Dualing GANs

Yujia Li[1,*], Alexander Schwing[3], Kuan-Chieh Wang[1,2] and Richard Zemel[1,2]

[1]University of Toronto [2]Vector Institute [3]University of Illinois at Urbana-Champaign [*]Now at DeepMind

Introduction

GAN training suffers from instability due to its saddle point formulation:

$$\max_\theta \min_{\mathbf{w}} f(\theta, \mathbf{w})$$

θ and $\mathbf{w}$ are generator and discriminator parameters respectively, f is the GAN loss. Typical GAN training alternates gradient updates to θ and $\mathbf{w}$

$$\theta \to \theta + \eta_\theta \nabla_\theta f(\theta, \mathbf{w}), \quad \mathbf{w} \to \mathbf{w} - \eta_{\mathbf{w}} \nabla_{\mathbf{w}} f(\theta, \mathbf{w}).$$

However, to solve the saddle point problem ideally for each θ we want to solve for $\mathbf{w}^*(\theta) = \mathrm{argmin}_{\mathbf{w}} f(\theta, \mathbf{w})$, and then optimize $\max_\theta f(\theta, \mathbf{w}^*(\theta))$. For any $\mathbf{w}$ obtained from gradient updates, we have $f(\theta, \mathbf{w}) \geq f(\theta, \mathbf{w}^*(\theta))$, therefore the outer optimization becomes a maximization of an upper bound, leading to instability.

In this paper we propose to dualize the inner part $\min_{\mathbf{w}} f(\theta, \mathbf{w})$ into $\max_\lambda g(\theta, \lambda)$ which is always a lower bound on $f(\theta, \mathbf{w}^*(\theta))$ and solve the much more stable maximization problem

$$\max_\theta \max_\lambda g(\theta, \lambda).$$

This formulation allows us to:

- Solve the instability problem for GANs with linear discriminators.
- Improve stability for GANs with nonlinear discriminators.

GANs with Linear Discriminators

We start from linear discriminators that rely on a scoring function $F(\mathbf{w}, \mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. Any differentiable nonlinear feature $\phi(\mathbf{x})$ can be used in place of $\mathbf{x}$. The discriminator

$$D_{\mathbf{w}}(\mathbf{x}) = p_{\mathbf{w}}(y = 1|\mathbf{x}) = \sigma(F(\mathbf{w}, \mathbf{x})) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}}.$$

The GAN loss on a batch of data $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and latent samples $\{\mathbf{z}_1, ..., \mathbf{z}_n\}$ is

$$\min_{\mathbf{w}} \frac{C}{2}\|\mathbf{w}\|_2^2 + \frac{1}{2n}\sum_i \log\left(1 + e^{-\mathbf{w}^\top \mathbf{x}_i}\right) + \frac{1}{2n}\sum_i \log\left(1 + e^{\mathbf{w}^\top G_\theta(\mathbf{z}_i)}\right).$$

The loss is convex in $\mathbf{w}$, we can derive the standard dual problem to be

$$\begin{aligned} \max_\lambda \quad & g(\theta, \lambda) = -\frac{1}{2C}\left\|\sum_i \lambda_{\mathbf{x}_i}\mathbf{x}_i - \sum_i \lambda_{\mathbf{z}_i} G_\theta(\mathbf{z}_i)\right\|_2^2 \\ & + \frac{1}{2n}\sum_i H(2n\lambda_{\mathbf{x}_i}) + \frac{1}{2n}\sum_i H(2n\lambda_{\mathbf{z}_i}), \\ \text{s.t.} \quad & \forall i, \quad 0 \leq \lambda_{\mathbf{x}_i} \leq \frac{1}{2n}, \quad 0 \leq \lambda_{\mathbf{z}_i} \leq \frac{1}{2n}. \end{aligned}$$

$H(u) = -u\log u - (1-u)\log(1-u)$ is the binary entropy, and the optimal $\mathbf{w}^*$ can be obtained from the optimal solution $(\lambda_{\mathbf{z}}^*, \lambda_{\mathbf{x}}^*)$ for this dual problem as

$$\mathbf{w}^* = \frac{1}{C}\left(\sum_i \lambda_{\mathbf{x}_i}^* \mathbf{x}_i - \sum_i \lambda_{\mathbf{z}_i}^* G_\theta(\mathbf{z}_i)\right).$$

Properties:

- The $\|\sum_i \lambda_{\mathbf{x}_i}\mathbf{x}_i - \sum_i \lambda_{\mathbf{z}_i} G_\theta(\mathbf{z}_i)\|_2^2$ encourages moment matching.
- The entropy terms encourage the λ's to be close to the mean.

For training we optimize $\max_\theta \max_\lambda g(\theta, \lambda)$, which is very stable.

GANs with Non-Linear Discriminators

In general the scoring function $F(\mathbf{w}, \mathbf{x})$ may be nonlinear in $\mathbf{w}$ and typically implemented by a neural network. In this case the GAN loss

$$f(\theta, \mathbf{w}) = \frac{C}{2}\|\mathbf{w}\|_2^2 + \frac{1}{2n}\sum_i \log\left(1 + e^{-F(\mathbf{w}, \mathbf{x}_i)}\right) + \frac{1}{2n}\sum_i \log\left(1 + e^{F(\mathbf{w}, G_\theta(\mathbf{z}_i))}\right)$$

is not convex in $\mathbf{w}$, therefore hard to dualize directly.

Proposed solution: approximate f locally around any point $\mathbf{w}_k$ using a model function $m_{k,\theta}(\mathbf{s}) \approx f(\theta, \mathbf{w}_k + \mathbf{s})$, then dualize $m_{k,\theta}(\mathbf{s})$. The optimization problem for the discriminator becomes

$$\min_{\mathbf{s}} m_{k,\theta}(\mathbf{s}) \quad \text{s.t.} \quad \frac{1}{2}\|\mathbf{s}\|_2^2 \leq \Delta_k,$$

where $\frac{1}{2}\|\mathbf{s}\|^2 \leq \Delta_k$ is a trust-region constraint that ensures the quality of the approximation. The overall algorithm is shown below:

GAN optimization with model function

Initialize θ, $\mathbf{w}_0$, $k = 0$ and iterate

1. One or few gradient ascent steps on $f(\theta, \mathbf{w}_k)$ w.r.t. θ
2. Find step $\mathbf{s}$ using $\min_{\mathbf{s}} m_{k,\theta}(\mathbf{s})$ s.t. $\frac{1}{2}\|\mathbf{s}\|_2^2 \leq \Delta_k$
3. Update $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \mathbf{s}$
4. $k \leftarrow k + 1$

We explore two approximations:

(A). Cost function linearization: Linearize f as

$$m_{k,\theta}(\mathbf{s}) = f(\mathbf{w}_k, \theta) + \nabla_{\mathbf{w}} f(\mathbf{w}_k, \theta)^\top \mathbf{s}$$

We can solve for the optimal $\mathbf{s}^* = -\frac{\sqrt{2\Delta_k}}{\|\nabla_{\mathbf{w}} f(\mathbf{w}_k, \theta)\|_2}\nabla_{\mathbf{w}} f(\mathbf{w}_k, \theta)$ analytically. This $\mathbf{s}^*$ has the same form and direction as a gradient update used in standard GANs.

(B). Score function linearization: Linearize F and keep the loss

$$F(\mathbf{w}_k + \mathbf{s}, \mathbf{x}) \approx \hat{F}(\mathbf{s}, \mathbf{x}) = F(\mathbf{w}_k, \mathbf{x}) + \mathbf{s}^\top \nabla_{\mathbf{w}} F(\mathbf{w}_k, \mathbf{x}), \quad \forall \mathbf{x}.$$

Model function is a more accurate approximation compared to (A).

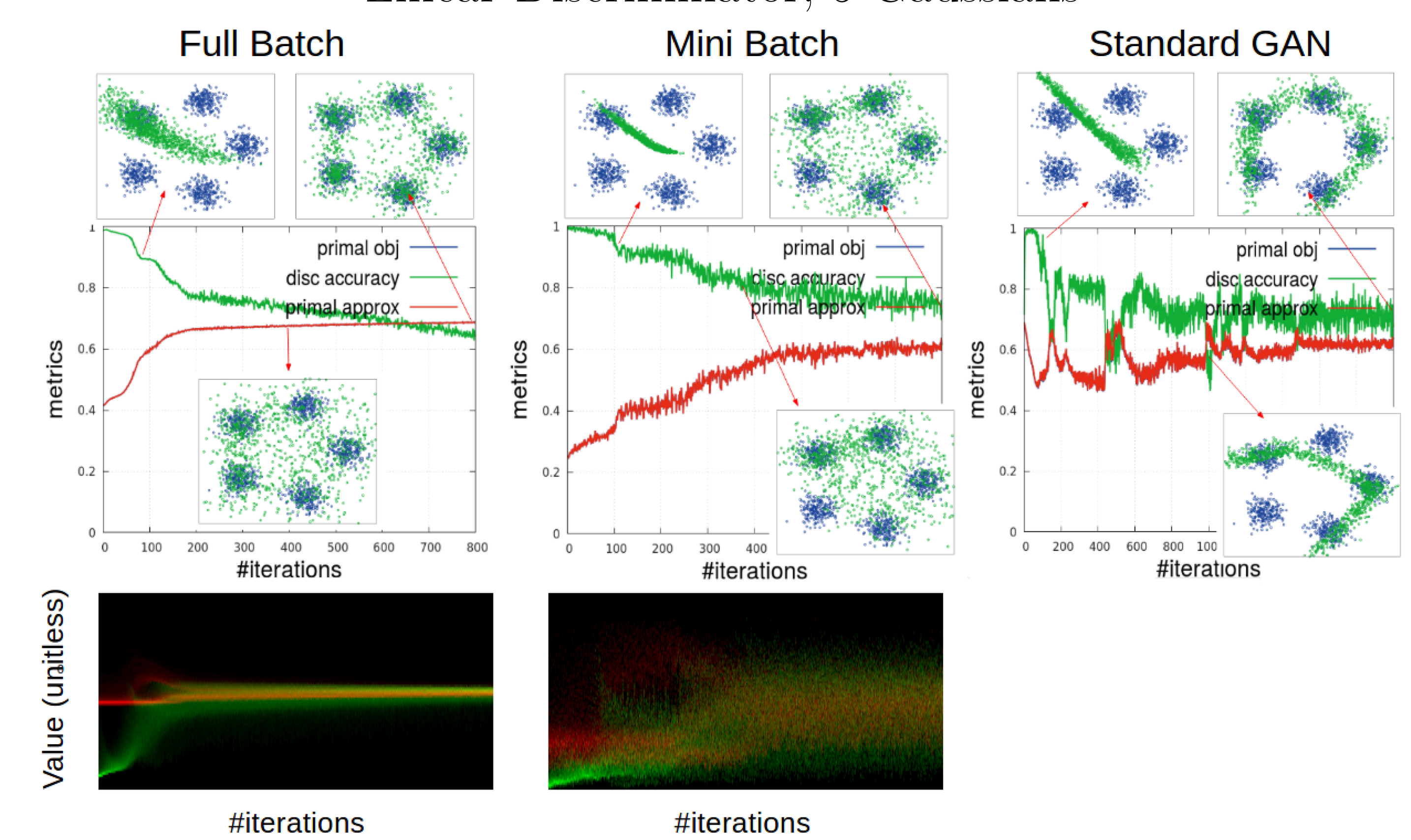
$$\begin{aligned} m_{k,\theta}(\mathbf{s}) = & \frac{C}{2}\|\mathbf{w}_k + \mathbf{s}\|_2^2 + \frac{1}{2n}\sum_i \log\left(1 + e^{-F(\mathbf{w}_k, \mathbf{x}_i) - \mathbf{s}^\top \nabla_{\mathbf{w}} F(\mathbf{w}_k, \mathbf{x}_i)}\right) \\ & + \frac{1}{2n}\sum_i \log\left(1 + e^{F(\mathbf{w}_k, G_\theta(\mathbf{z}_i)) + \mathbf{s}^\top \nabla_{\mathbf{w}} F(\mathbf{w}_k, G_\theta(\mathbf{z}_i))}\right). \end{aligned}$$

This m is convex in $\mathbf{s}$ and can be dualized. See paper for details.

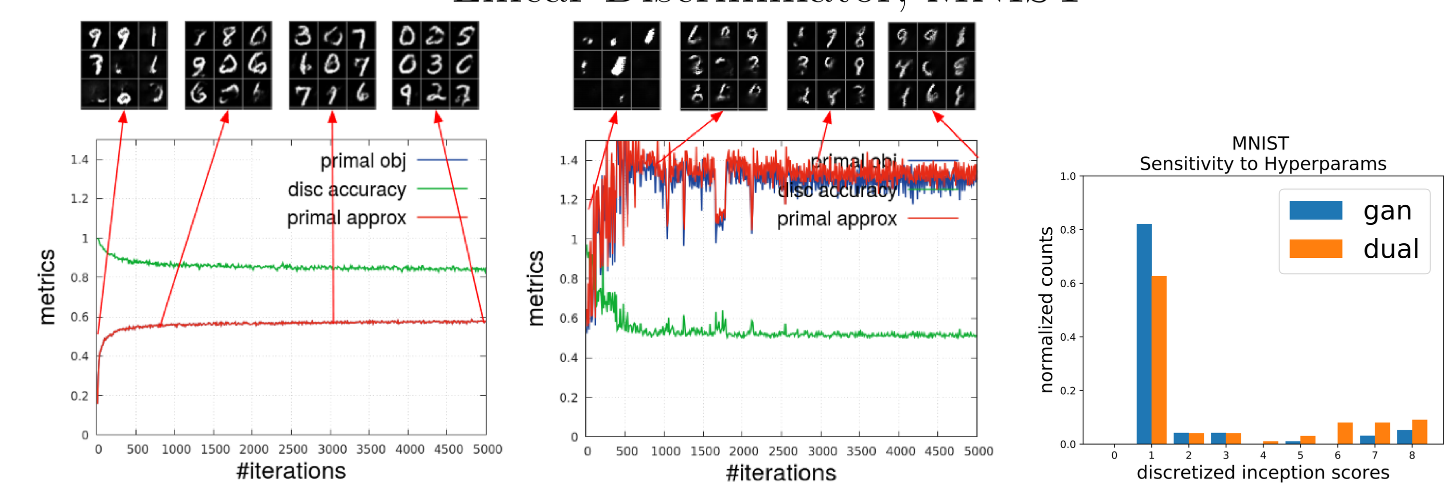
With approximations the dual is not exact, but solving the dual may still be better than taking gradient steps for the primal.

Experiments

Linear Discriminator, 5 Gaussians

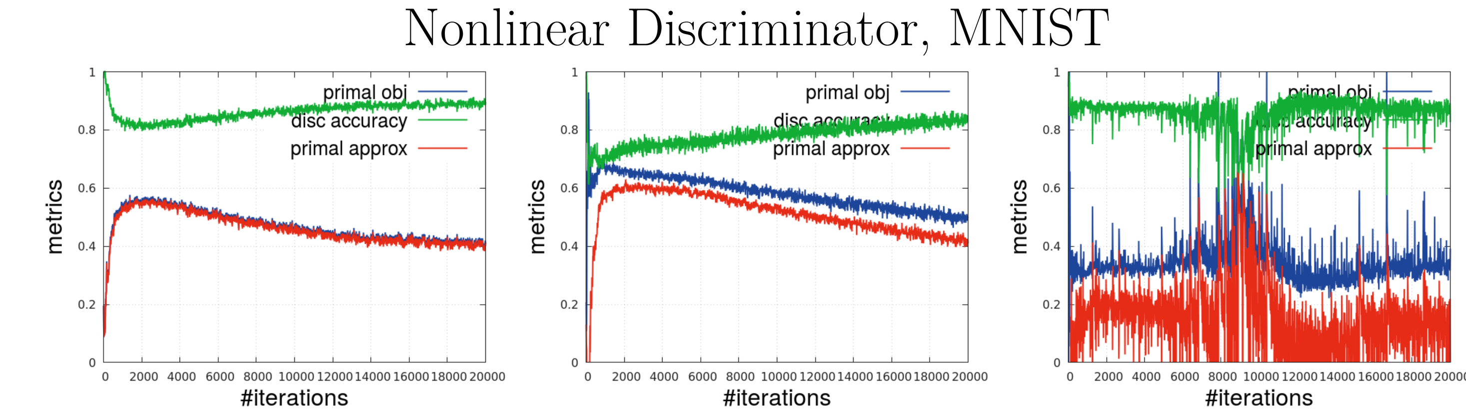


Linear Discriminator, MNIST



Dual GAN | Standard GAN | Sensitivity to Hyperparams

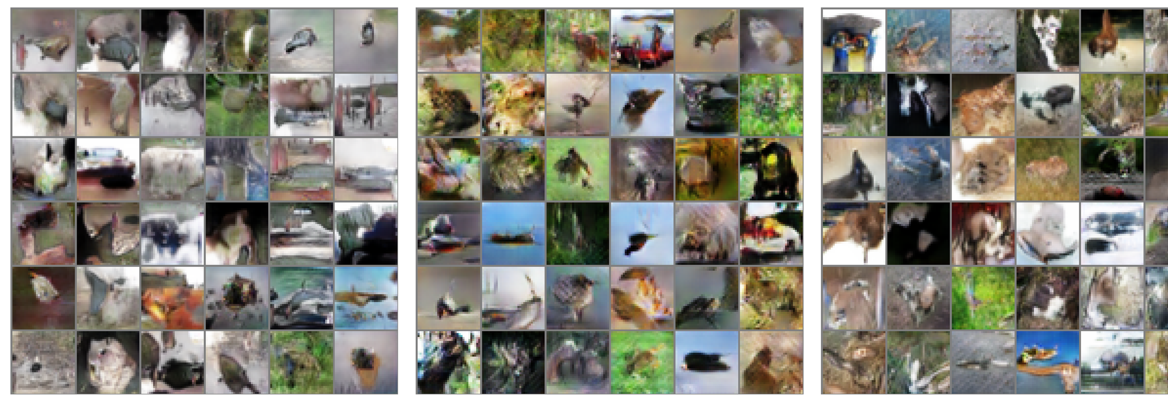
Nonlinear Discriminator, MNIST



Score Lin. | Cost Lin. | Standard GAN

Nonlinear Discriminator, CIFAR-10

Score Type	Std. GAN	Score Lin	Cost Lin	Real Data
Inception (end)	5.61±0.09	5.40±0.12	5.43±0.10	10.72 ± 0.38
Our classifier (end)	3.85±0.08	3.52±0.09	4.42±0.09	8.03 ± 0.07
Inception (avg)	5.59±0.38	5.44±0.08	5.16±0.37	-
Our classifier (avg)	3.64±0.47	3.70±0.27	4.04±0.37	-



Score Lin. | Cost Lin. | Standard GAN