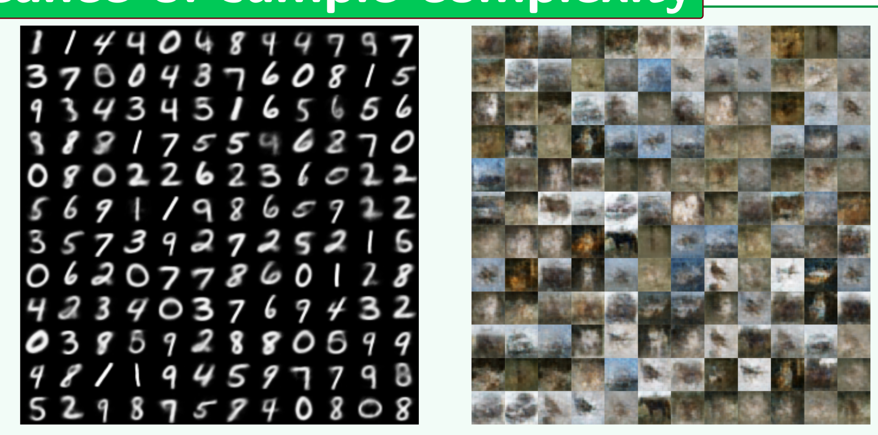
| Divergence | Sample Comp. | Computation | Integrate Final Loss |
|---|---|---|---|
| f-Div (EXPL) | $O(1/\epsilon^2)$ | MC, $O(n)$ | no |
| f-Div (IMPL) | N/A | N/A | N/A |
| Wasserstein | $O(1/\epsilon^{d+1})$ | Sinkhorn, $O(n^2)$ | in base distance |
| MMD | $O(1/\epsilon^2)$ | analytic, $O(n^2)$ | in kernel |
| Adversarial | $O(p/\epsilon^2)$ | SGD | in discriminator |

EXPL and IMPL stand for explicit and implicit models, and $p$ is the VC- dimension/number of parameters of the discriminator.

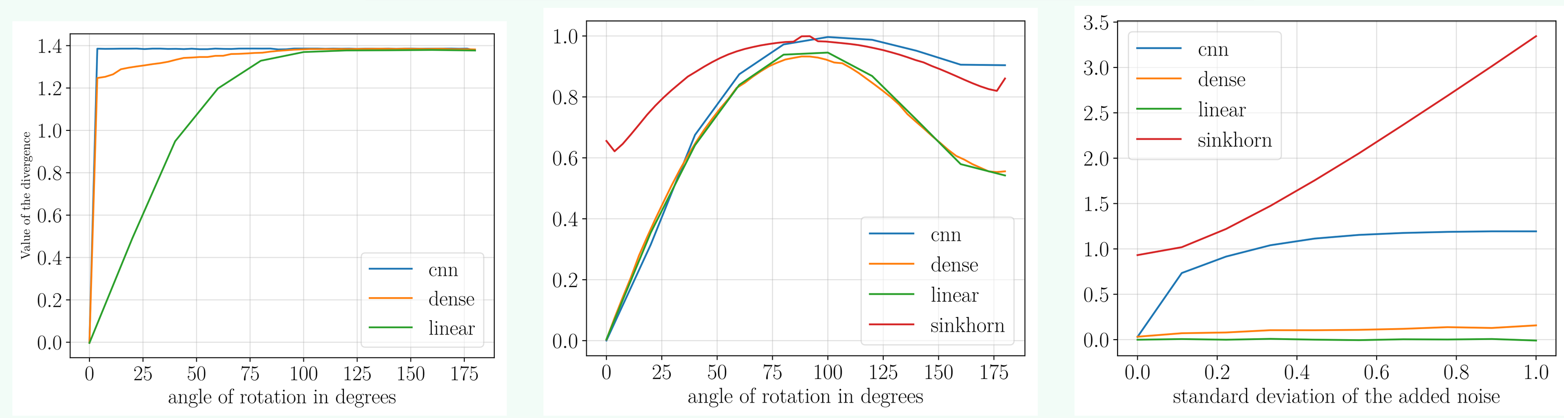
## Experiments

### 1. Importance of sample complexity

Images generated by the network after training with the Sinkorn-Autodiff algorithm on MNIST dataset (left) and CIFAR-10 dataset (right).

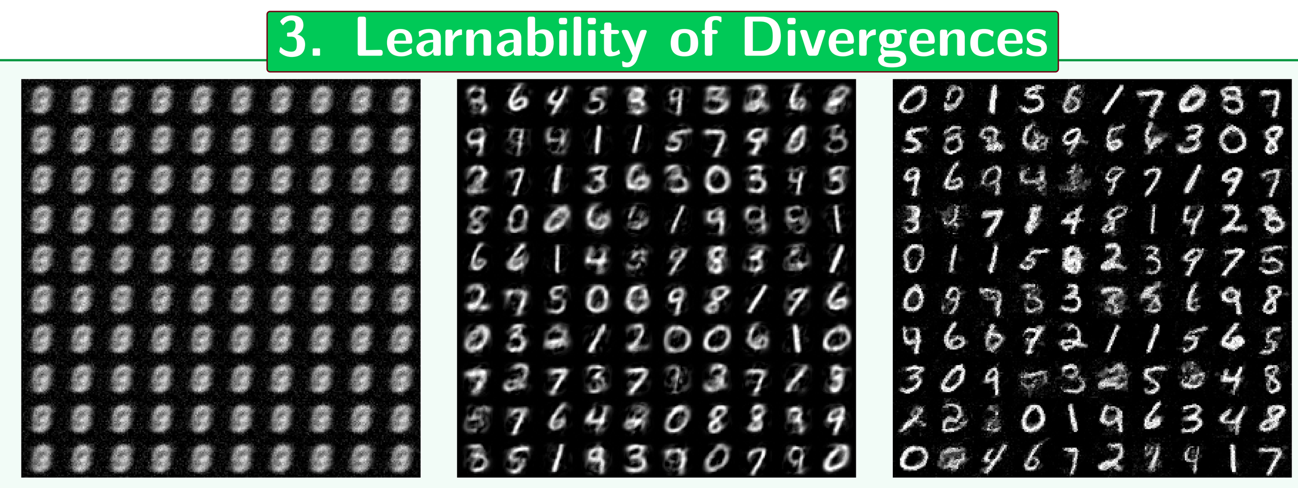


### 2. Robustness to Transformations




**JSD**(rotation) **WGAN**(rotation) **WGAN**(noise)

### 3. Learnability of Divergences



**Linear** **Dense** **CNN**

## References

[1] A. Osokin, F. Bach, and S. Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. *arXiv preprint arXiv:1703.02403*, 2017.