Bridging the Gap Between f-GANs and Wasserstein GANs

Jiaming Song, Stefano Ermon

Stanford University

http://tsong.me

Generative Adversarial Networks

$$T(\boldsymbol{x})$$

$$x_{\text{real}} \sim P \longleftrightarrow x_{\text{fake}} \sim Q$$

$$\max_{T \in \mathcal{F}} \mathbb{E}_P[T] - \mathbb{E}_Q[\cdot]$$

f-GAN

- $\mathcal{F} =$ "all" functions
- Use $\mathbb{E}_Q[f^\star(T(\boldsymbol{x}))]$
- Estimate f-divergences

WGAN

- $\mathcal{F} =$ 1-Lipschitz func.
- Use $\mathbb{E}_Q[T(\boldsymbol{x})]$
- Estimate Wasserstein distances

We generalize and extend f-GAN and WGAN objectives (and respective notions of distances)

A Generalized Objective

We are interested in "distance" between P and Q

P: real distribution $\quad Q$: generated distribution

T: discriminator $\quad r$: "weights"

$$-\mathbb{E}_{\boldsymbol{x}\sim Q}[r(\boldsymbol{x})\cdot T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P}[T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim Q}[f(r(\boldsymbol{x}))]$$

"weighted generator" → $\min_{Q}$ distribution matching

"real samples" → $\max_{T}$ estimate divergence

"convex regularization over weights" → $\min_{r\in\mathcal{R}}$ re-weight generator

A Generalized Objective

P: real distribution $\quad Q$: generated distribution

T: discriminator $\quad r$: "weights"

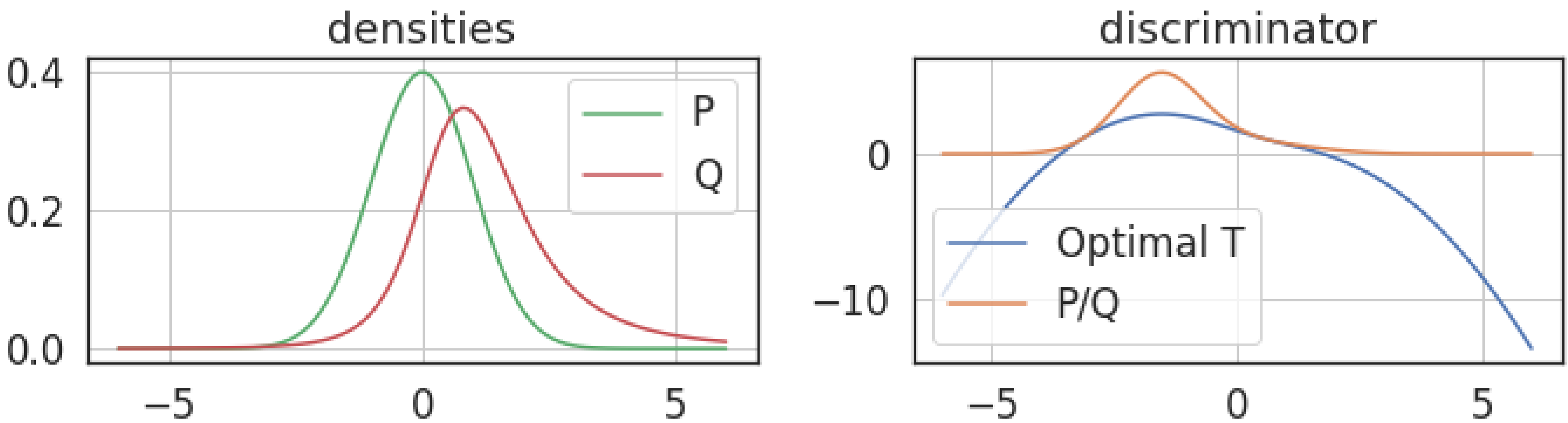
$$\min_{Q} \max_{T} \min_{r \in \mathcal{R}} -\mathbb{E}_{\boldsymbol{x}\sim Q}[r(\boldsymbol{x}) \cdot T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P}[T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim Q}[f(r(\boldsymbol{x}))]$$

"weighted generator" — "real samples" — "convex regularization over weights"

The above objective can generalize f-GAN and WGANs!

Discriminator and sample quality

- Higher T(x) -> x is more likely to come from P (“higher quality”)
- f-GAN estimates density ratio P/Q, WGAN does not!



Intuition: re-weighting based on “sample quality”

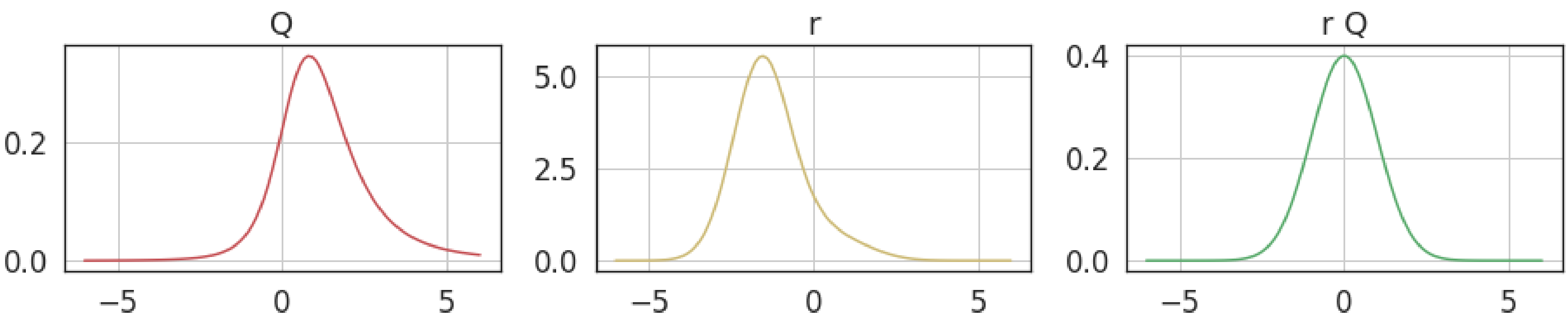
Re-weighting the Generated Samples

Weights $r(\boldsymbol{x})$: produce non-negative weight for a sample $\boldsymbol{x}$.

$$\mathbb{E}_Q[r(\boldsymbol{x}) \cdot T(\boldsymbol{x})] \quad \to \quad \mathbb{E}_{rQ}[T(\boldsymbol{x})]$$

"weights"

"importance weights"
if r is density ratio



"distribution" "weights" "re-weighted"

A Generalized Objective

$$\min_{Q} \max_{T} \min_{r \in \mathcal{R}} -\mathbb{E}_{\boldsymbol{x} \sim Q}[r(\boldsymbol{x}) \cdot T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim P}[T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim Q}[f(r(\boldsymbol{x}))]$$

"weighted generator" "real samples" "convex regularization over weights"

"valid set of possible weights that we consider"

What set should we choose?

$\mathcal{R} = \{\mathbb{1}\}$

function that only outputs 1

$\mathcal{R} = \Delta(Q)$

valid density ratios over Q

$\mathcal{R} = L^{\infty}(Q)$

"almost all" functions

Generalization of *f*-GAN and WGANs

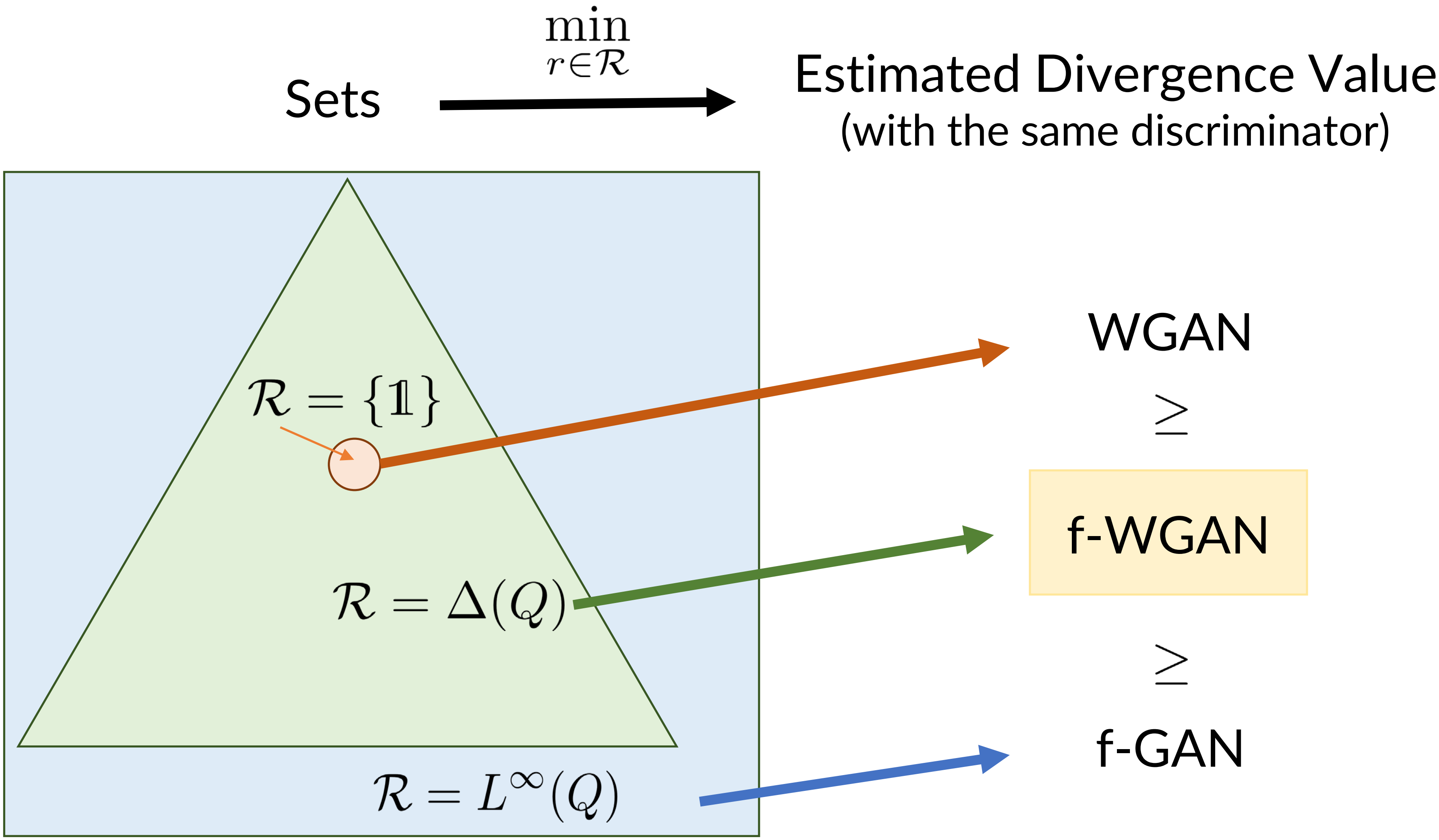
$$\min_{Q} \max_{T} \min_{r \in \mathcal{R}} -\mathbb{E}_{\boldsymbol{x} \sim Q}[r(\boldsymbol{x}) \cdot T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim P}[T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim Q}[f(r(\boldsymbol{x}))]$$

- Different choices of $\mathcal{R}$ give different objectives!

$\mathcal{R} = \{\mathbb{1}\}$		$\mathcal{R} = \Delta(Q)$		$\mathcal{R} = L^{\infty}(Q)$
function that only outputs 1	$\subseteq$	valid density ratios over Q	$\subseteq$	"almost all" functions
$\min_{r \in \mathcal{R}} \downarrow$		$\min_{r \in \mathcal{R}} \downarrow$		$\min_{r \in \mathcal{R}} \downarrow$
WGAN	$\geq$	f-WGAN	$\geq$	f-GAN

(when estimated with the same discriminator)

Beyond f-GAN and WGANs



Training the f-WGAN objective

- How do we find the optimal weights?

$$\min_{Q} \max_{T} \min_{r \in \Delta(Q)} -\mathbb{E}_{\boldsymbol{x}\sim Q}[r(\boldsymbol{x}) \cdot T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P}[T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim Q}[f(r(\boldsymbol{x}))]$$

Need to optimize this now

- For certain *f* we can find optimal weights analytically!
- e.g. KL divergence, chi^2 divergence
- Both place higher weights to “better” samples!

Why do we assign higher weights to "better" samples?

Generator has access to discriminator values

↓

Use them to improve generator distribution

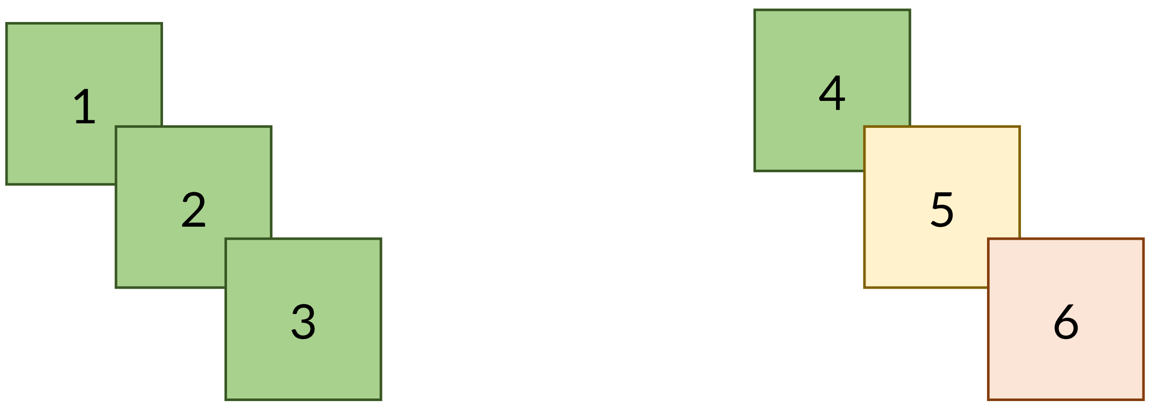
↓

Discriminator anticipates that generator is using its information

Both are learning with explicit awareness of the opponent!

Algorithm

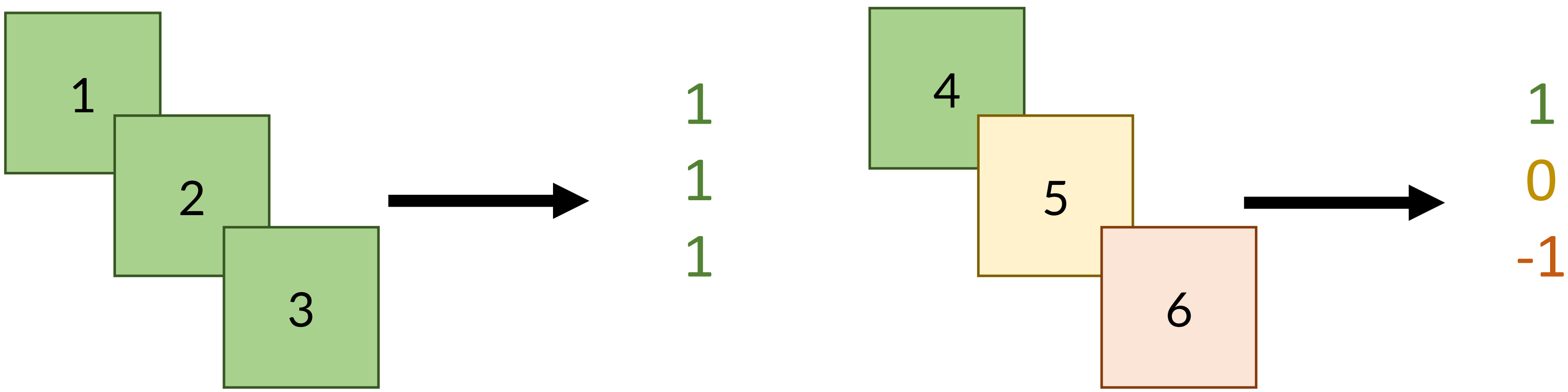
- Sample a batch of real / fake samples



$x_{\text{real}} \sim P$ $\quad$ $x_{\text{fake}} \sim Q$

Algorithm

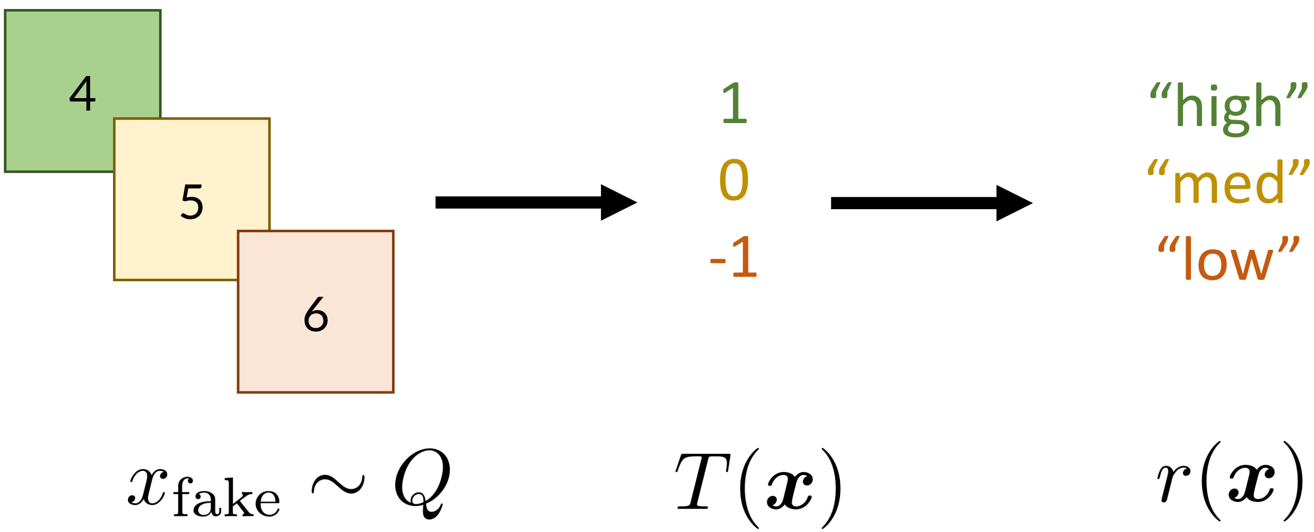
- Obtain discriminator values for the samples



$x_{\text{real}} \sim P$ $T(\boldsymbol{x})$ $x_{\text{fake}} \sim Q$ $T(\boldsymbol{x})$

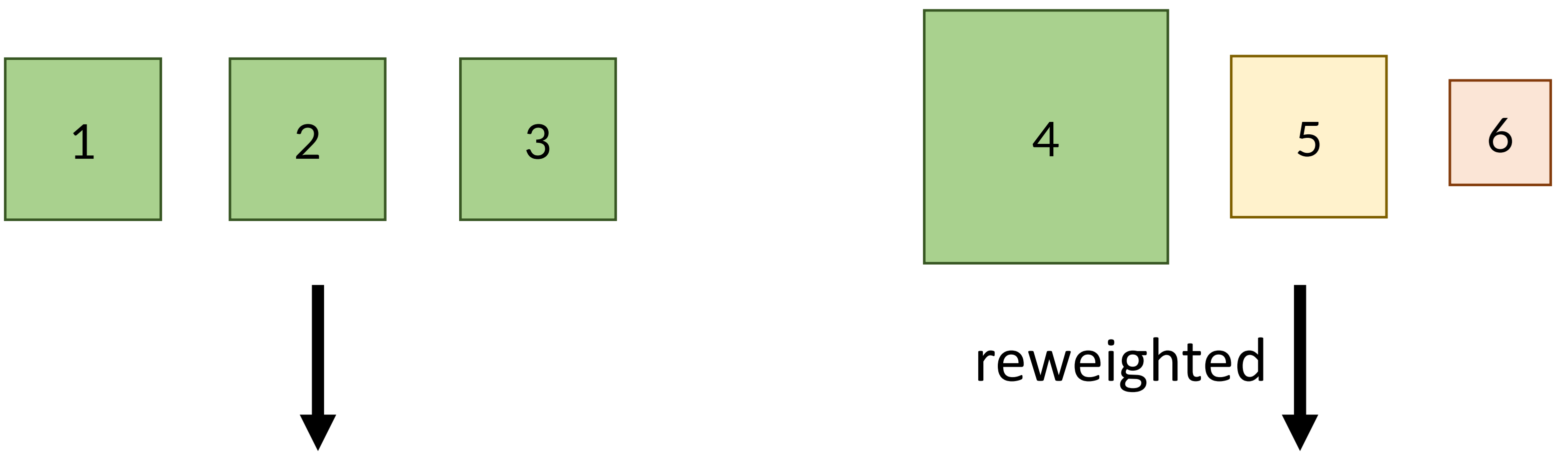
Algorithm

- Compute the weights within the batch



Algorithm

- Assign weights to objective and optimize new objective



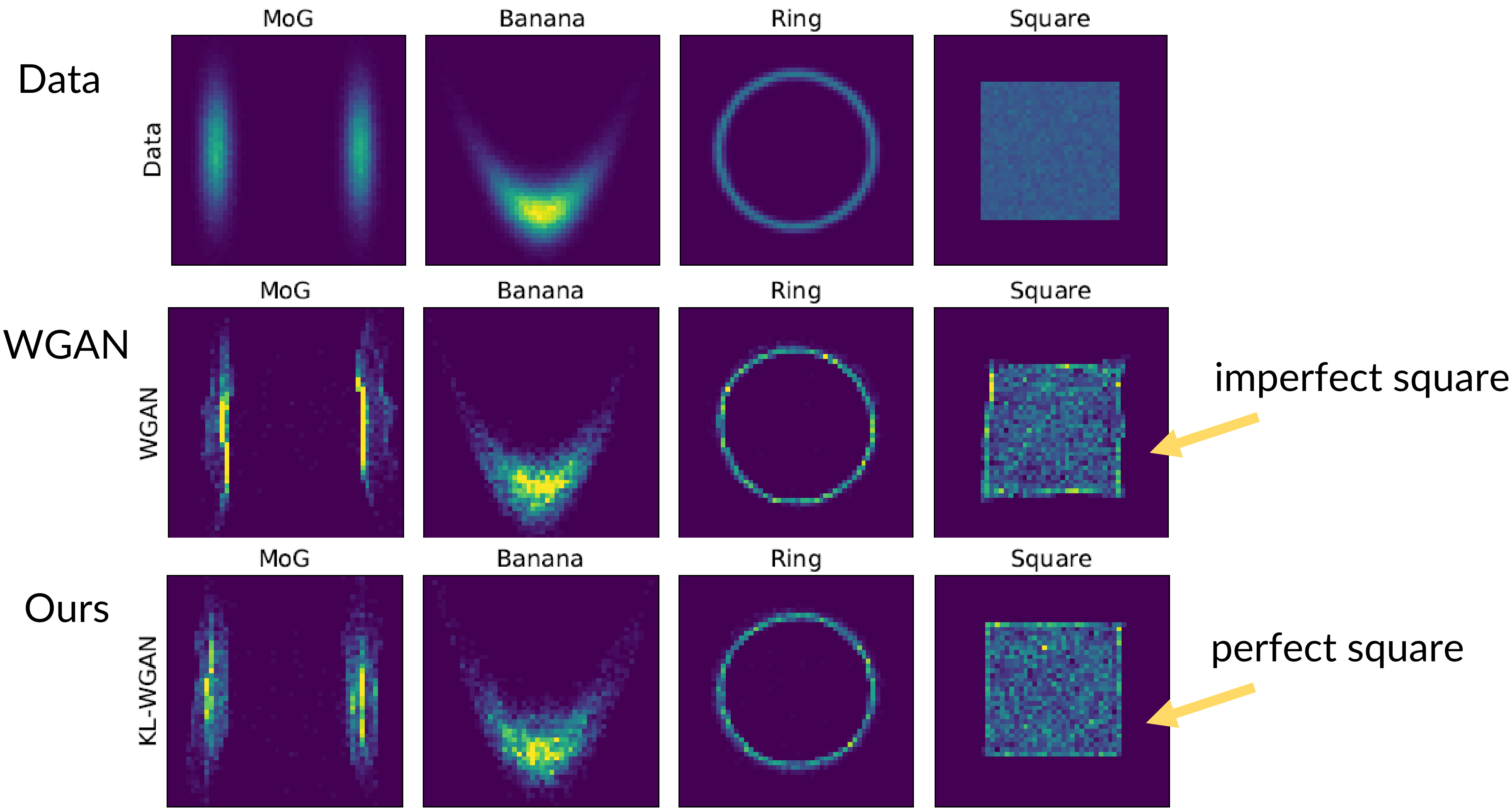
$$E_P[T(\boldsymbol{x})] \qquad E_Q[r(\boldsymbol{x})T(\boldsymbol{x})]$$

- Optimize the new objective

Experiments

- Synthetic data generation (vs. WGAN)
- Density ratio estimation (vs. f-GAN)
- Image generation (vs. WGAN)

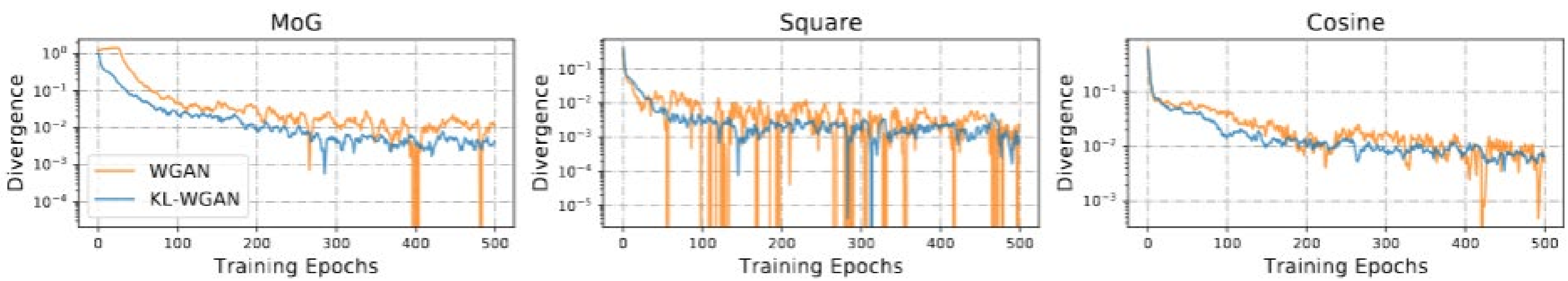
Experiments (synthetic data generation)



KL-WGAN learns better shapes than WGAN!

Experiments (synthetic data generation)

Metric	GAN	MoG	Banana	Rings	Square
NLL	W	2.65 ± 0.00	3.61 ± 0.02	$\mathbf{4.25} \pm 0.01$	3.73 ± 0.01
	KL-W	$\mathbf{2.54} \pm 0.00$	$\mathbf{3.57} \pm 0.00$	$\mathbf{4.25} \pm 0.00$	$\mathbf{3.72} \pm 0.00$
MMD	W	25.45 ± 7.78	3.33 ± 0.59	2.05 ± 0.47	2.42 ± 0.24
	KL-W	$\mathbf{6.51} \pm 3.16$	$\mathbf{1.45} \pm 0.12$	$\mathbf{1.20} \pm 0.10$	$\mathbf{1.10} \pm 0.23$

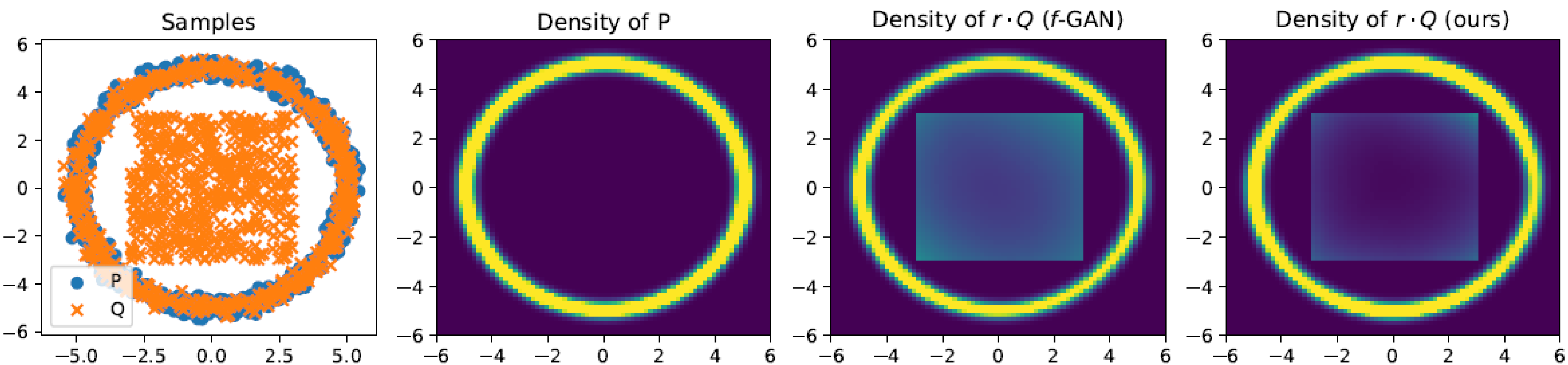


- Estimated divergences are more stable

Experiments (density ratio estimation)

1. Given samples from P and Q
2. Train time: estimate ratio of P / Q
3. Test time: observe ratio times ground truth density of Q
4. Closer to ground truth density of P -> better!

Experiments (density ratio estimation)



f-GAN Ours

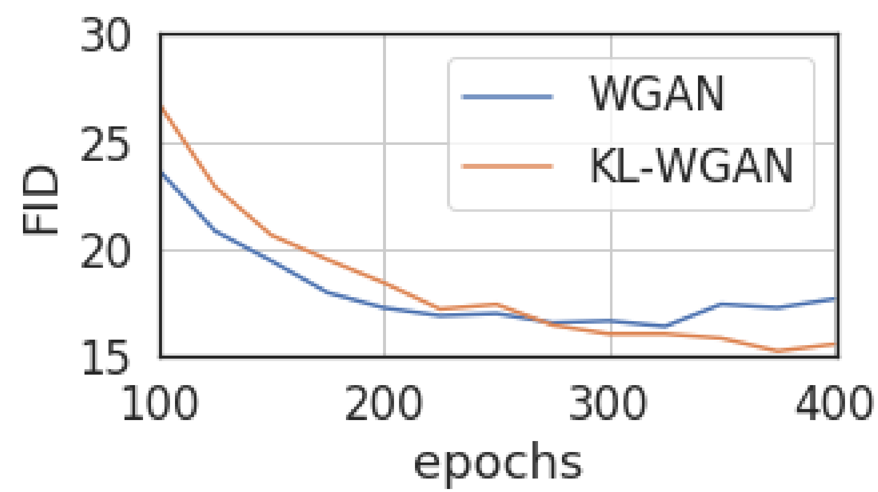
Samples Ground truth Estimates

- Better density ratio estimates than KL-GAN!

Experiments (image generation)

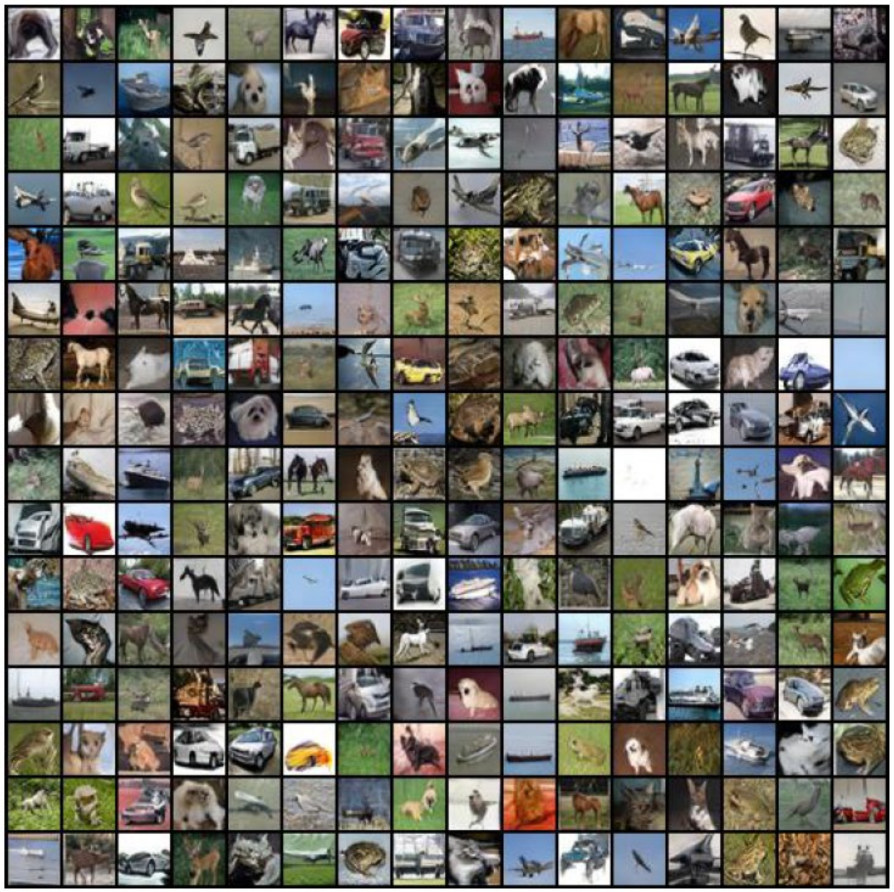
Method	Inception score	FID score
CIFAR10 Unconditional		
WGAN-GP	$7.86 \pm .07$	-
Fisher GAN	$7.90 \pm .05$	-
MoLM	$7.90 \pm .10$	18.9
SNGAN	$8.22 \pm .05$	21.7
Sphere GAN	$8.39 \pm .08$	**17.1**
NCSN	**8.91**	25.32
BigGAN*	$8.60 \pm .10$	16.38
KL-BigGAN*	$\mathbf{8.66} \pm .09$	**15.23**
CIFAR10 Conditional		
Fisher GAN	$8.16 \pm .12$	-
WGAN-GP	$8.42 \pm .10$	-
χ^2-GAN	$8.44 \pm .10$	-
SNGAN	$8.60 \pm .08$	17.5
BigGAN	**9.22**	**14.73**
BigGAN*	$9.08 \pm .11$	9.51
KL-BigGAN*	$\mathbf{9.20} \pm .09$	**9.17**

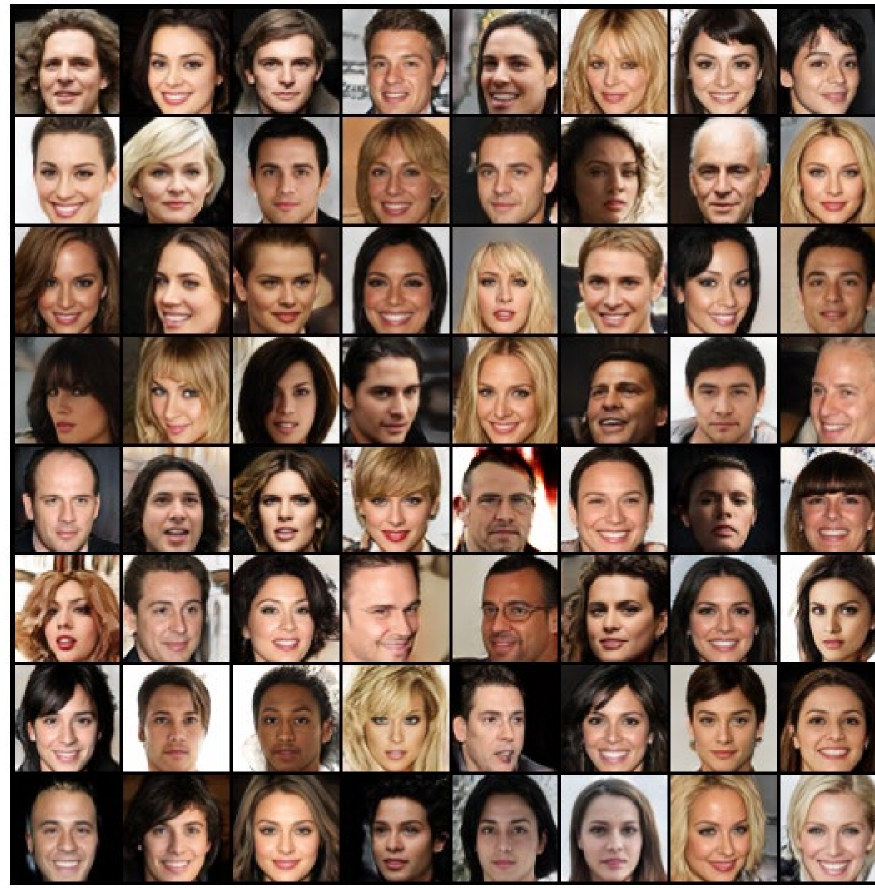
Method	Image Size	FID score
BigGAN	64×64	18.07 ± 0.47
KL-BigGAN	64×64	$\mathbf{17.70} \pm 0.32$



- Better image generation quality than WGAN counterparts!
- In comparison, f-GAN (KL) has stability issues.

Experiments (image generation)





Takeaway

- Generalization of f-GAN and WGANs
- Introduce new family of objectives
- Better density ratio estimates than f-GAN
- Better sample quality than WGAN

Paper: https://arxiv.org/abs/1910.09779

Code: https://github.com/ermongroup/f-wgan

Contact: tsong@cs.stanford.edu