Generative Adversarial Networks

Part 1

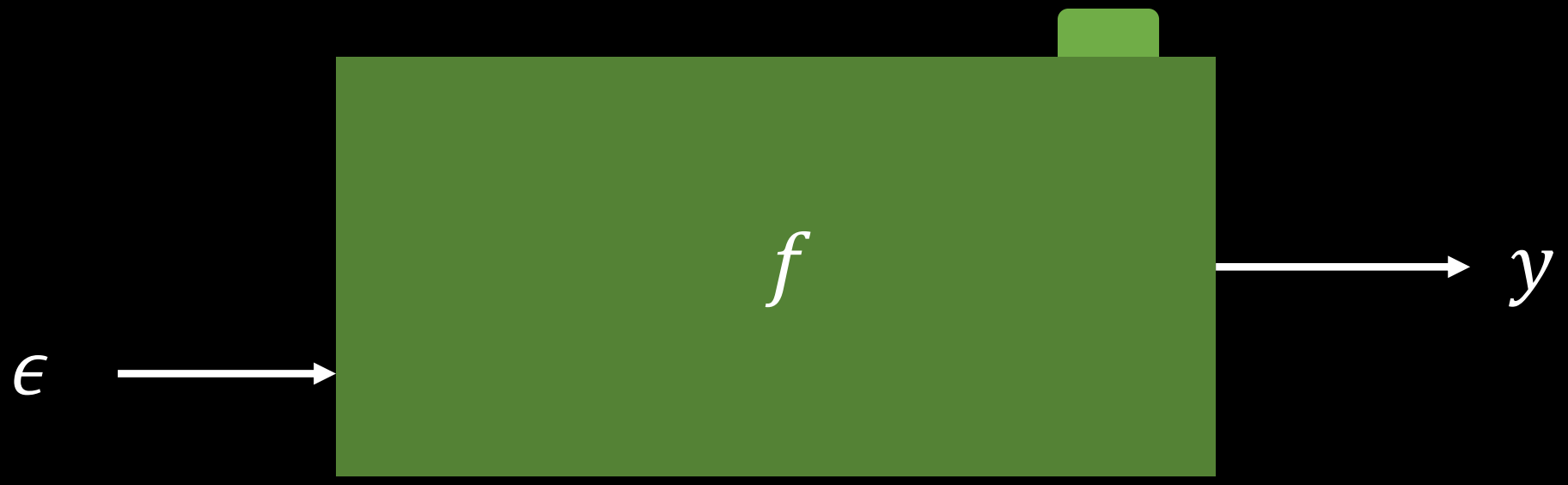
Sebastian Nowozin
Microsoft Research Cambridge
MLSS 2018 – Madrid – 7th September 2018

Outline

- Probabilistic models
- Example applications of GANs
- Principles of estimation
- GAN models
  - Divergences and f-GAN family
  - IPM family: MMD
  - IPM family: Wasserstein GANs
- Problems and Fixes: Mode Collapse, Instability
- Implicit models more generally
- Open research problems

Probabilistic Models

x
f
y


f
y
ϵ

x
ϵ
f
y

Taxonomy

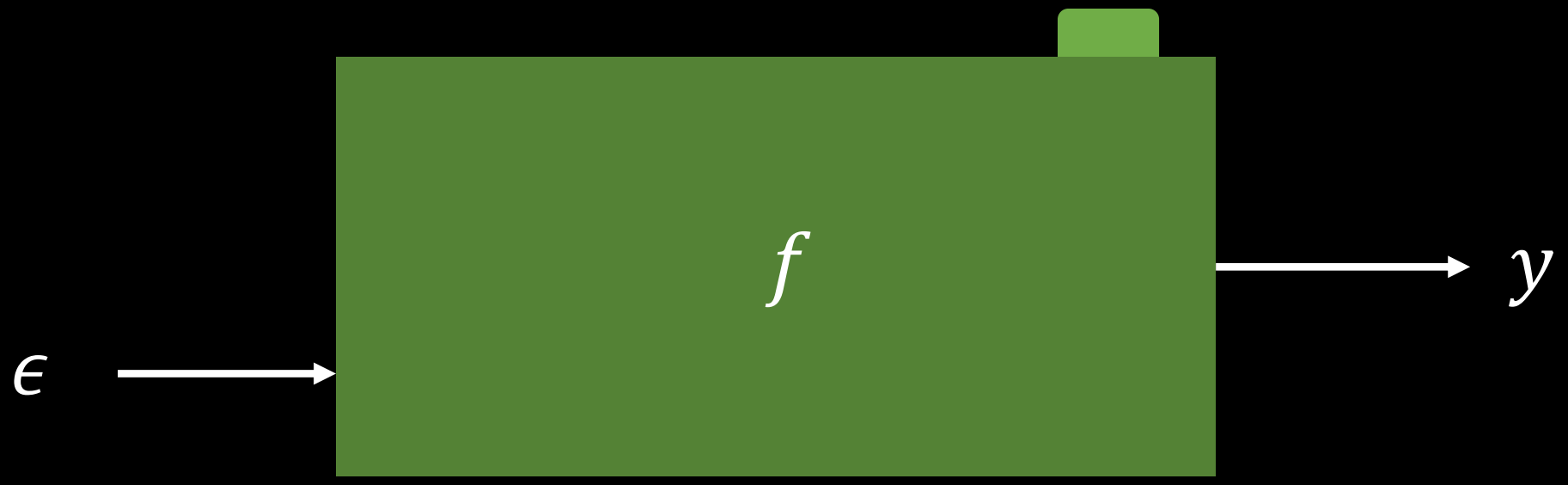
Non-probabilistic

Probabilistic

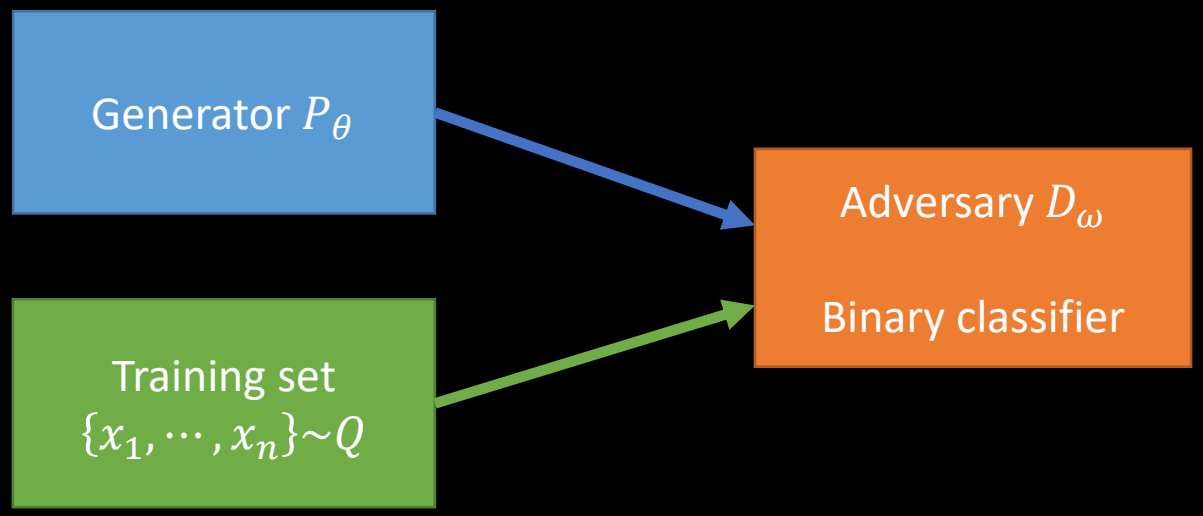
Generative

Discriminative

"Conditionally Generative"


f
y
ϵ

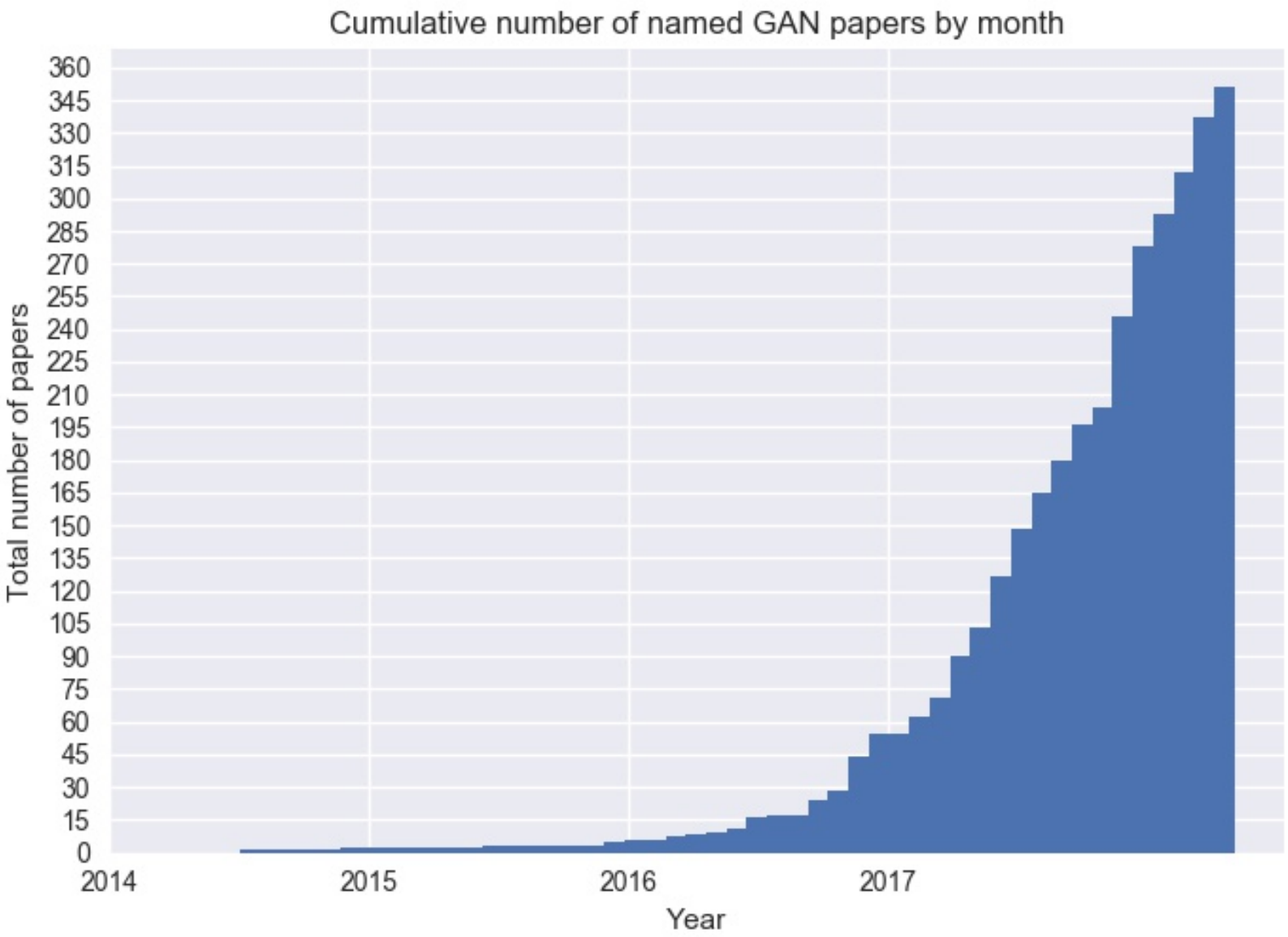
GAN Training Objective [Goodfellow et al., 2014]



- Generator tries to fool discriminator (i.e. generate realistic samples)
- Discriminator tries to distinguish fake from real samples
- Saddle-point problem

$$\min_\theta \max_\omega \mathbb{E}_{x \sim P_\theta}[\log D_\omega(x)] + \mathbb{E}_{x \sim Q}[\log(1 - D_\omega(x))]$$

GAN Research



Cumulative number of GAN models (Credit: Bruno Gavranović)

Applications of GANs

Generating Face Images, 2014-2018

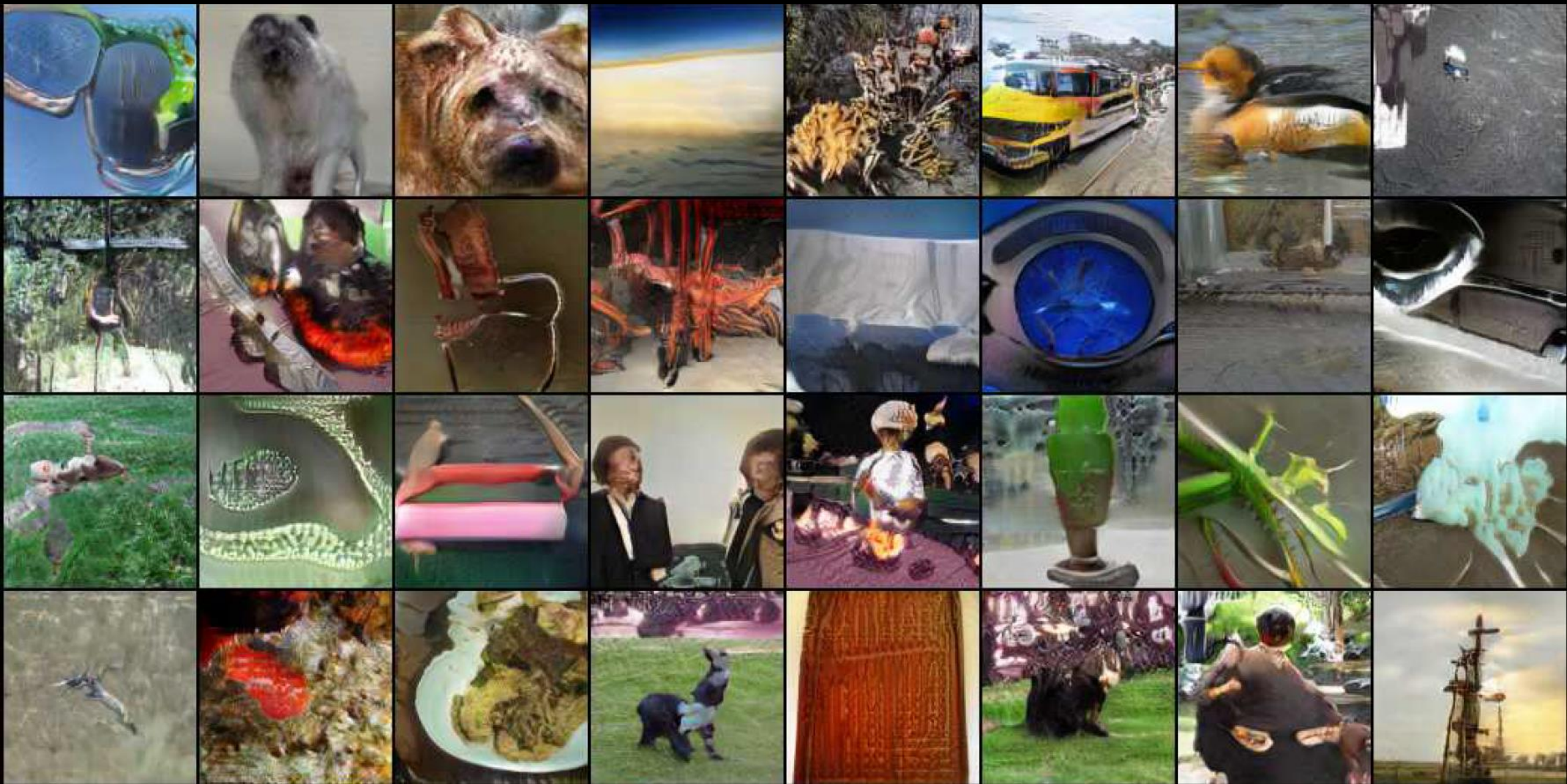
[Goodfellow et al., 2014]
University of Montreal

[Radford et al., 2015]
Facebook AI Research

[Roth et al., 2017]
Microsoft and ETHZ

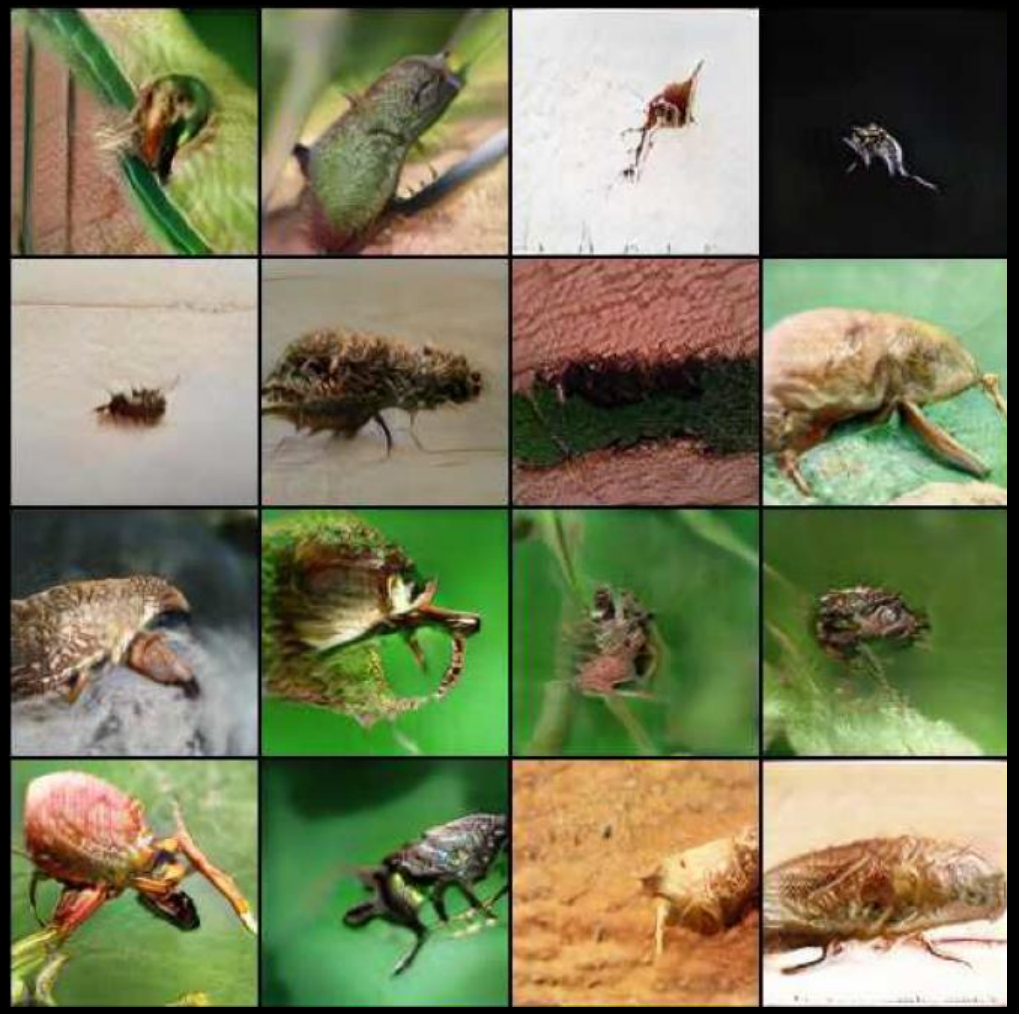
[Karras et al., 2018]
NVIDIA

ImageNet 1000 classes, 128x128, unconditional generation, ResNet 55 [Mescheder et al., ICML 2018], arXiv:1801.04406



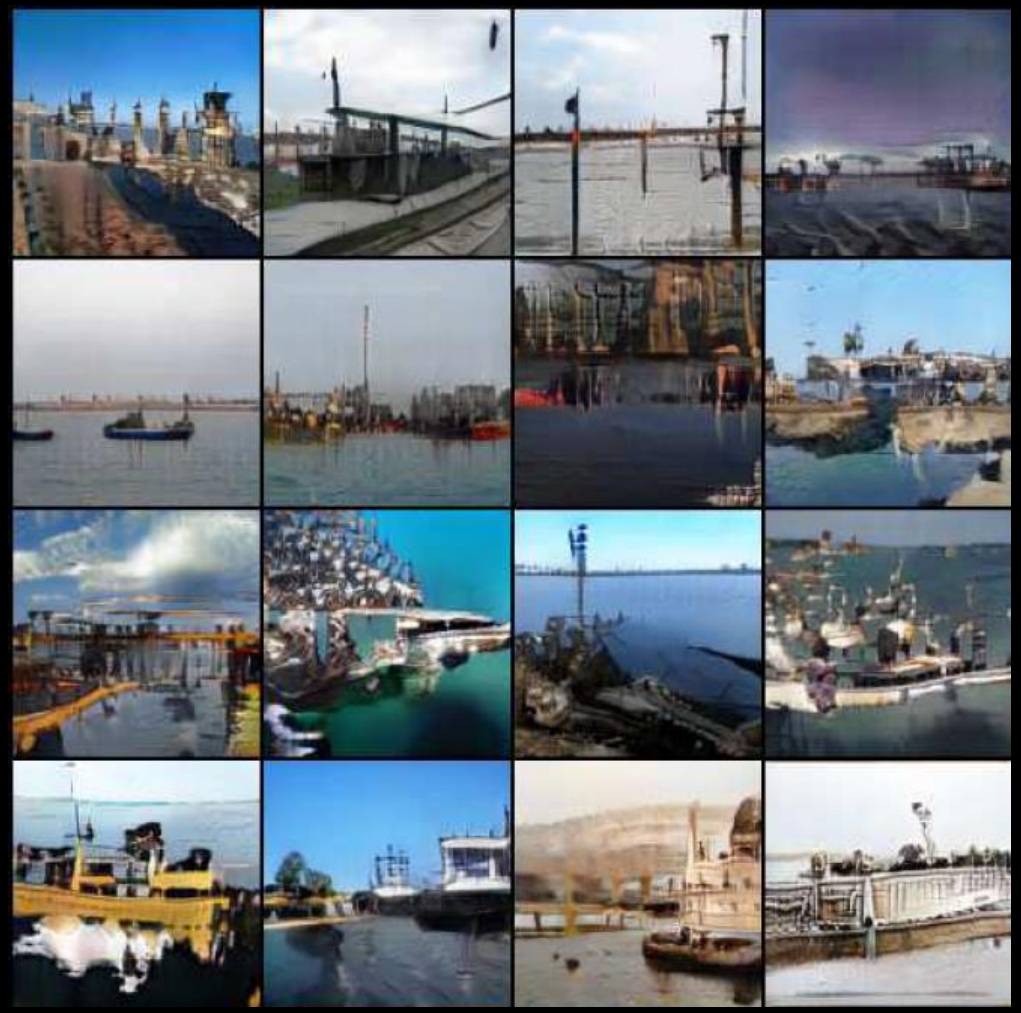
ImageNet conditional generation [Mescheder et al., ICML 2018], arXiv:1801.04406

papillon



weevil

ImageNet conditional generation [Mescheder et al., ICML 2018], arXiv:1801.04406

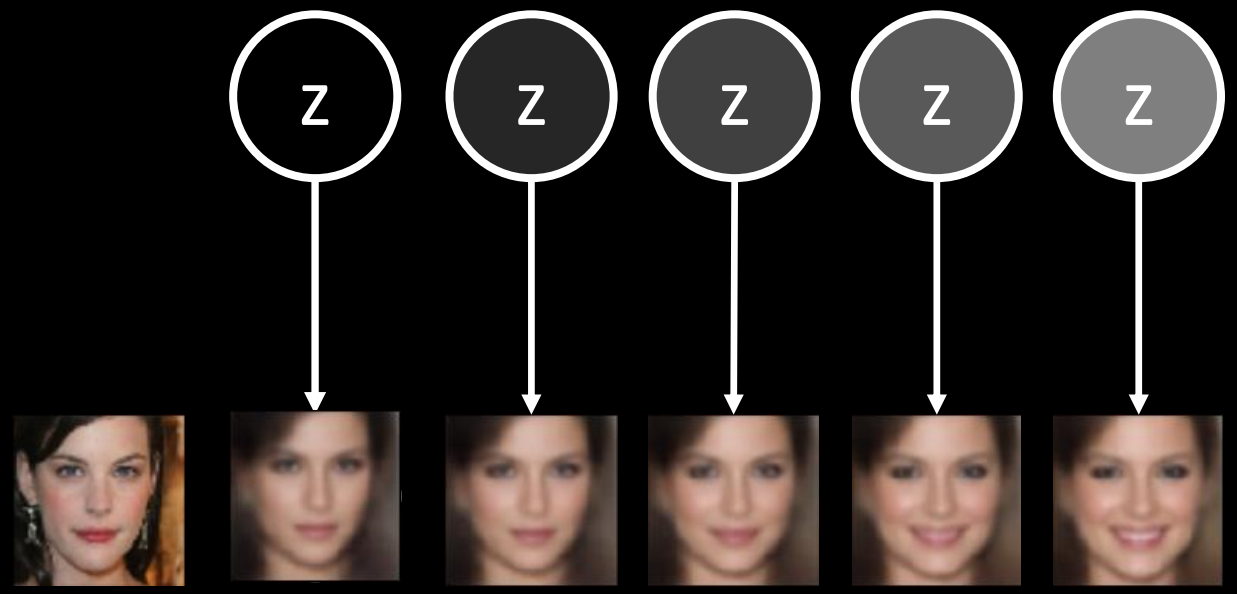


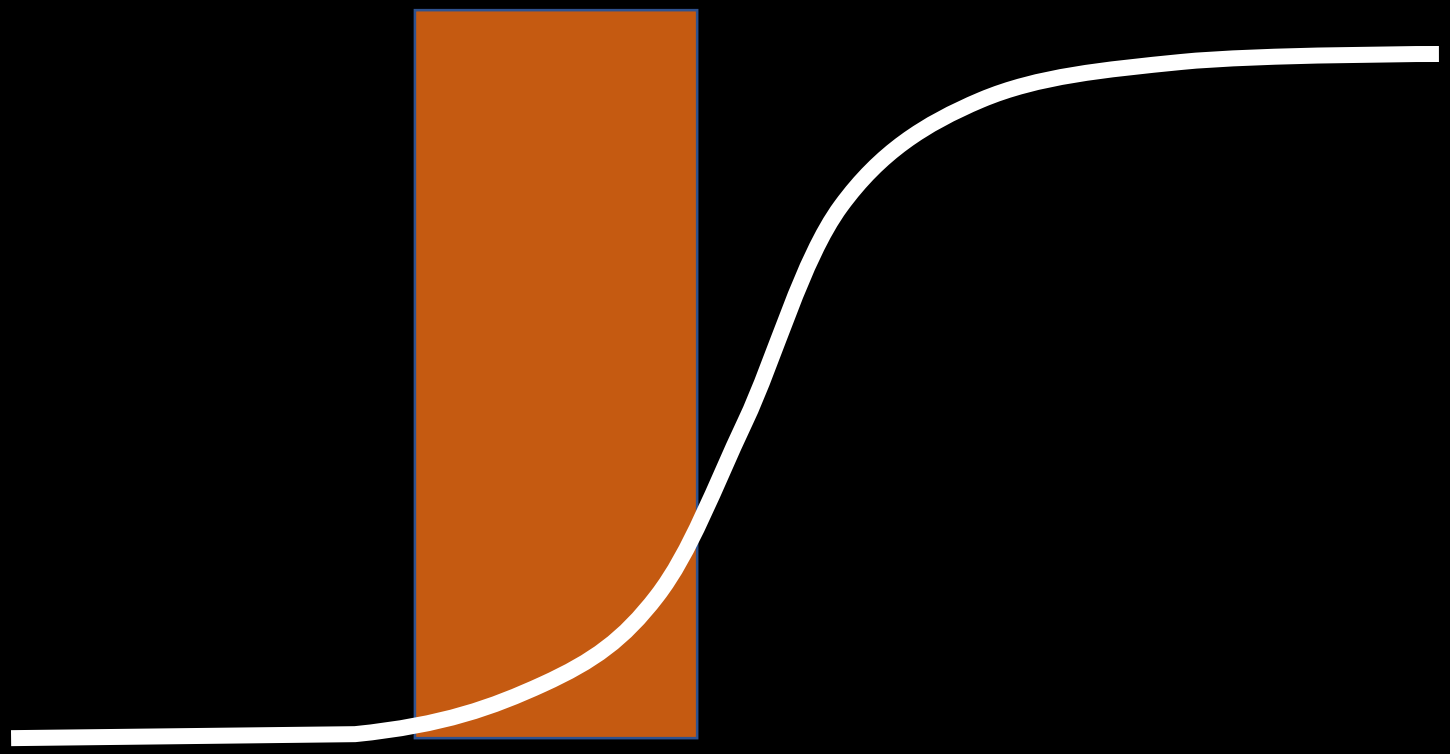
dock

pizza

Unsupervised learning ⟷ Representation learning

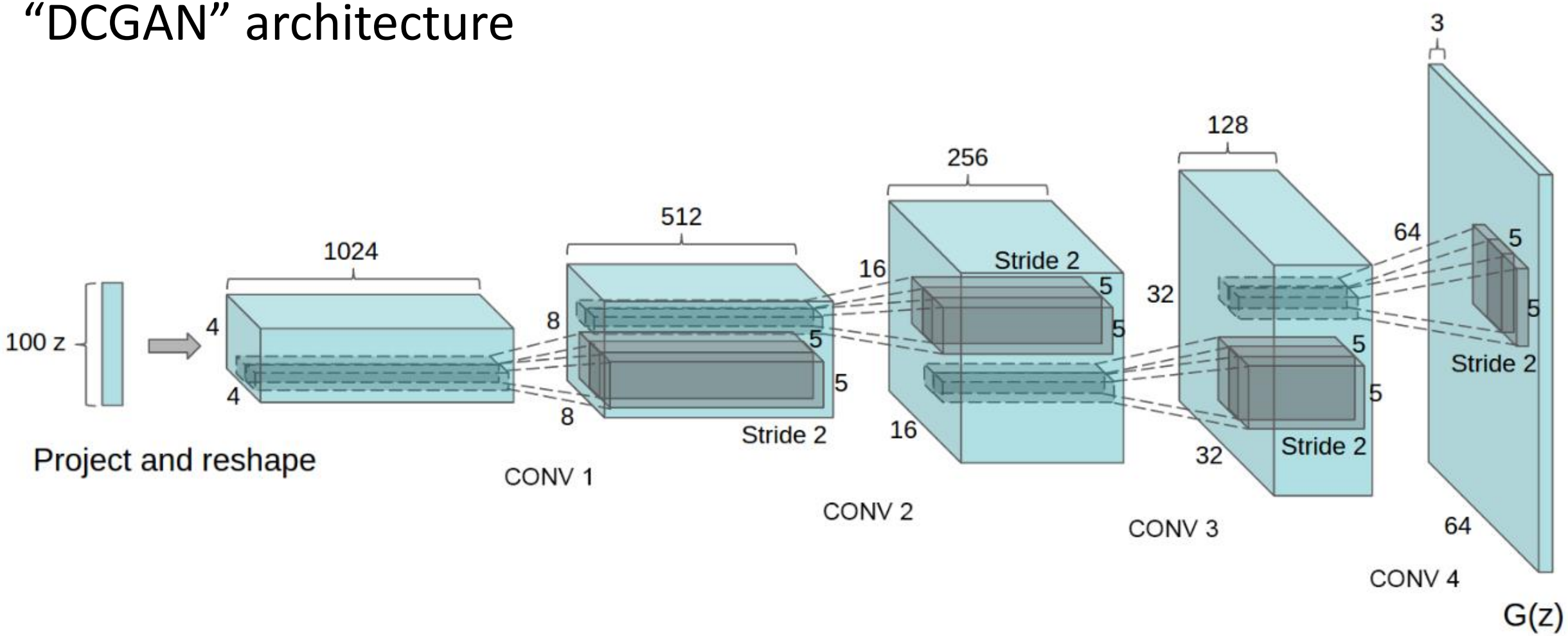
Representation Learning





Natural Images (Radford et al., 2015, arXiv:1511.06434)

“DCGAN” architecture



Linear interpolation in latent space [Radford et al., 2015]

Vector Arithmetic [Radford et al., 2015]

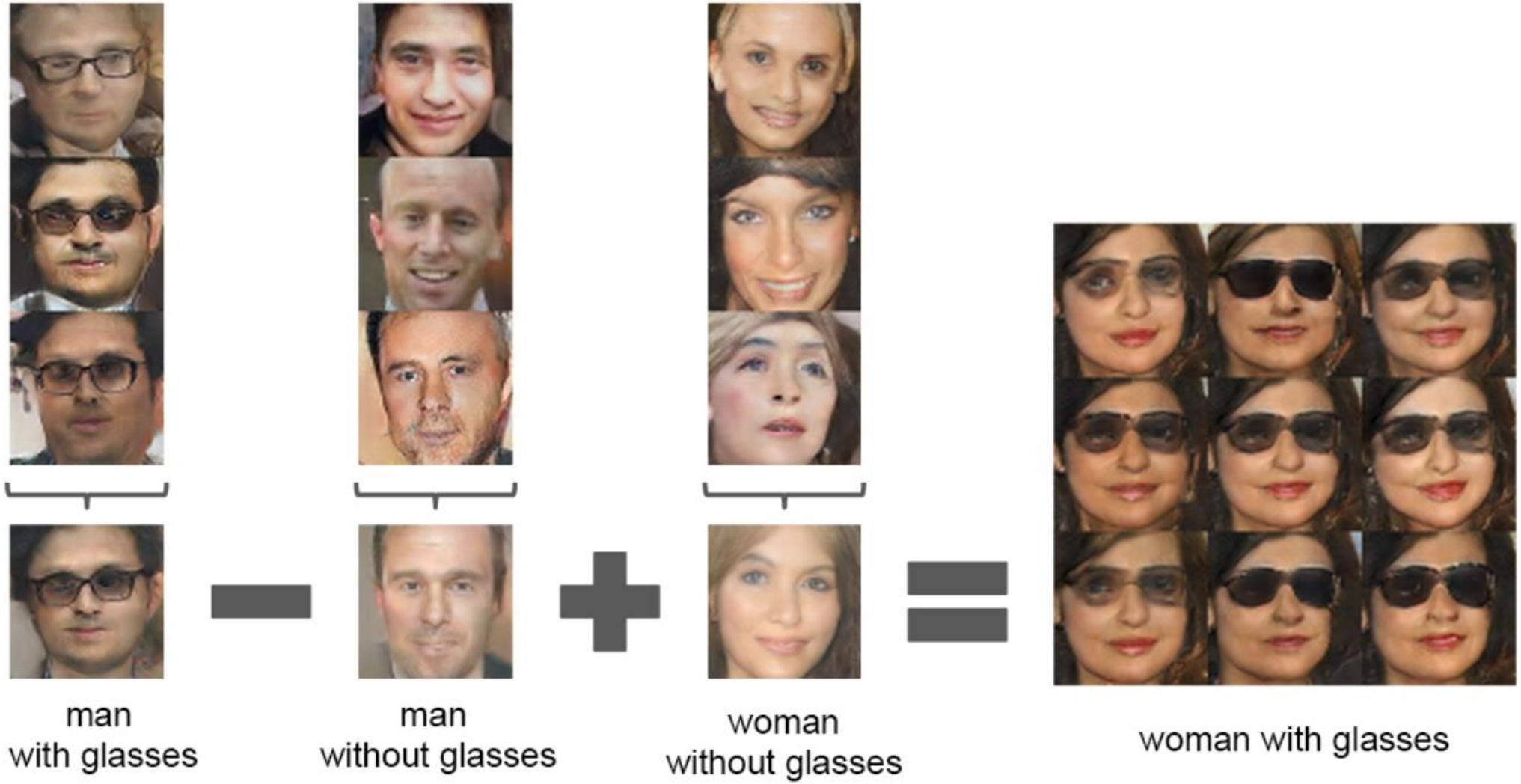
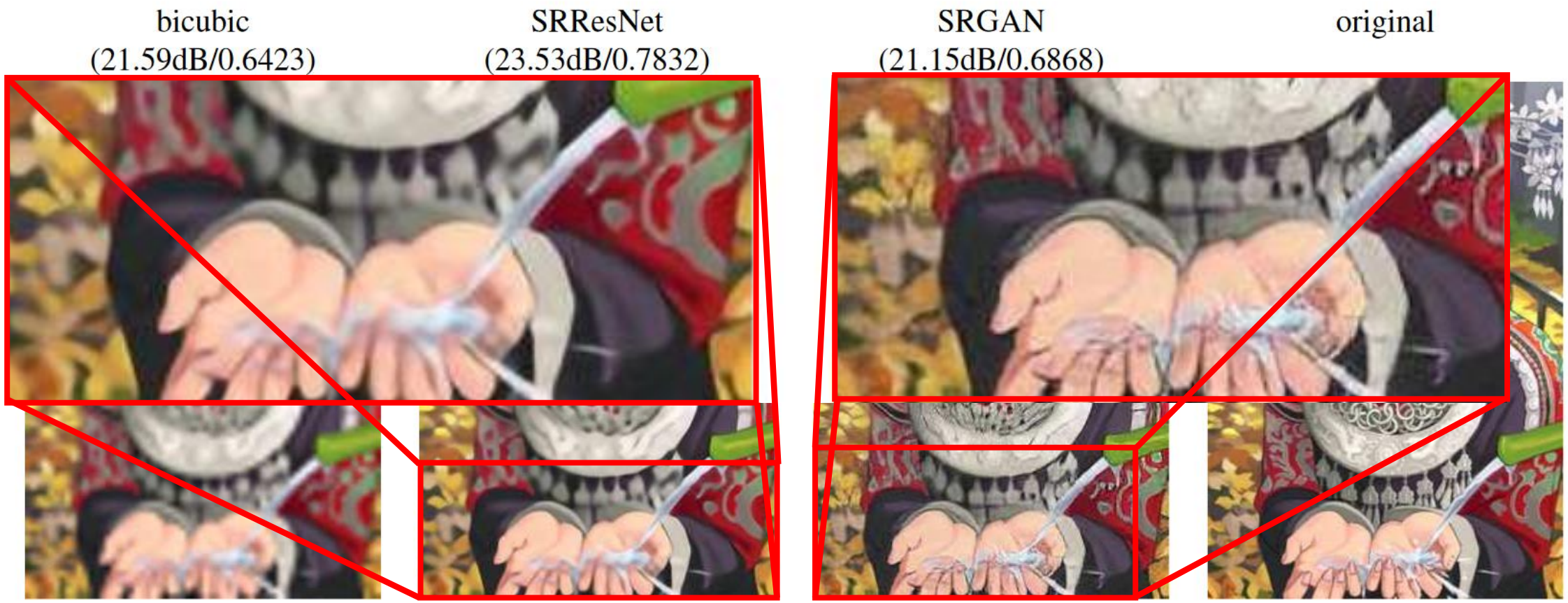
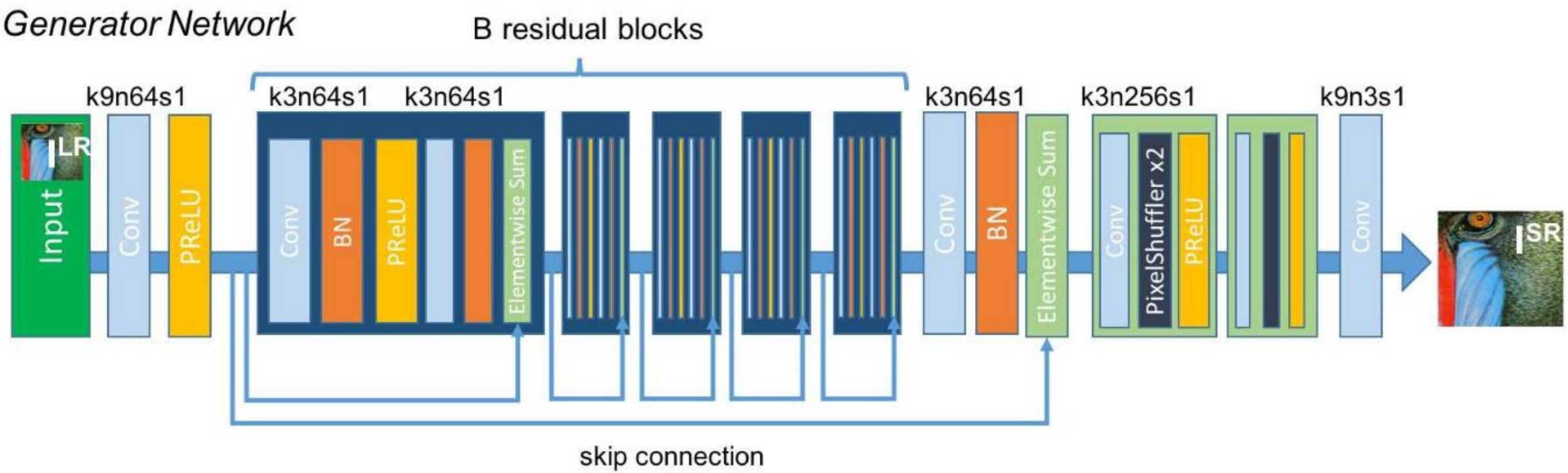


Image Super-Resolution [Ledig et al., CVPR 2017]



Loss = MSE + Adversarial loss

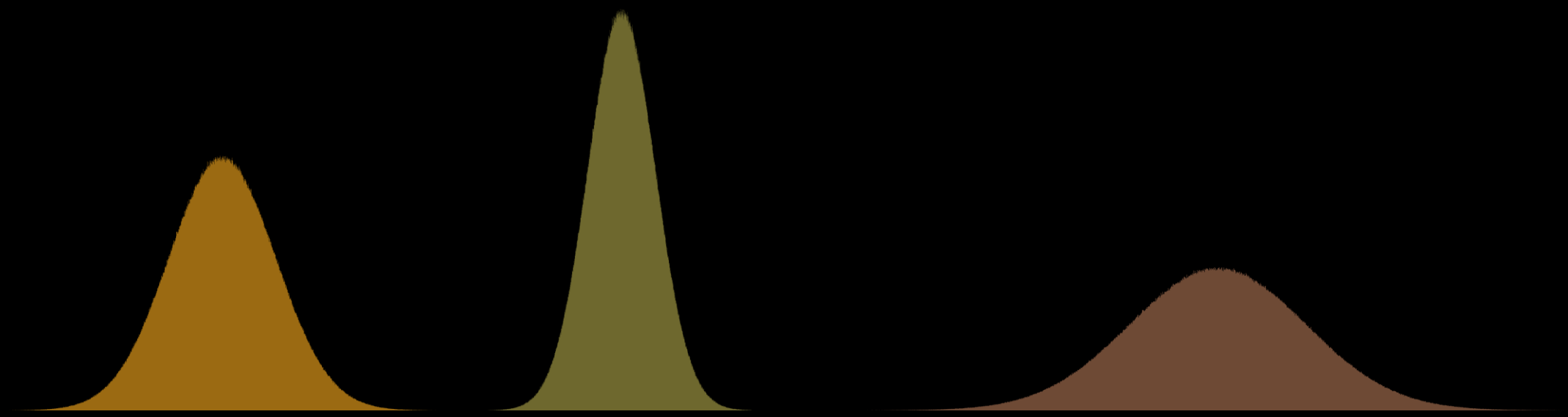
Image Super-Resolution [Ledig et al., CVPR 2017]



Principles of Estimation

1. Explicit parametric models and likelihood
2. Implicit models
3. Misspecification
4. Dimensional misspecification

Classic parametric models



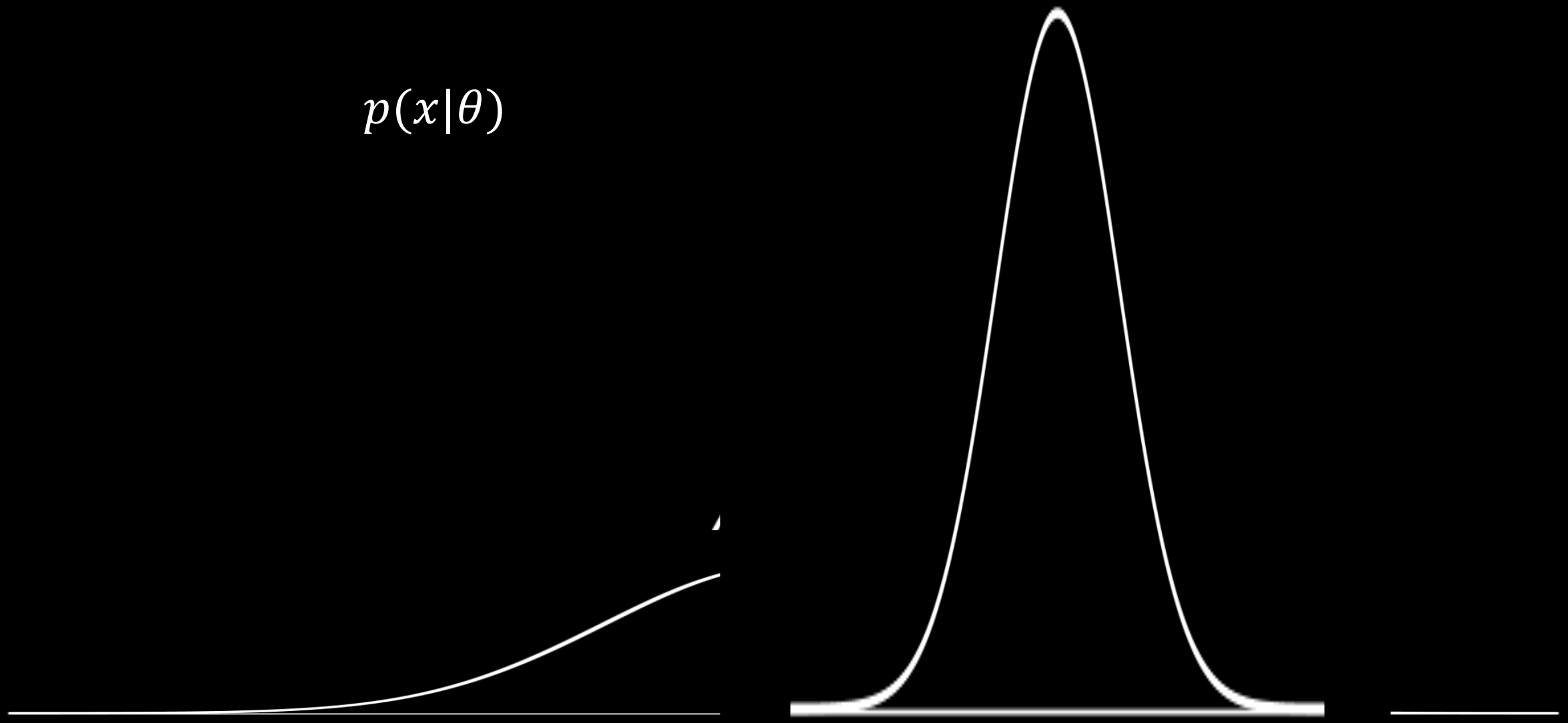
- Density function available
- Limited expressive power
- Mature field in statistics and learning theory

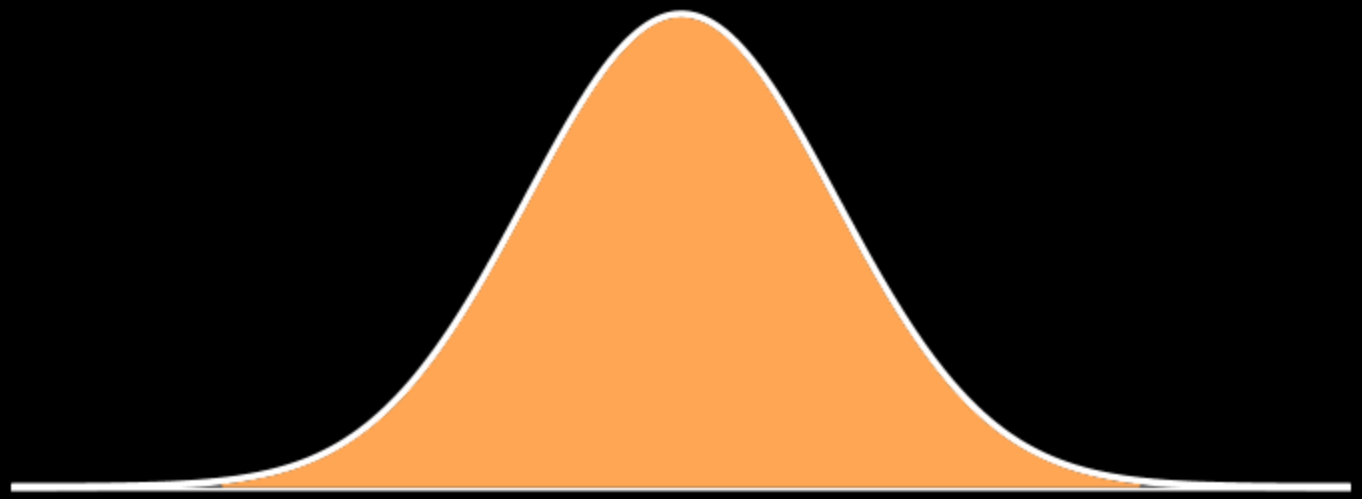
Likelihood and Maximum Likelihood Estimation (MLE)

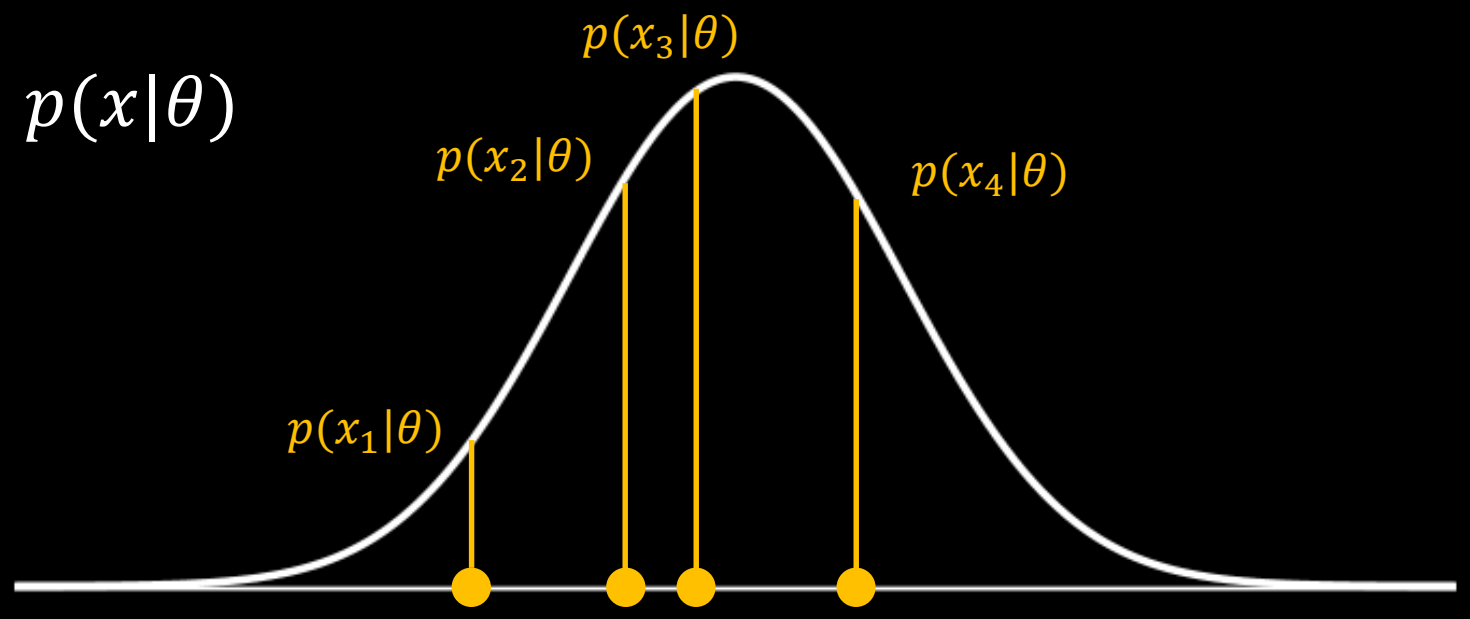
[Fisher, 1929]

- Likelihood is a function of the model parameters
- MLE extremely successful:
  e.g. least squares and cross-entropy are MLE estimators
- Direct link to the Kullback Leibler divergence

$p(x|\theta)$







Maximize the *likelihood function*

$$L(\theta) = \prod_i p(x_i|\theta)$$

$$\log L(\theta) = \log \prod_i p(x_i|\theta)$$

$$\log L(\theta) = \sum_i \log p(x_i|\theta)$$

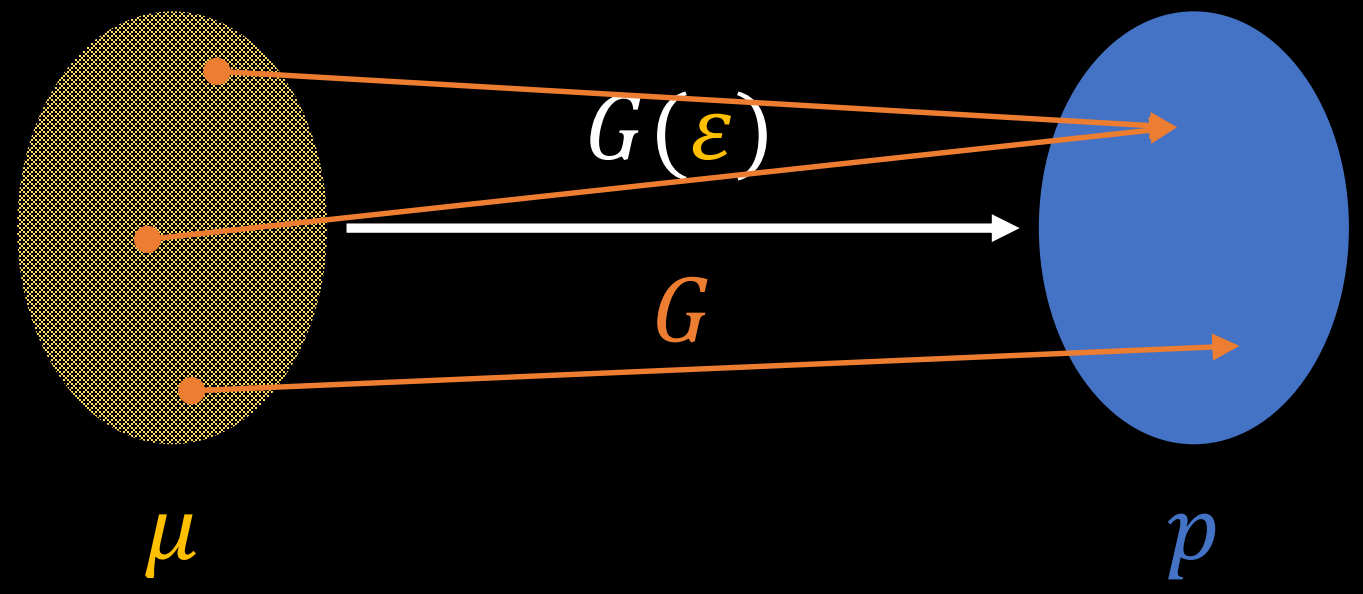
Implicit Model / Neural Sampler / Likelihood-free Model

- Highly expressive model class (universal)
- Density function not defined or intractable
- Lack of theory and learning algorithms compared to explicit models
- Basis for generative adversarial networks (GANs) and other recent models

Implicit Models

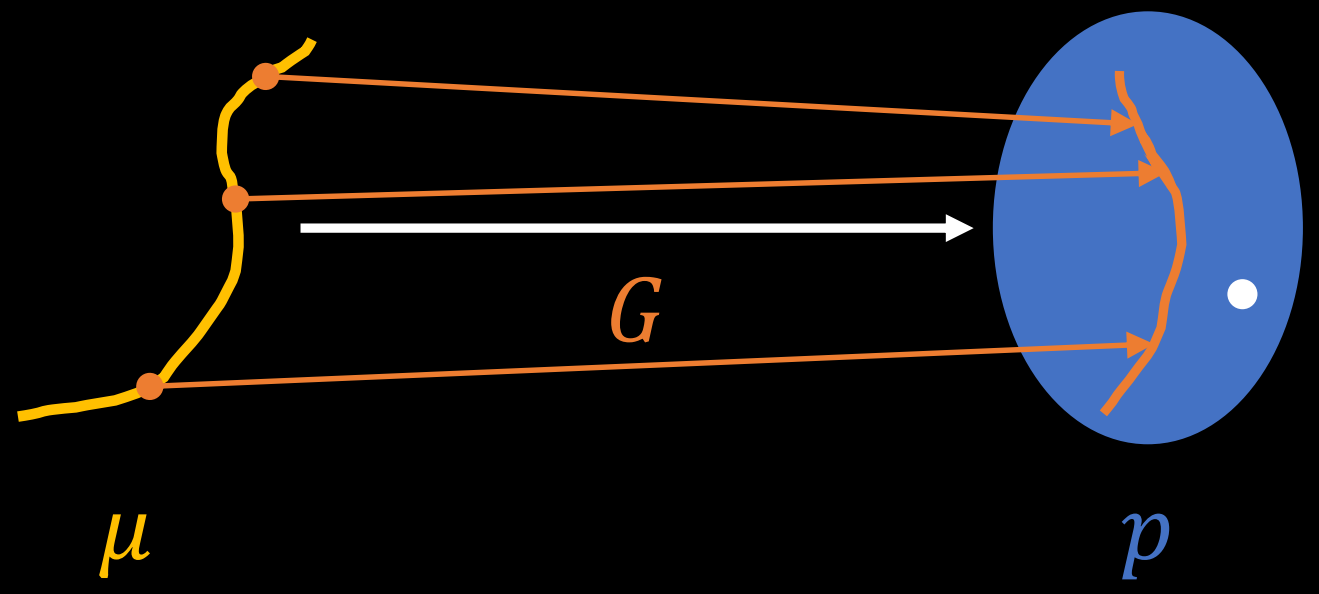
1. Diggle and Gratton (1984). Monte Carlo methods of inference for implicit statistical models. *JRSS B*
2. Goodfellow et al. (2014). Generative Adversarial Nets. NIPS
3. Mohamed and Lakshminarayanan (2016). Learning in Implicit Generative Models. *arXiv:1610.03483*

Problem 1: Non-Invertible Map



$$p(x) = \int_{z \in G^{-1}(x)} \mu(z)\, \mathrm{d}z$$

Problem 2: Lack of Density



$p(x)$ not defined a.e.

Misspecification

- A model is *well-specified* if and only if the model family contains the *true distribution*

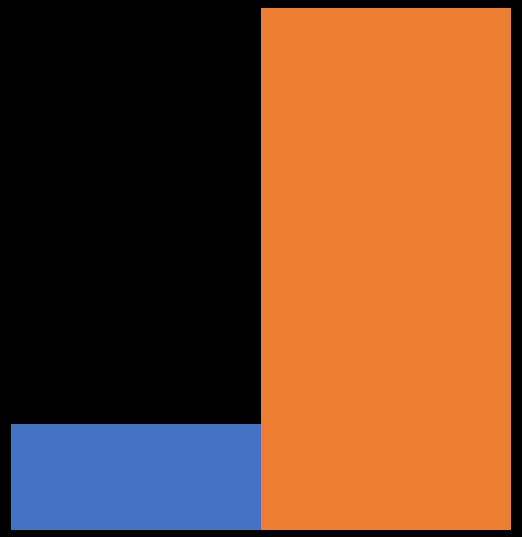
$$P^* \in \{ P_\theta \mid \theta \in \Theta \}$$

- If this is not the case, the model is *misspecified*
- For real data, most models are misspecified

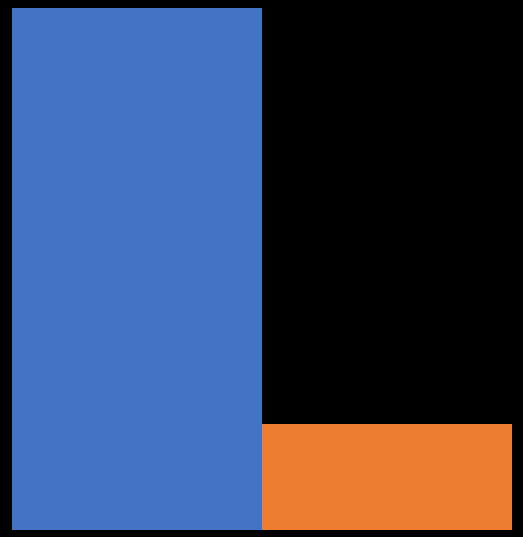
Well-developed theory in:
[White, “Estimation, Inference, and Specification Analysis”, Cambridge University Press, 1994]

EURO CENT

Misspecification: Elementary Example



a



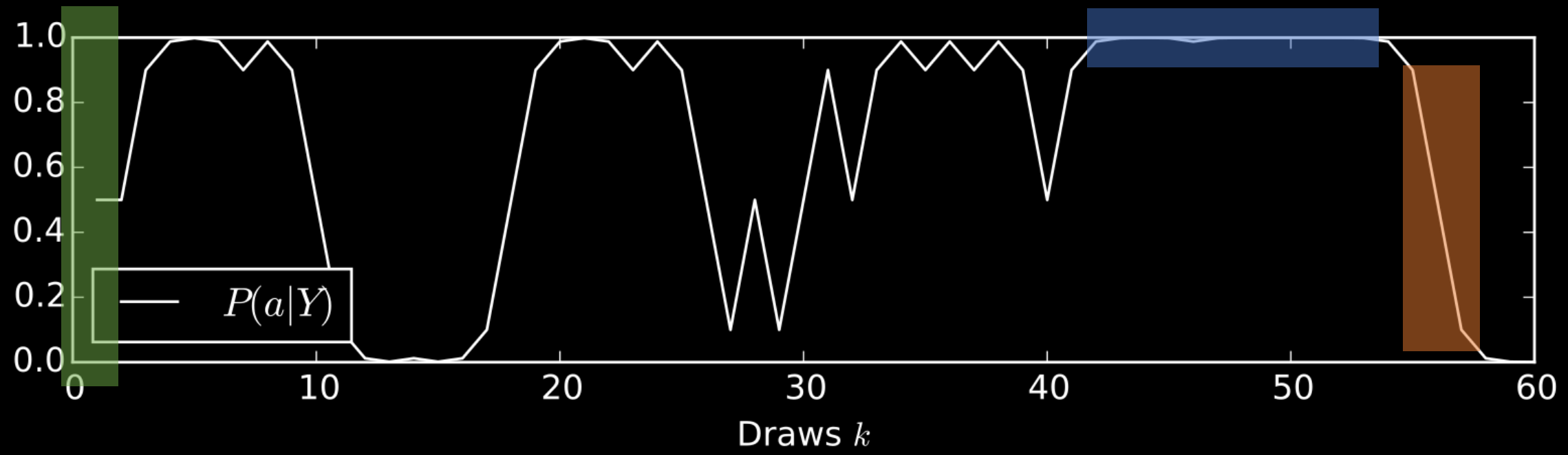
b

Example due to Cosma Shalizi and Bradford DeLong

Model

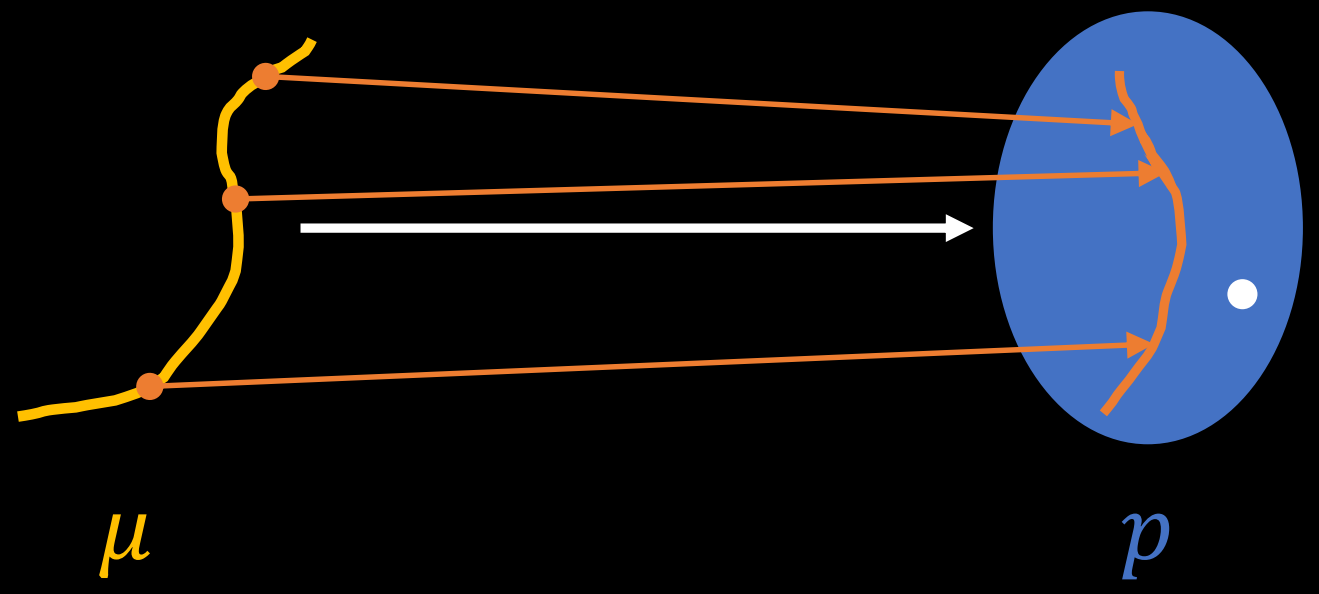
- $Y = (y_1, y_2, \dots)$ coin outcomes
- n: total coin flips
- k: total heads observed
- $P(Y|a) = \binom{n}{k} p_a^k (1 - p_a)^{n-k}$
- $P(a|Y) = \frac{P(Y|a)P(a)}{P(Y)}$
- Prior: uniform $p(a) = p(b) = {}^1/_2$

Posterior under a fair coin



1. Prior predictive at time zero is perfect.
2. Posterior beliefs concentrate, distance to truth increases in expectation but beliefs revised infinitely often.
3. Posterior beliefs are revised entirely by single observations.

Dimensional Misspecification

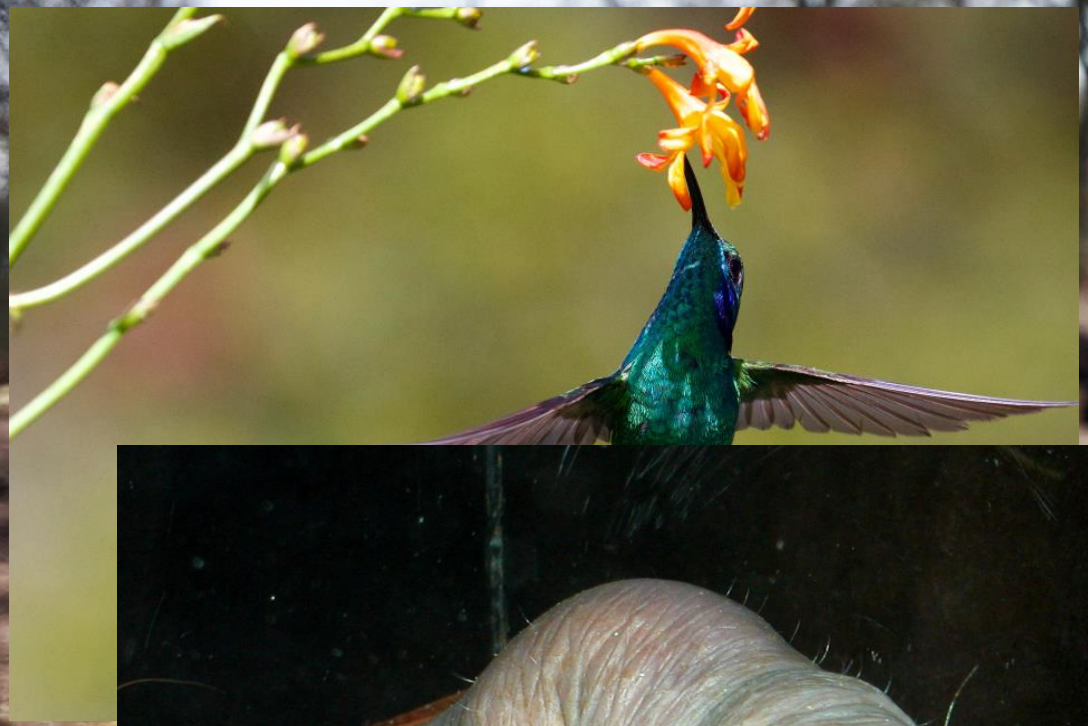


GAN Models

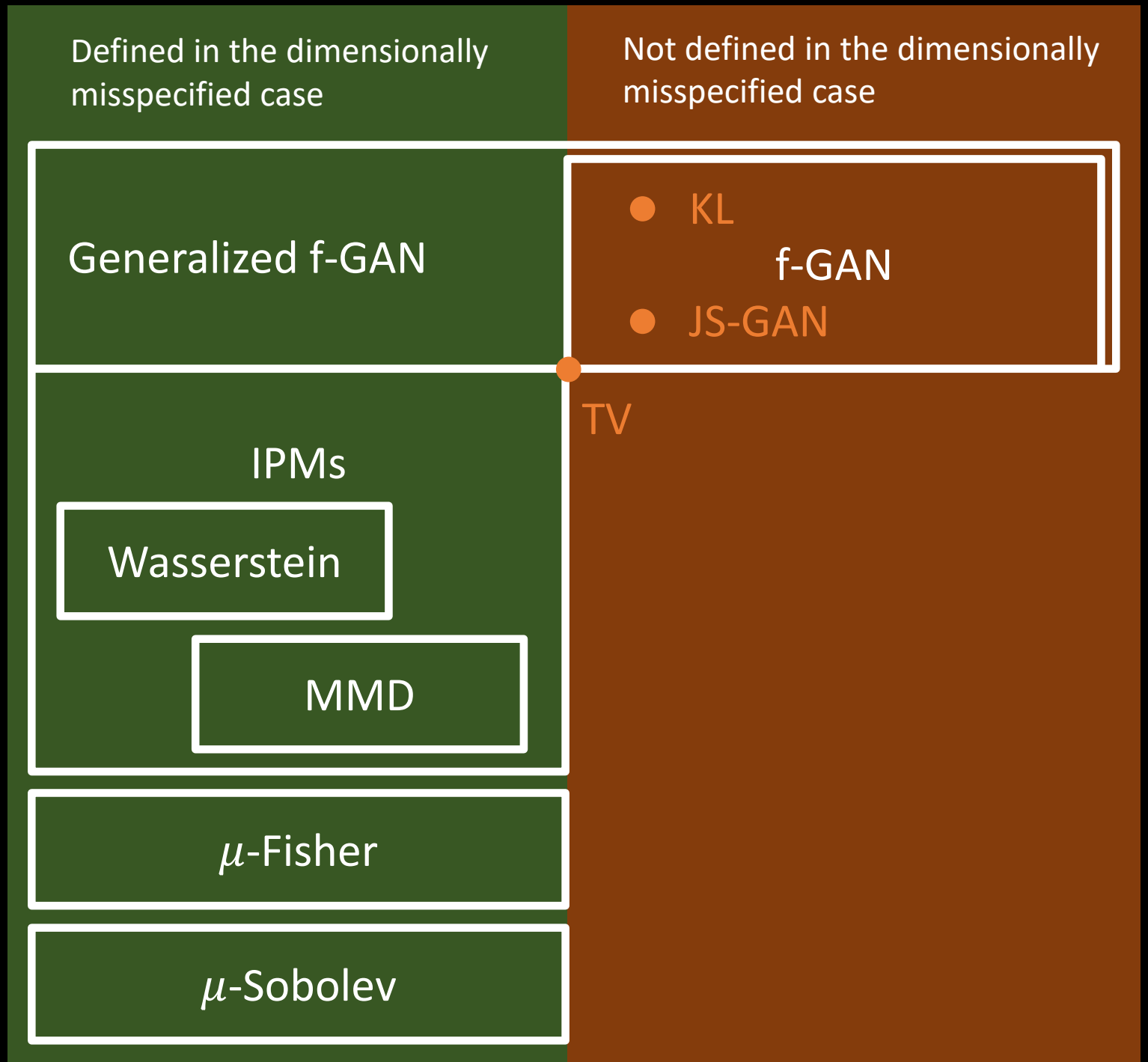
1. Divergences
2. Integral Probability Metrics
3. Generalizations and other variants

The GAN Zoo

https://github.com/hindupuravinash/the-gan-zoo



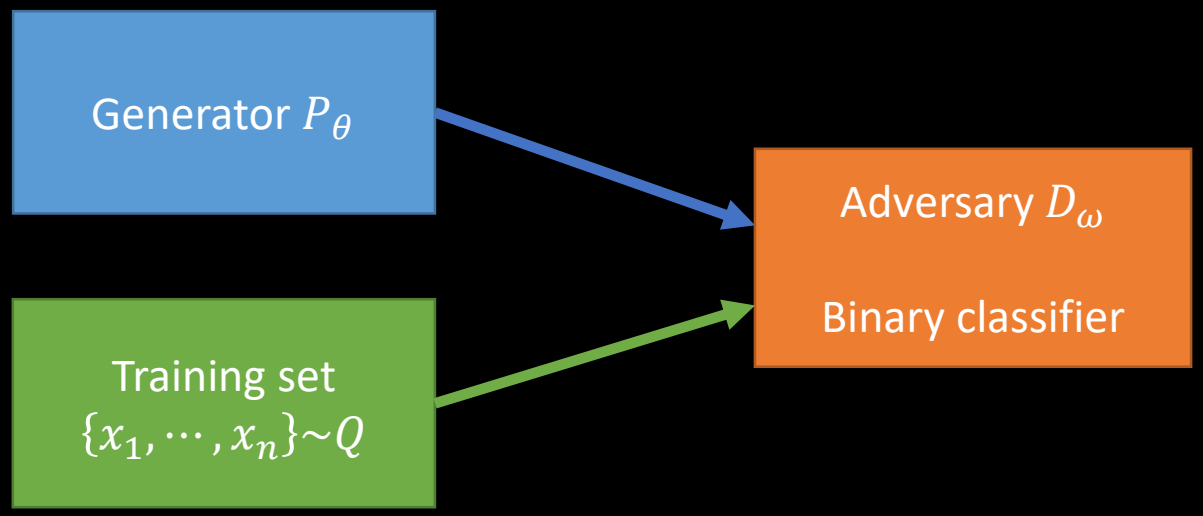
Divergences


Defined in the dimensionally misspecified case
Not defined in the dimensionally misspecified case
Generalized f-GAN
KL
f-GAN
JS-GAN
TV
IPMs
Wasserstein
MMD
μ-Fisher
μ-Sobolev

Jensen-Shannon GAN

[Goodfellow et al., NIPS 2013]

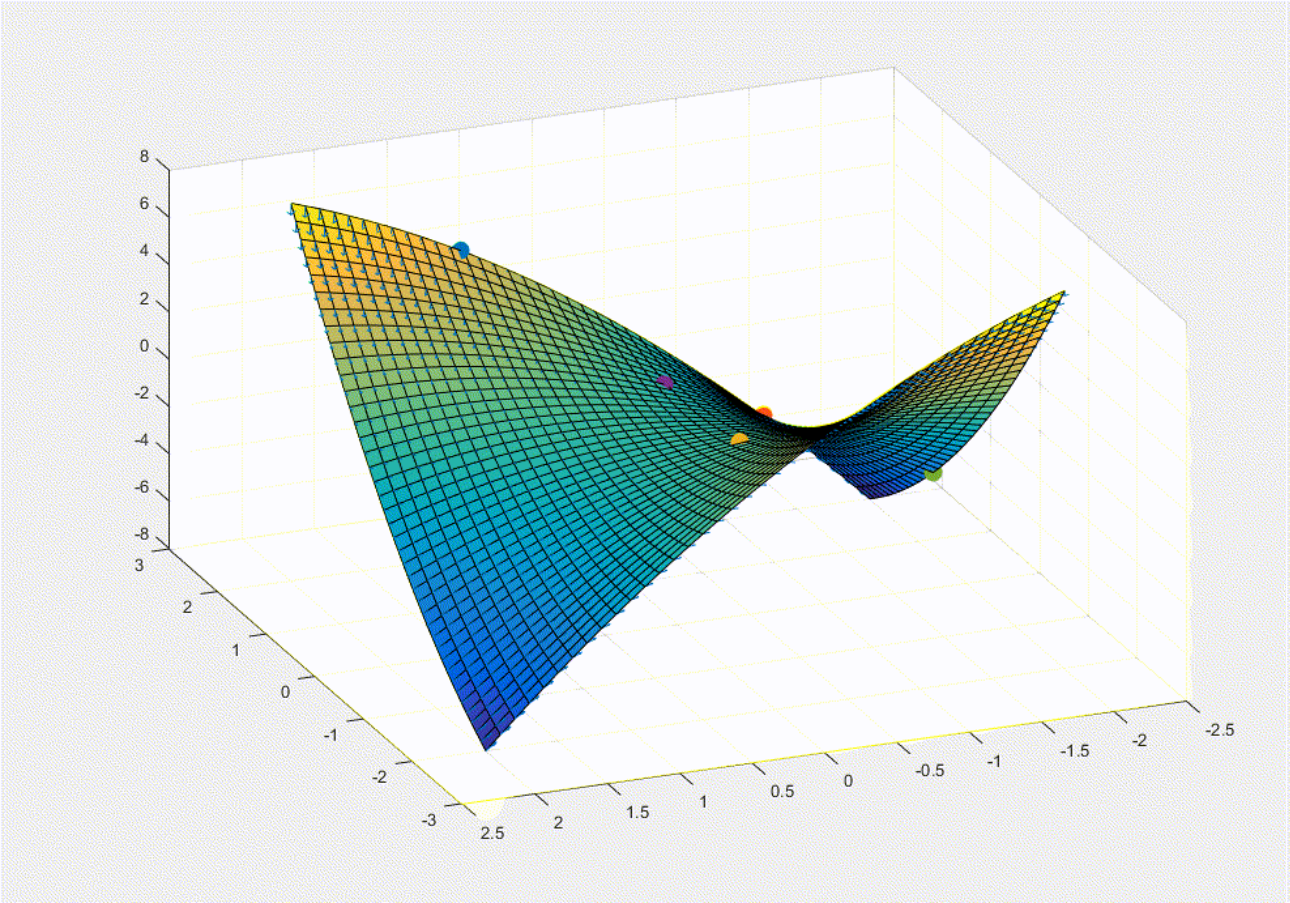
GAN Training Objective [Goodfellow et al., 2014]

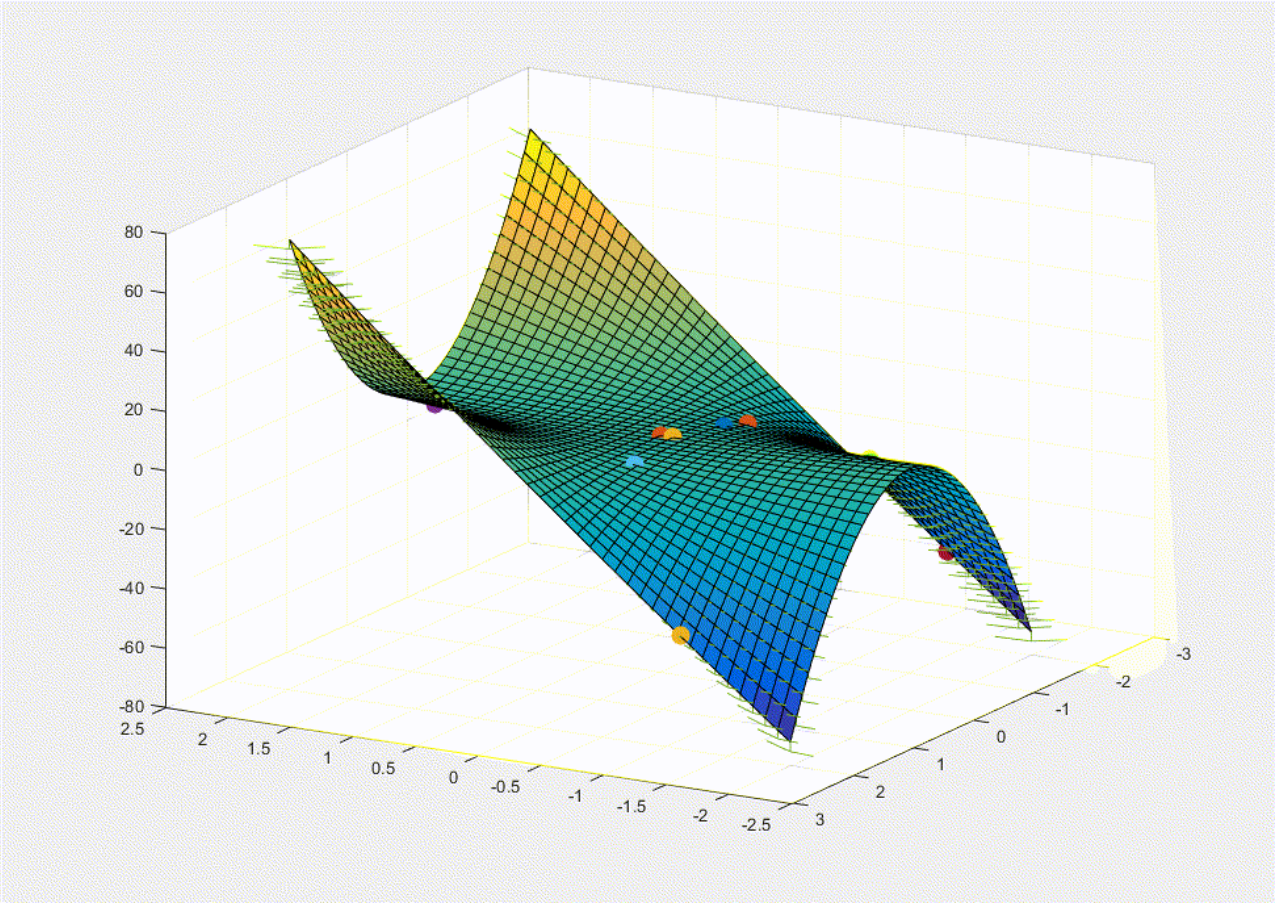


- Generator tries to fool discriminator (i.e. generate realistic samples)
- Discriminator tries to distinguish fake from real samples
- Saddle-point problem

$$\min_\theta \max_\omega \mathbb{E}_{x \sim P_\theta}[\log D_\omega(x)] + \mathbb{E}_{x \sim Q}[\log(1 - D_\omega(x))]$$

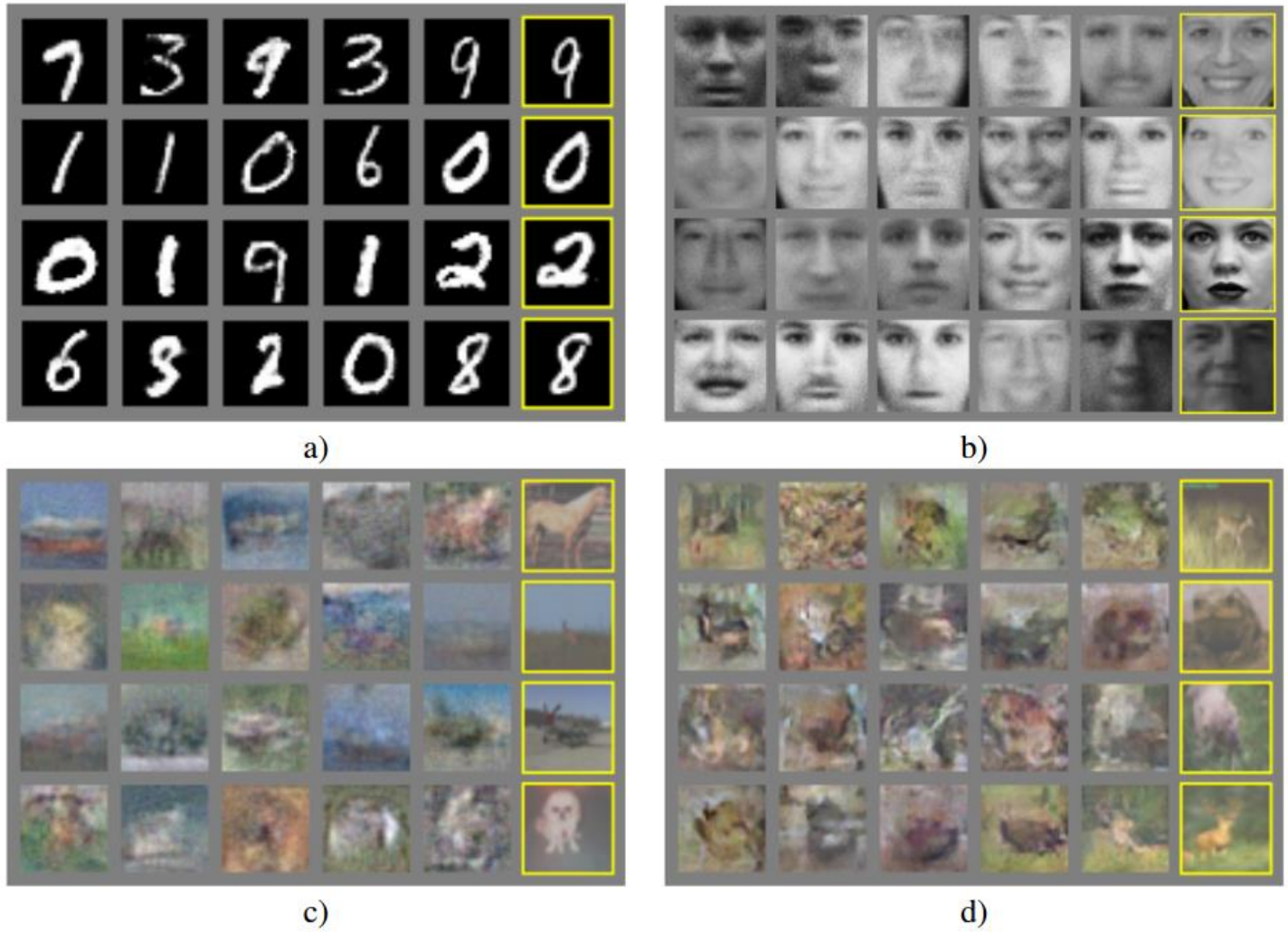
Optimization of Saddle Point Objectives





- Practical difficulty: no monotonic objective
- Theoretical difficulties: which algorithm to use?

Goodfellow’s 2014 results



- JS-GAN

f-GAN

[Nowozin, Cseke, Tomioka, NIPS 2016]

f-Divergences

- Divergence between two probability densities
  [Csiszar and Shields, 2004], [Liese and Vajda, IEEE Inf Th, 2006]

$$D_f(P \parallel Q) = \int_{\mathcal{X}} q(x)\, f\left(\frac{p(x)}{q(x)}\right)$$

- Scalar "*generator function*" $f: \mathbb{R}_+ \to \mathbb{R}$ is
  - Convex, lower-semicontinuous
- Assumptions
  - Both distributions have a density function wrt Lebesgue measure

f-divergences

Name	$D_f(P\|Q)$	Generator $f(u)$
Total variation	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \,\mathrm{d}x$	$\frac{1}{2}\lvert u-1\rvert$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x$	$u\log u$
Reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\,\mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)}\,\mathrm{d}x$	$(u-1)^2$
Neyman χ^2	$\int \frac{(p(x)-q(x))^2}{q(x)}\,\mathrm{d}x$	$\frac{(1-u)^2}{u}$
Squared Hellinger	$\int \left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2 \mathrm{d}x$	$(\sqrt{u}-1)^2$
Jeffrey	$\int (p(x)-q(x))\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x$	$(u-1)\log u$
Jensen-Shannon	$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x$	$-(u+1)\log\frac{1+u}{2}+u\log u$
Jensen-Shannon-weighted	$\int p(x)\pi\log\frac{p(x)}{\pi p(x)+(1-\pi)q(x)}+(1-\pi)q(x)\log\frac{q(x)}{\pi p(x)+(1-\pi)q(x)}\,\mathrm{d}x$	$\pi u\log u-(1-\pi+\pi u)\log(1-\pi+\pi u)$
GAN	$\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x-\log(4)$	$u\log u-(u+1)\log(u+1)$
α-divergence ($\alpha\notin\{0,1\}$)	$\frac{1}{\alpha(\alpha-1)}\int\left(p(x)\left[\left(\frac{q(x)}{p(x)}\right)^{\alpha}-1\right]-\alpha(q(x)-p(x))\right)\mathrm{d}x$	$\frac{1}{\alpha(\alpha-1)}\left(u^{\alpha}-1-\alpha(u-1)\right)$

Estimating f-divergences from samples

[Nguyen, Wainwright, Jordan, 2010]

- Divergence between two distributions

$$D_f(P \parallel Q) = \int_{\mathcal{X}} q(x)\, f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$$

f : generator function (convex & $f(1)=0$)

- Every convex function f has a *Fenchel conjugate* f^* so that

$$f(u) = \sup_{t \in \mathrm{dom}_{f^*}} \{tu - f^*(t)\}$$

"any convex f can be represented as point-wise max of *linear* functions"

Estimating f-divergences from samples (cont)

[Nguyen, Wainwright, Jordan, 2010]

$$
\begin{aligned}
D_f(P \parallel Q) &= \int_{\mathcal{X}} q(x)\, f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \\
&= \int_{\mathcal{X}} q(x) \sup_{t \in \mathrm{dom}_{f^*}} \left\{ t \frac{p(x)}{q(x)} - f^*(t) \right\} \mathrm{d}x \\
&\geq \sup_{T \in \mathcal{T}} \left(\int_{\mathcal{X}} p(x)\, T(x)\, \mathrm{d}x - \int_{\mathcal{X}} q(x)\, f^*\big(T(x)\big)\, \mathrm{d}x \right) \\
&= \sup_{T \in \mathcal{T}} \Big(\underbrace{\mathbb{E}_{x \sim P}[T(x)]}_{} - \underbrace{\mathbb{E}_{x \sim Q}[f^*(T(x))]}_{} \Big)
\end{aligned}
$$

Approximate using: samples from P samples from Q

f-GAN and GAN objectives [Nowozin, Cseke, Tomioka, NIPS 2016]

- GAN

$$\min_{\theta} \max_{\omega} \mathbb{E}_{x \sim P_\theta}[\log D_\omega(x)] + \mathbb{E}_{x \sim Q}[\log(1 - D_\omega(x))]$$

- f-GAN

$$\min_{\theta} \max_{\omega} \left(\mathbb{E}_{x \sim P_\theta}[T_\omega(x)] - \mathbb{E}_{x \sim Q}[f^*(T_\omega(x))] \right)$$

  - GAN discriminator-variational function correspondence: $\log D_\omega(x) = T_\omega(x)$
  - GAN minimizes the Jensen-Shannon divergence (which was also pointed out in Goodfellow et al., 2014)

Kullback Leibler example

$$D_{\mathrm{KL}}(P \parallel Q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$$

$$D_{\mathrm{KL}}(P \parallel Q) = \int p(x) \left(-\log \frac{q(x)}{p(x)} \right) \mathrm{d}x$$

$$f^*(t) = \sup_{u \in \mathrm{dom}_f} \{ut - f(u)\}$$

$$f^*(t) = \sup_{u>0} \{ ut + \log u \}$$

Kullback Leibler example

$$f^*(t) = \sup_{u>0} \{ ut + \log u \}$$

$$\frac{\partial}{\partial u}(ut + \log u) = 0 \quad \Rightarrow \quad u = -\frac{1}{t}$$

$$f^*(t) = \sup_{u>0} \{ ut + \log u \} = -\frac{1}{t}t + \log -\frac{1}{t} = -1 - \log -t$$

$$\mathrm{dom}_{f^*} = \{ t \mid t < 0 \}$$

Kullback Leibler example

- Generic form

$$D_f(P \parallel Q) = \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{x \sim P}[T(x)] - \mathbb{E}_{x \sim Q}[f^*(T(x))] \right)$$

- Kullback-Leibler f-functions

$f(u) = -\log u$ and $f^*(t) = -1 - \log(-t)$

- Yields

$$D_{\mathrm{KL}}(P \parallel Q) = \sup_{T<0} \left(\mathbb{E}_{x \sim P}[T(x)] + \mathbb{E}_{x \sim Q}\left[\log\left(-T(x)\right)\right] + 1 \right)$$

f-GAN

Name	Output activation g_f	dom_{f^*}	Conjugate $f^*(t)$	$f'(1)$
Total variation	$\frac{1}{2}\tanh(v)$	$-\frac{1}{2} \leq t \leq \frac{1}{2}$	t	0
Reverse KL	v	$\mathbb{R}$	$\exp(t-1)$	1
Kullback-Leibler (KL)	$-\exp(v)$	$\mathbb{R}_-$	$-1-\log(-t)$	-1
Pearson χ^2	v	$\mathbb{R}$	$\frac{1}{4}t^2+t$	0
Neyman χ^2	$1-\exp(v)$	$t<1$	$2-2\sqrt{1-t}$	0
Squared Hellinger	$1-\exp(v)$	$t<1$	$\frac{t}{1-t}$	0
Jeffrey	v	$\mathbb{R}$	$W(e^{1-t})+\frac{1}{W(e^{1-t})}+t-2$	0
Jensen-Shannon	$\log(2)-\log(1+\exp(-v))$	$t<\log(2)$	$-\log(2-\exp(t))$	0
Jensen-Shannon-weighted	$-\pi\log\pi-\log(1+\exp(-v))$	$t<-\pi\log\pi$	$(1-\pi)\log\frac{1-\pi}{1-\pi e^{t/\pi}}$	0
GAN	$-\log(1+\exp(-v))$	$\mathbb{R}_-$	$-\log(1-\exp(t))$	$-\log(2)$
α-div. ($\alpha<1, \alpha\neq 0$)	$\frac{1}{1-\alpha}-\log(1+\exp(-v))$	$t<\frac{1}{1-\alpha}$	$\frac{1}{\alpha}(t(\alpha-1)+1)^{\frac{\alpha}{\alpha-1}}-\frac{1}{\alpha}$	0
α-div. ($\alpha>1$)	v	$\mathbb{R}$	$\frac{1}{\alpha}(t(\alpha-1)+1)^{\frac{\alpha}{\alpha-1}}-\frac{1}{\alpha}$	0

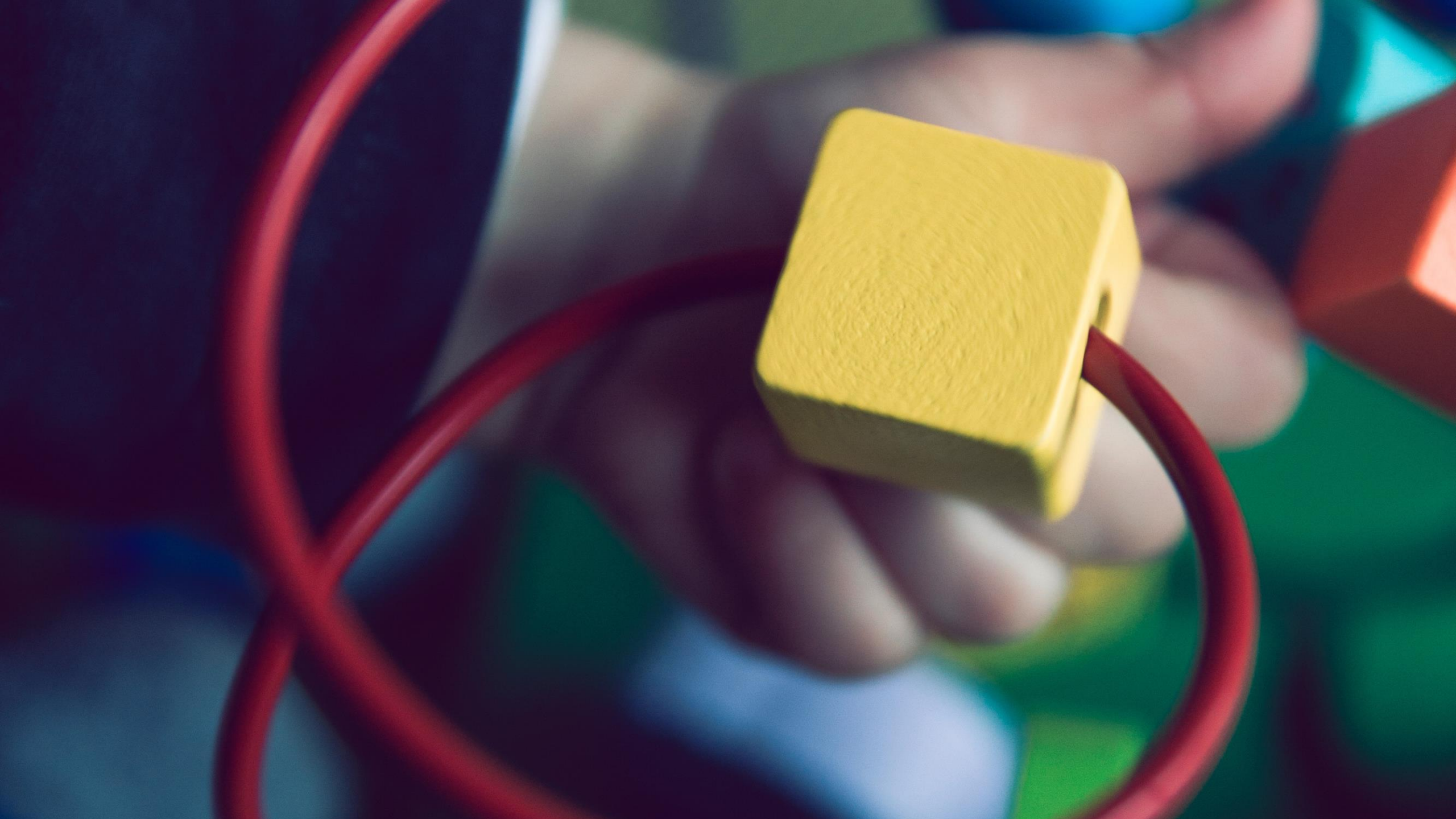
f-divergences, key properties

- Invariance to coordinate transformations
- Come in pairs:

$$D_f(P \parallel Q) = D_g(Q \parallel P)$$

$$g(u) = u\, f\left(\frac{1}{u}\right)$$

- KL
  f-GAN
- JS-GAN



f-GAN 1D example

Jupyter notebook

https://github.com/nowozin/mlss2018-madrid-gan/blob/master/GAN%20-%201D%20Example.ipynb

f-GAN CIFAR-10 example

Jupyter notebook

https://github.com/nowozin/mlss2018-madrid-gan/blob/master/GAN%20-%20CIFAR.ipynb

DCGAN architecture on TinyImageNet

- 64x64 DCGAN, 1.5M iterations

Integral Probability Metrics (IPM)

[Sriperumbudur et al., JMLR 2010]

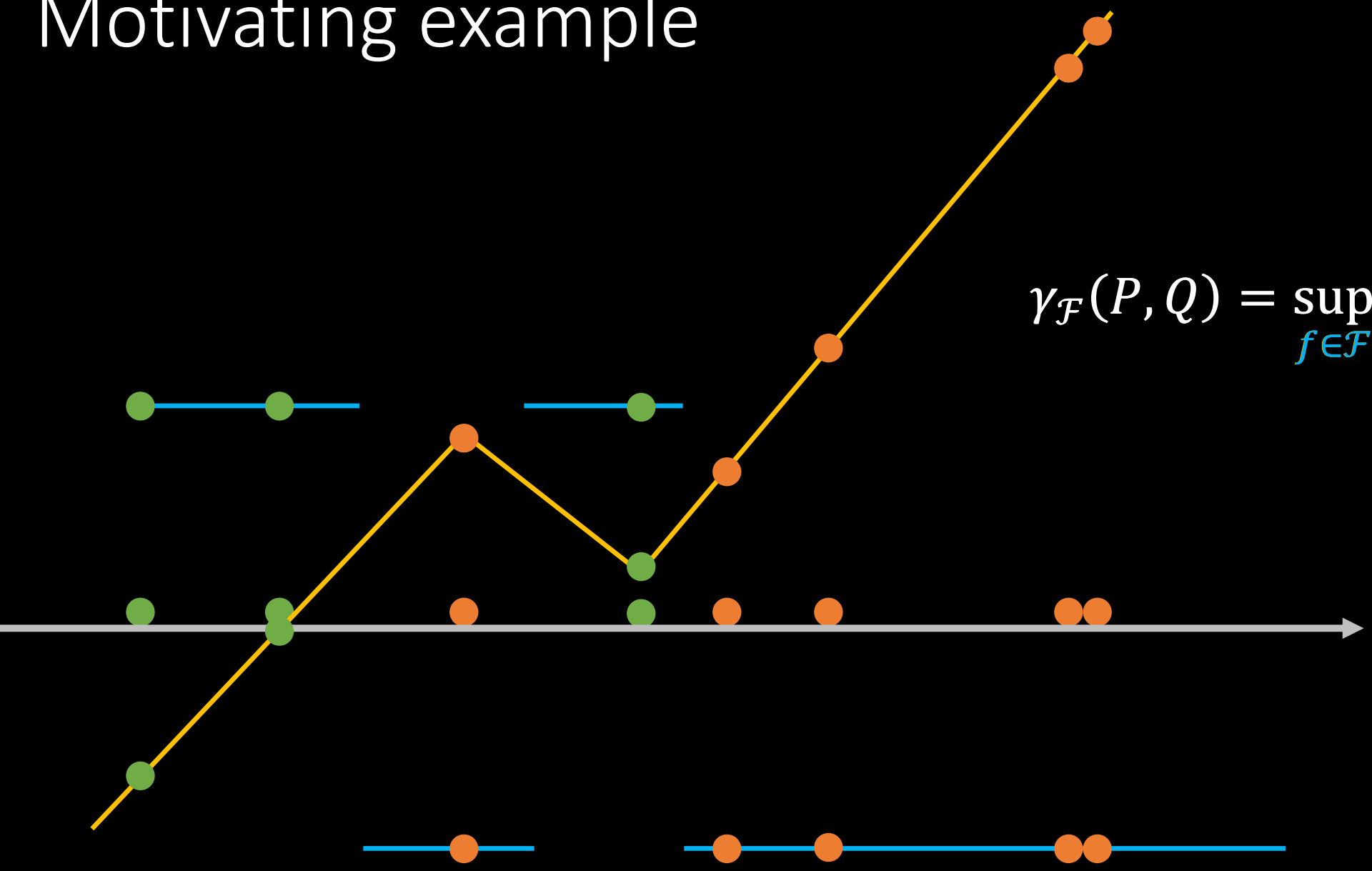
Integral Probability Metrics (IPMs)

- $\mathcal{F}$ class of real-valued bounded measurable functions
- P, Q probability measures

$$\gamma_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \int f \mathrm{d}P - \int f \mathrm{d}Q \right|$$

- See [Sriperumbudur et al., 2009, arXiv:0901.2698] and [Sriperumbudur et al., 2012, Electr. Journal of Stat.]
- Choice of $\mathcal{F}$ determines metric

Motivating example



$$\gamma_{\mathcal{F}}(P,Q) = \sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}P - \int f \, \mathrm{d}Q \right|$$

KL
f-GAN
JS-GAN
TV
IPMs

IPMs: Maximum Mean Discrepancy (MMD)

RKHS Norm

- [Gretton et al., "A Kernel Two-sample Test", JMLR 2012]
- If $\mathcal{F}$ is a unit-ball in a reproducing kernel Hilbert space $\mathcal{H}$ we have

$$\gamma_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} \left| \int f \mathrm{d}P - \int f \mathrm{d}Q \right| = \left\| \mu_P - \mu_Q \right\|_{\mathcal{H}}$$

- Kernel mean embedding of a probability measure

$$\mu_P = \int k(x,\cdot) P(\mathrm{d}x)$$

- Estimator given sample $X = (x_1, \cdots, x_N)$ and $Y = (y_1, \cdots, y_M)$

$$\widehat{\mathrm{MMD}}^2(X,Y) = \frac{1}{N(N-1)} \sum_{n\neq n'} k(x_n, x_{n'}) + \frac{1}{M(M-1)} \sum_{m\neq m'} k(y_m, y_{m'}) - \frac{2}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} k(x_n, y_m)$$

Kernel MMD Training in Deep Learning

- Deep generative models
  [Li et al., 2015], [Dziugaite et al., 2015]
- Deep discriminative models
  [Bouchacourt et al., NIPS 2016]
- Use for model criticism
  [Sutherland et al., ICLR 2017]

- More discriminative kernel functions
  [Li et al., NIPS 2017, arXiv:1705.08584]

[Dziugaite et al., 2015]

- Neural MNIST/faces samples (RBF kernel)

DISCO Nets: DISsimilarity COefficients Networks

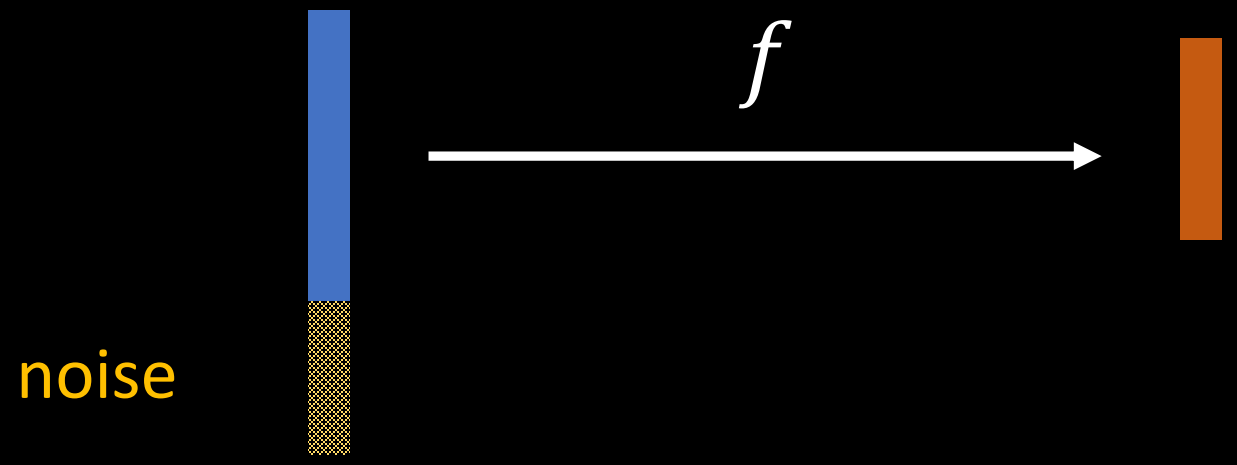
Diane Bouchacourt (Oxford)
Pawan Kumar (Oxford)
Sebastian Nowozin

NIPS 2016

arXiv:1606.02556

$$f\colon \mathcal{X} \to \mathcal{Y}$$

$$f\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$$



$$\min_f \ \mathbb{E}_{(x,y)} \mathbb{E}_{\varepsilon} \left[\ell(y, f(x, \varepsilon)) - \frac{1}{2} \mathbb{E}_{\varepsilon', \varepsilon''} [\ell(f(x, \varepsilon'), f(x, \varepsilon''))] \right]$$

Minimize Expected loss minus Diversity

Constructing Divergence from Loss

- Loss $\Delta(y, y')$
- True joint distribution $\mathrm{T}(x, y)$
- Model distribution $P(y|x)$
- Expected loss (diversity coefficient)

$$\mathrm{DIV}(Q, P) = \mathbb{E}_{x \sim T(x)}\left[\mathbb{E}_{y \sim Q(y|x)}\left[\mathbb{E}_{y' \sim P(y|x)}[\Delta(y, y')]\right]\right]$$

- Dissimilarity coefficient [Rao, 1982]

$$\mathrm{DISC}(Q, P) = \mathrm{DIV}(Q, P) - \gamma \mathrm{DIV}(P, P)$$

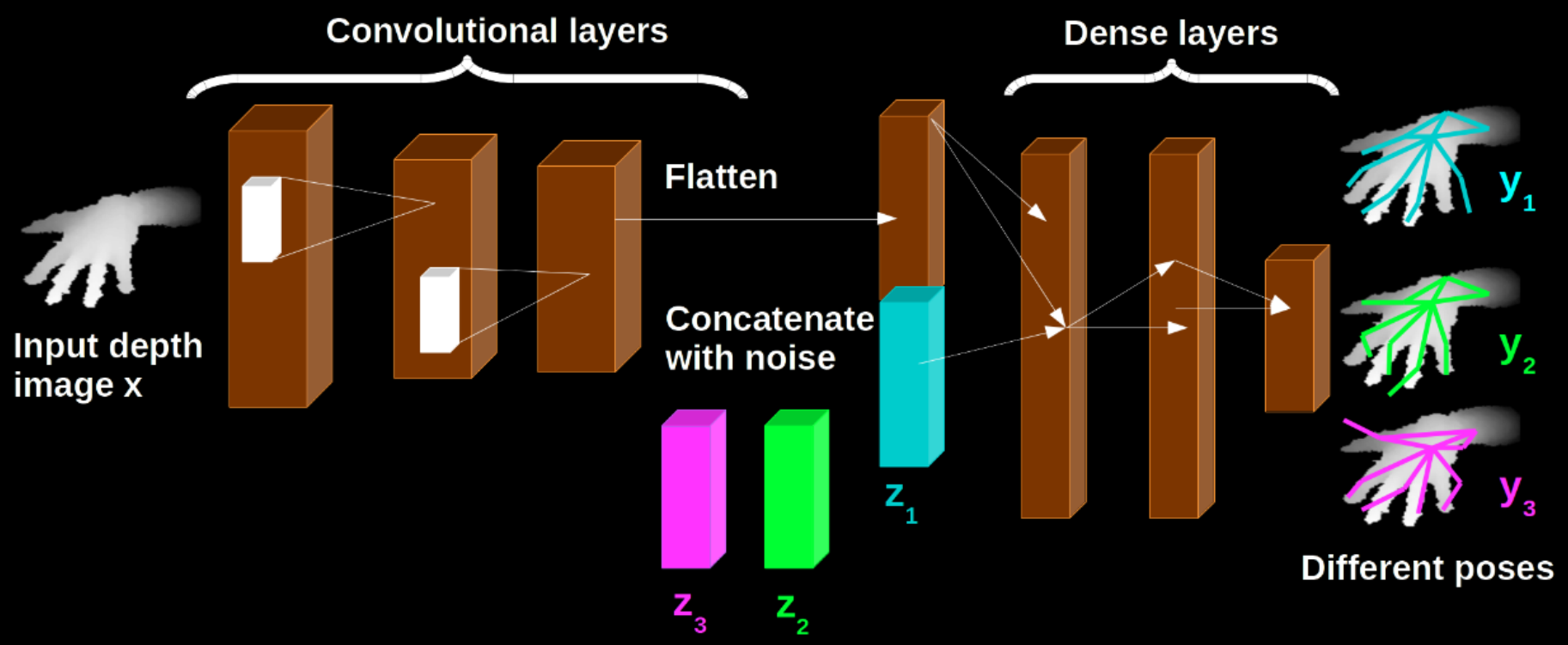
- $\gamma = \frac{1}{2}$

Relation to Scoring rules and Kernel MMD

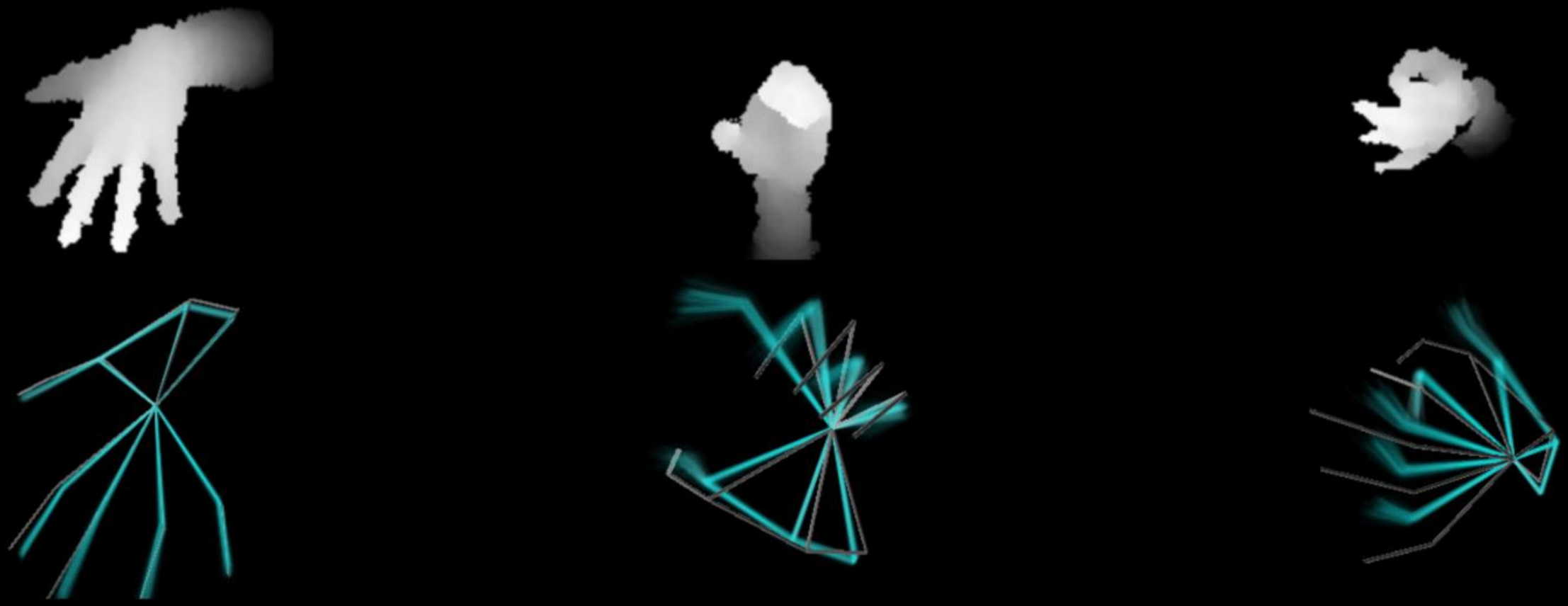
- Via [Gneiting and Raftery, 2007]:
  For $\Delta_\beta(y, y') = \|y - y'\|_2^\beta$, with $\beta \in (0,2)$ DISCO is a proper scoring rule.
- Via [Schölkopf, 2001]:
  For $k(y, y') = \|y - y'\|_2^\beta$, with $\beta \in (0,2)$, k is conditionally positive definite and DISCO is the kernel MMD objective with $k = \Delta$.

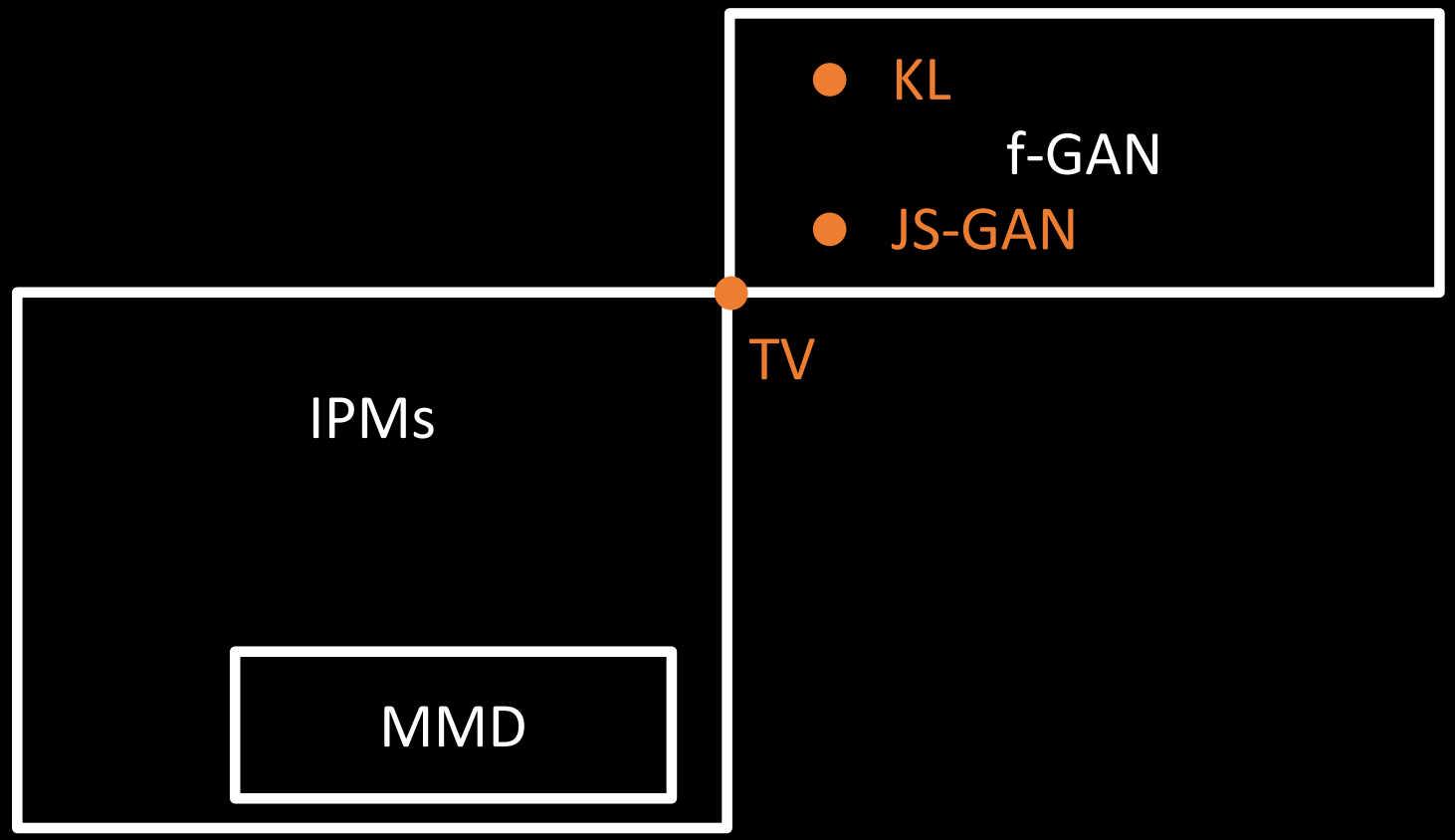
DISCO Nets

with Diane Bouchacourt, Pawan Kumar, NIPS 2016, arXiv:1606.02556

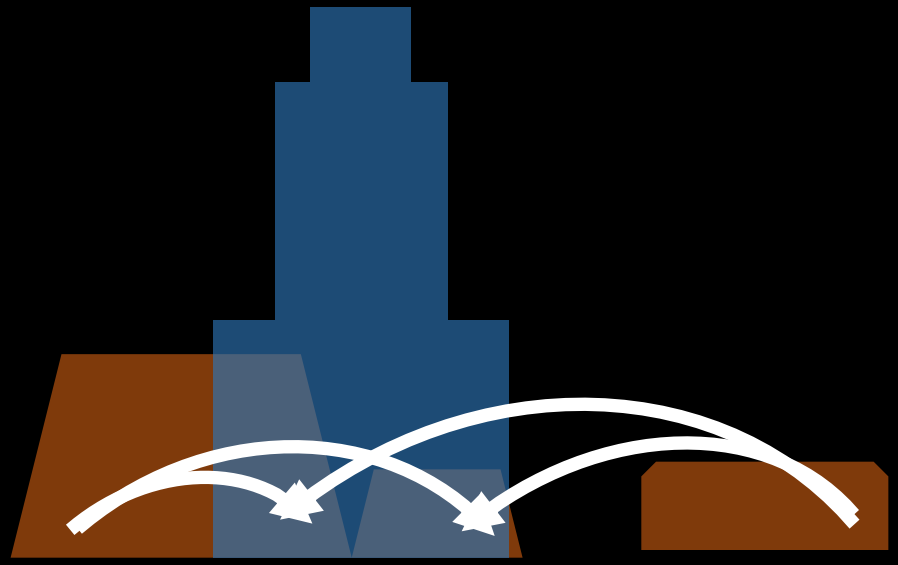


$$\mathrm{DISC}(Q,P) = \mathrm{DIV}(Q,P) - \gamma \mathrm{DIV}(P,P)$$



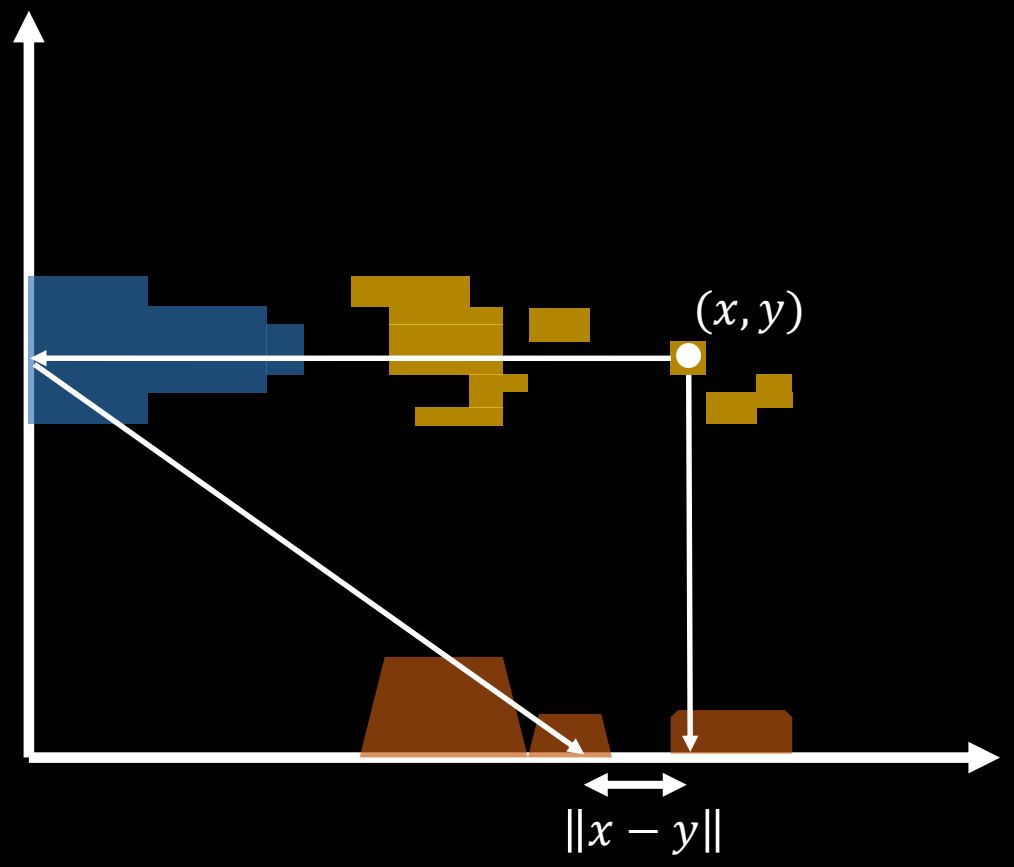

KL
f-GAN
JS-GAN
TV
IPMs
MMD

IPM: Wasserstein GAN



Earth Movers Distance

$$W(P,Q) = \inf_{U \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim U}[\|x-y\|]$$



Kantorovich-Rubinstein Duality

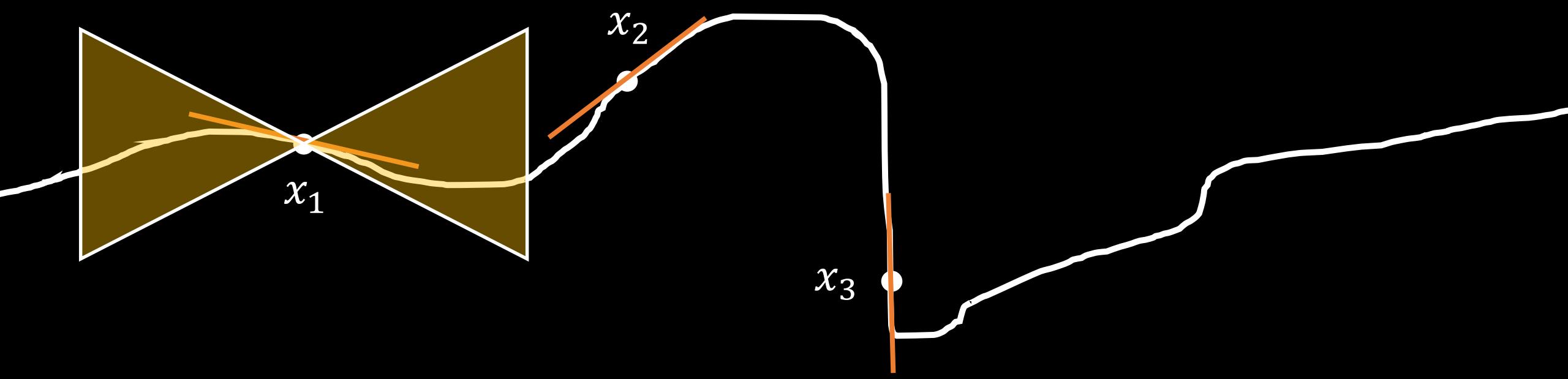
$$W(P, Q) = \inf_{U \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim U}[\|x - y\|]$$

Is the same as

$$W(P, Q) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[f(x)]$$

Survey on optimal transport [Villani, 2009]

Lipschitz norm of a function



$$\|f\|_{\mathrm{Lip}} = \sup_{x,x'} \frac{|f(x) - f(x')|}{|x - x'|}$$

$$\|f\|_{\mathrm{Lip}} = \sup_{x,x'} \frac{\|f(x) - f(x')\|}{\|x - x'\|}$$

Wasserstein GAN

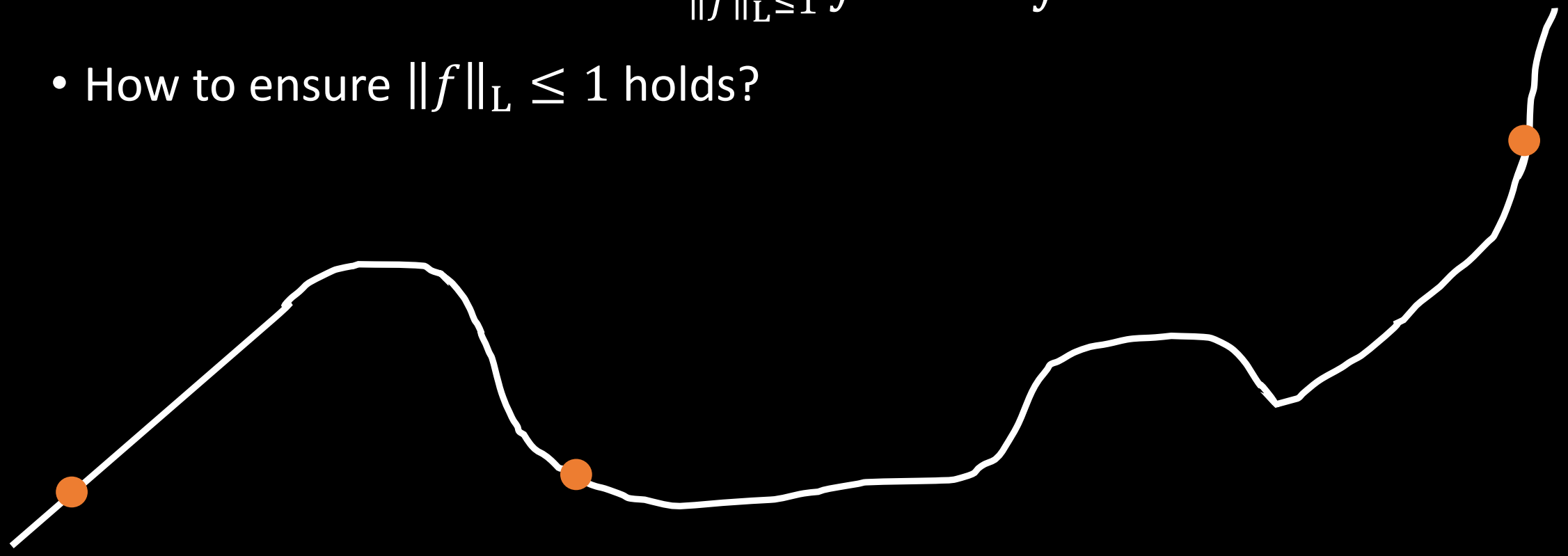
- [Arjovsky et al., arXiv:1701.07875], “WGAN”
- [Gulrajani et al., NIPS 2017], “WGAN-GP”

- Basic idea: approximate $W(P, Q)$

Wasserstein GAN

$$W(P,Q) = \sup_{\|f\|_{\mathrm{L}} \leq 1} \int f \mathrm{d}P - \int f \mathrm{d}Q$$

- How to ensure $\|f\|_{\mathrm{L}} \leq 1$ holds?



WGAN Solution [Arjosvky et al., 2017]

- Solution: weight clipping
- Justification: guarantees K-Lipschitz bounded functions
- Good results on DCGAN architectures
- *"Weight clipping is a clearly terrible way to enforce a Lipschitz constraint"*
- Problem 1 (practical): choice of clipping value (e.g. [-0.01, 0.01])
- Problem 2: leads to non-uniform bounding of gradients

WGAN-GP Solution [Guljarani et al., NIPS 2017]

- Approximate Lipschitz condition with a soft-penalty

$$\widetilde{W}(P,Q) = \boxed{\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[f(x)]} + \lambda \mathbb{E}_{x\sim M(P,Q)}[(\|\nabla_x f(x)\|_2 - 1)^2]$$

- IPM objective
- Choice of points to enforce gradient penalty at $\alpha \sim U(0,1)$, and $M(P,Q) = \alpha P + (1-\alpha)Q$
- Quadratic penalty function

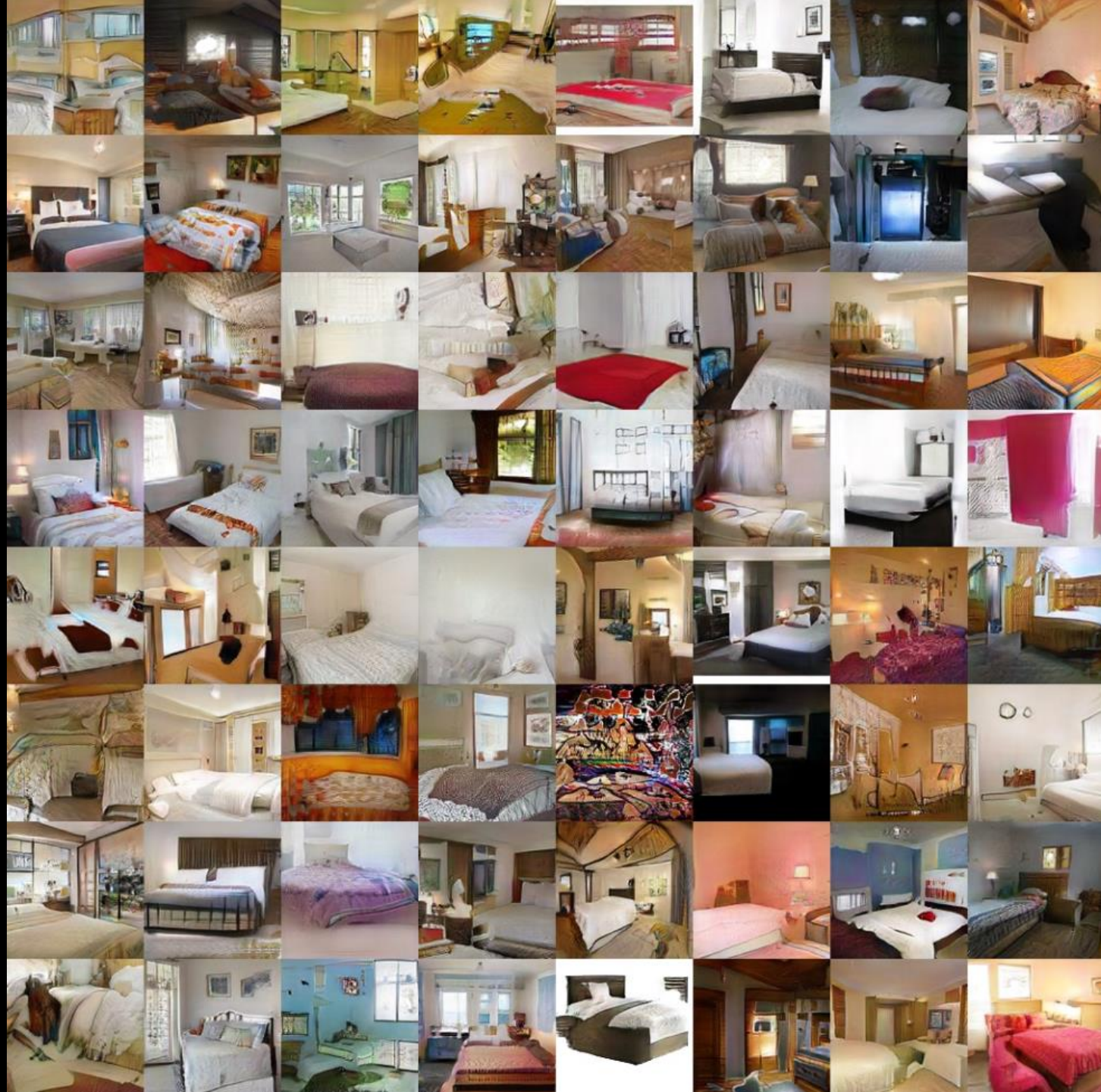
Algorithm 1 WGAN with gradient penalty. We use default values of $\lambda = 10$, $n_{\text{critic}} = 5$, $\alpha = 0.0001$, $\beta_1 = 0$, $\beta_2 = 0.9$.

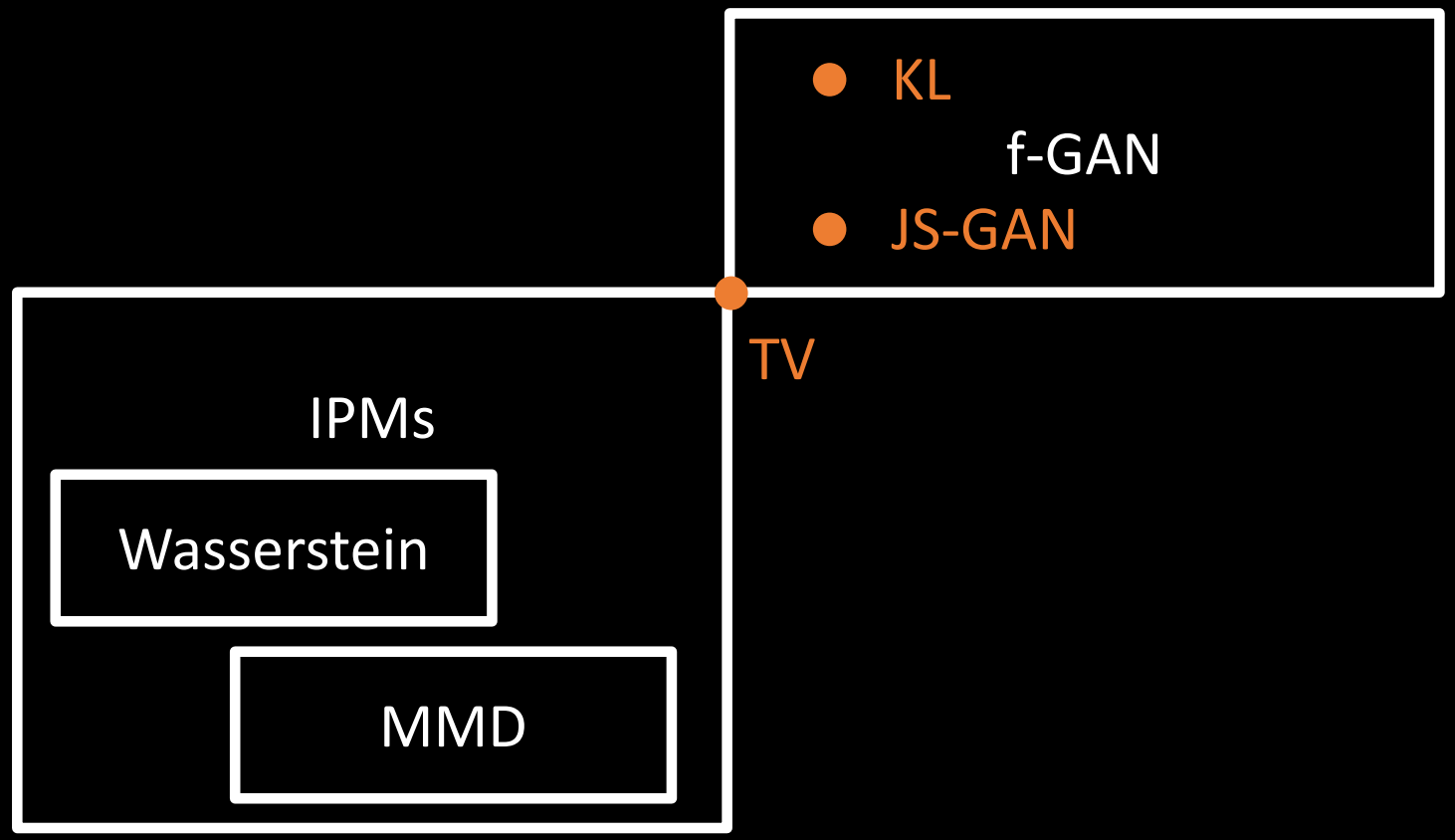
Require: The gradient penalty coefficient λ, the number of critic iterations per generator iteration n_{critic}, the batch size m, Adam hyperparameters α, β_1, β_2.
Require: initial critic parameters w_0, initial generator parameters θ_0.

```
while θ has not converged do
    for t = 1, ..., n_critic do
        for i = 1, ..., m do
            Sample real data x ~ P_r, latent variable z ~ p(z), a random number ε ~ U[0, 1].
            x̃ ← G_θ(z)
            x̂ ← εx + (1 − ε)x̃
            L^(i) ← D_w(x̃) − D_w(x) + λ(‖∇_x̂ D_w(x̂)‖_2 − 1)^2
        end for
        w ← Adam(∇_w (1/m) Σ_{i=1}^m L^(i), w, α, β_1, β_2)
    end for
    Sample a batch of latent variables {z^(i)}_{i=1}^m ~ p(z).
    θ ← Adam(∇_θ (1/m) Σ_{i=1}^m −D_w(G_θ(z)), θ, α, β_1, β_2)
end while
```

ResNet WGAN-GP

- [Guljarani et al., 2017]
- ResNet-101
- 128x128
- Gradient penalty, $\lambda = 10$

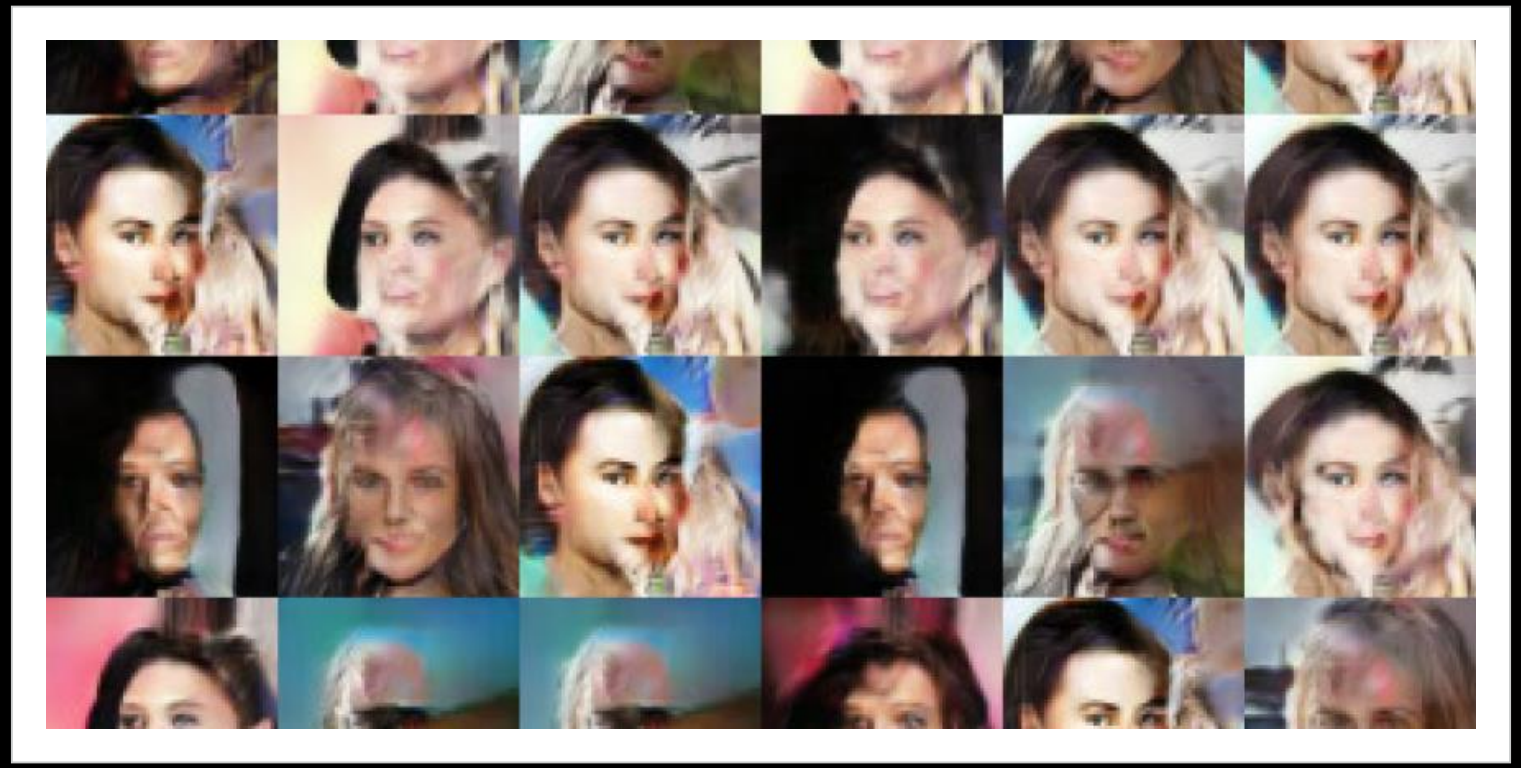



KL
f-GAN
JS-GAN
TV
IPMs
Wasserstein
MMD

Generative Adversarial Networks

Part 2

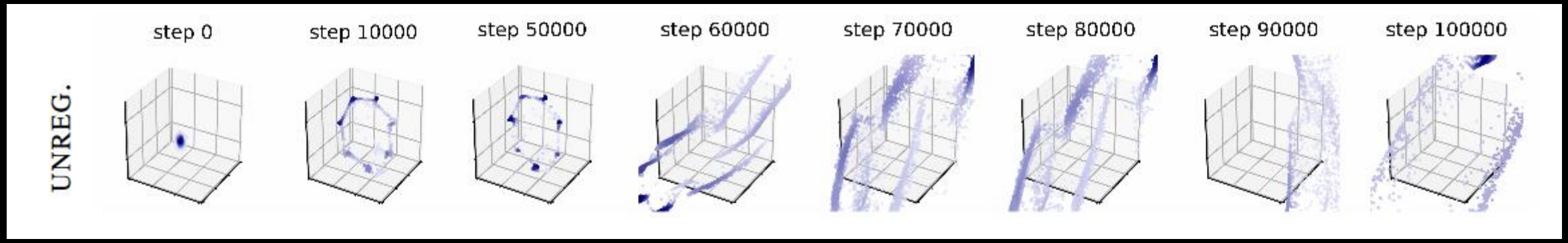
Sebastian Nowozin
Microsoft Research Cambridge
MLSS 2018 – Madrid – 7th September 2018



Mode Collapse

Empirically observed behavior where model produces only few distinct samples

Mode Collapse and Stability

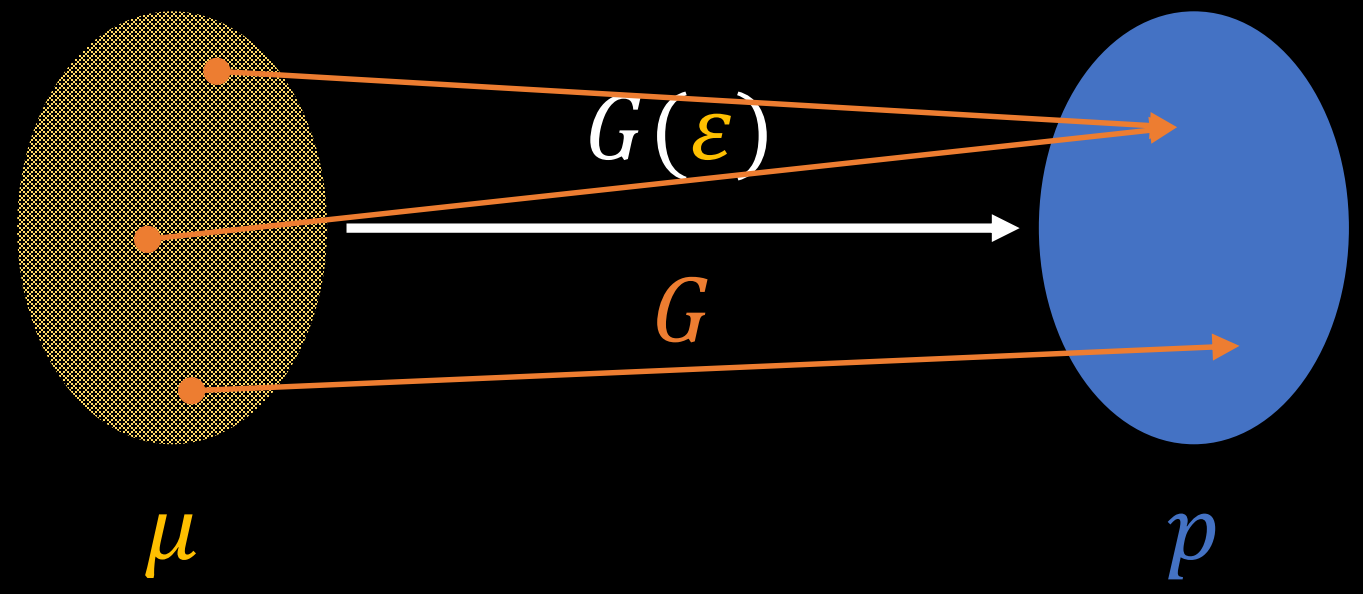


Unstable Training Behaviour example [Roth et al., 2017]

Understanding and Resolving GAN Mode Collapse and Instability

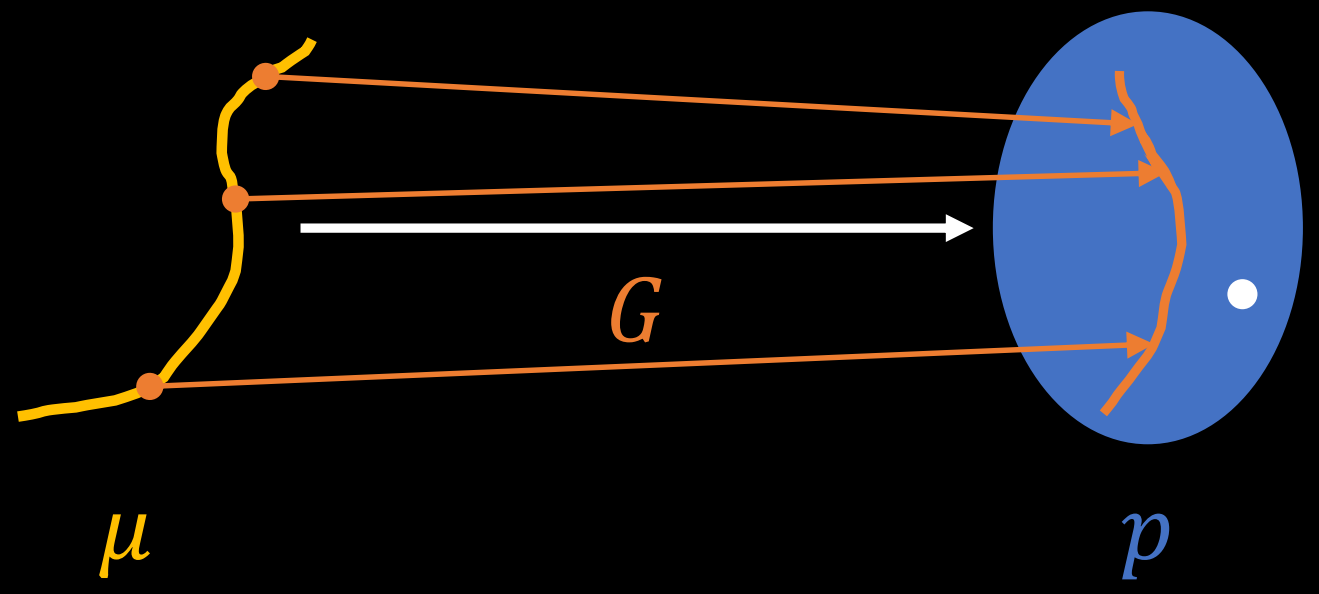
	Divergence Viewpoint	Algorithmic Viewpoint
Mode Collapse and Training Instability	[Arjovsky and Bottou, 2016], [Sønderby et al., 2016], [Roth et al., 2017], [Mescheder et al., 2018]	[Mescheder et al., 2017] [Mescheder et al., 2018]

Divergence View



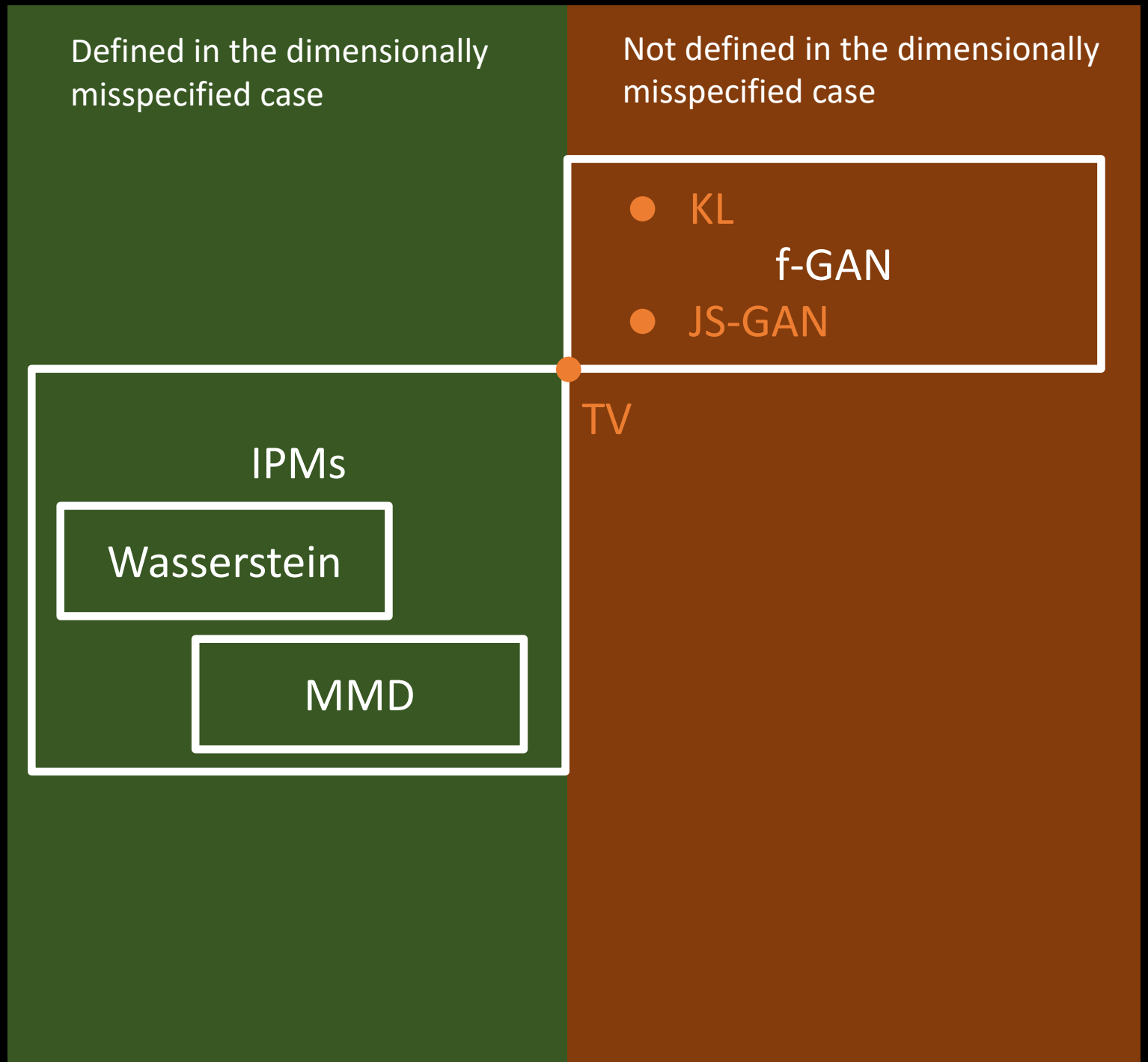
$$p(x) = \int_{z \in G^{-1}(x)} \mu(z)\, \mathrm{d}z$$

Divergence View (cont)



$p(x)$ not defined a.e.

Hence, f-divergence not defined!

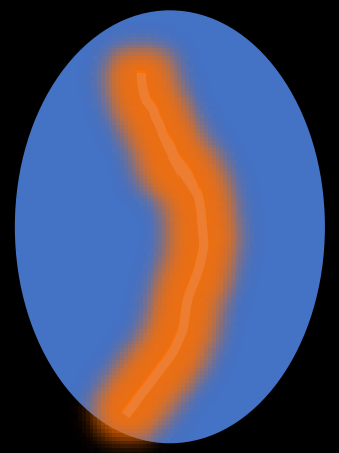

Defined in the dimensionally misspecified case
Not defined in the dimensionally misspecified case
KL
f-GAN
JS-GAN
TV
IPMs
Wasserstein
MMD

Fixing the Problem: Generalized f-Divergence

- Use generalized f-divergence

$$D_{f,K}(P,Q) = D_f(K * P, K * Q)$$

- Implementation: instance noise
  [Sønderby et al., 2016]
  [Arjovsky and Bottou, 2016]
- Implementation: analytic
  [Roth et al., 2017]
- Choice of kernel introduces local geometry



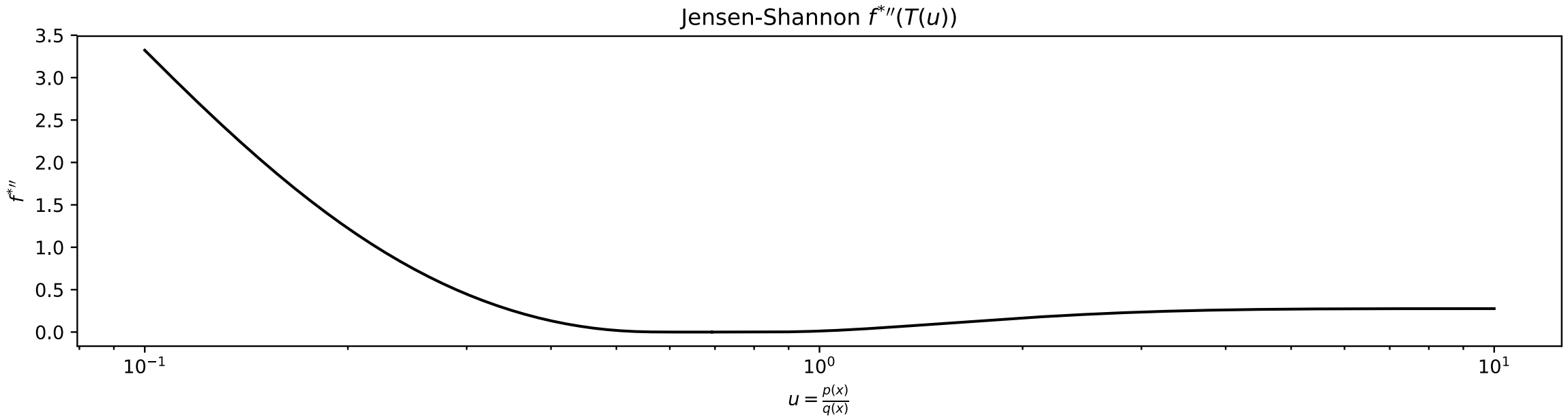
p

$p(x)$ not defined a.e.

Regularization [Roth et al., NIPS 2017]

- $\min_{\theta} \max_{\omega} \left(\mathbb{E}_{x \sim P_\theta}[T_\omega(x)] - \mathbb{E}_{x \sim Q}[f^*(T_\omega(x))] \right)$

$$- \frac{\gamma}{2} \mathbb{E}_{x \sim Q}[f^{*\prime\prime}(T_\omega(x)) \; \|\nabla T_\omega(x)\|^2]$$



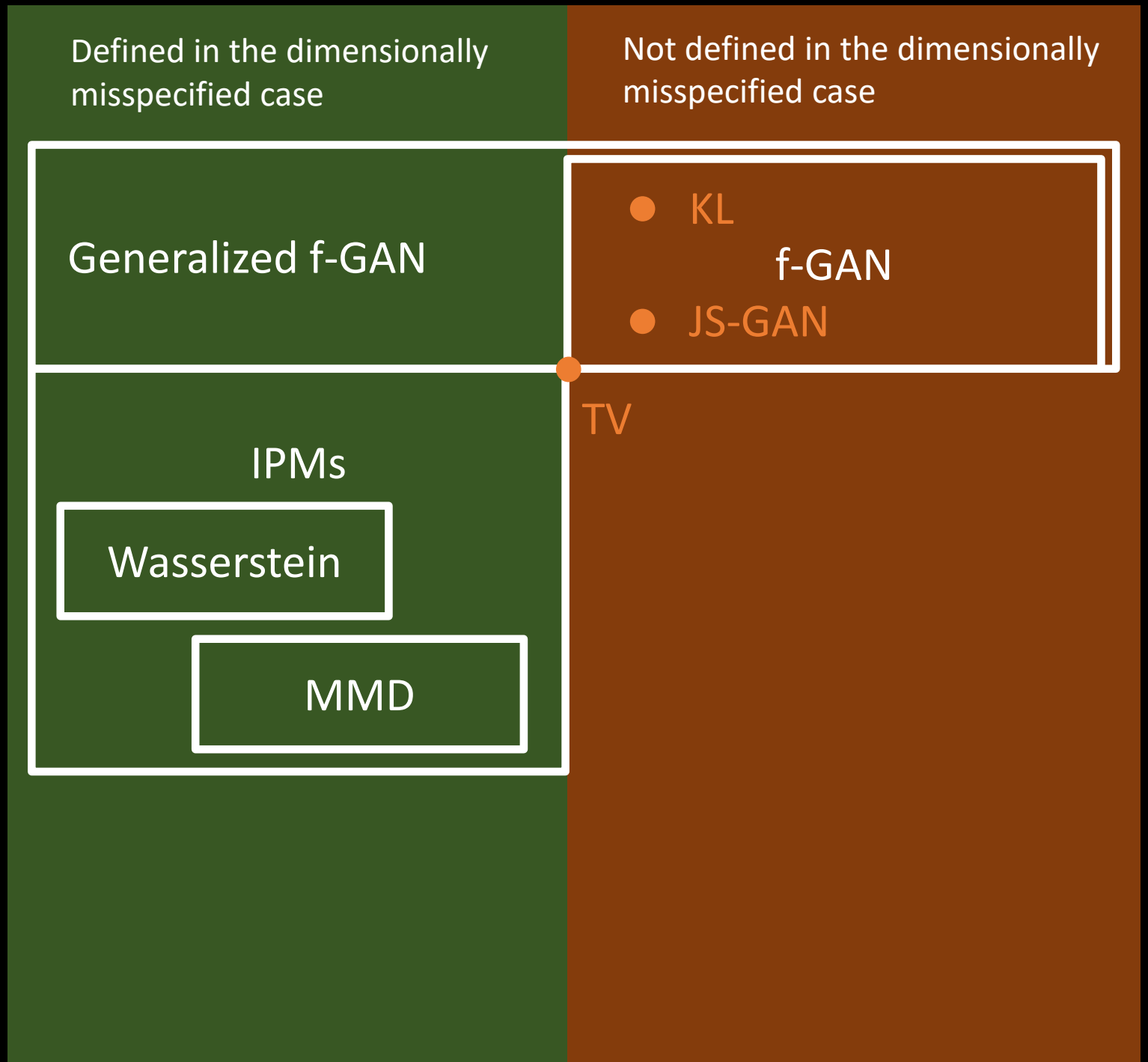
Regularization: Jensen-Shannon Case

- Classification form, Jensen-Shannon special case $D_\omega(x) = \sigma(T_\omega(x))$

$$\min_{\theta} \max_{\omega} \mathbb{E}_{x\sim P_\theta}[\log D_\omega(x)] + \mathbb{E}_{x\sim Q}[\log(1 - D_\omega(x))] - \frac{\gamma}{2}\Omega_{\mathrm{JS}}$$

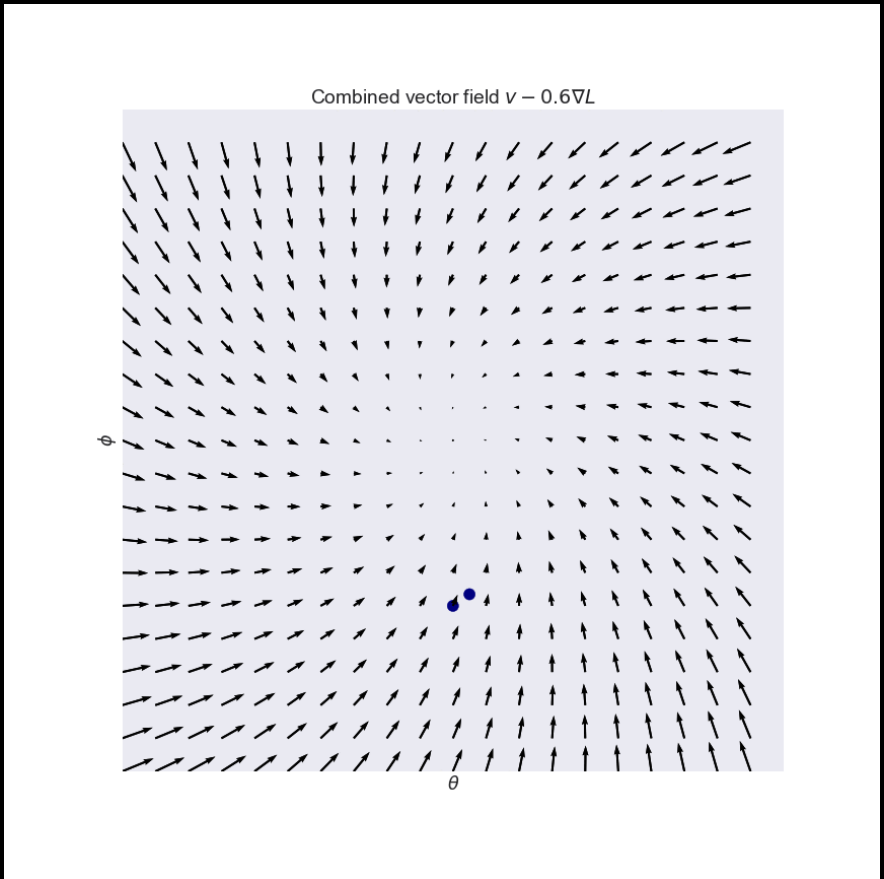
$$\Omega_{\mathrm{JS}} = \mathbb{E}_{x\sim P_\theta}[(1 - D_\omega(x))^2 \|\nabla T_\omega(x)\|^2] + \mathbb{E}_{x\sim Q}[D_\omega(x)^2 \|\nabla T_\omega(x)\|^2]$$

- “When uncertain, keep logits smooth”


Defined in the dimensionally misspecified case
Not defined in the dimensionally misspecified case
Generalized f-GAN
KL
f-GAN
JS-GAN
TV
IPMs
Wasserstein
MMD

Algorithmic View [Mescheder et al., 2017, 2018]

(with figures from Ferenc Huszar)

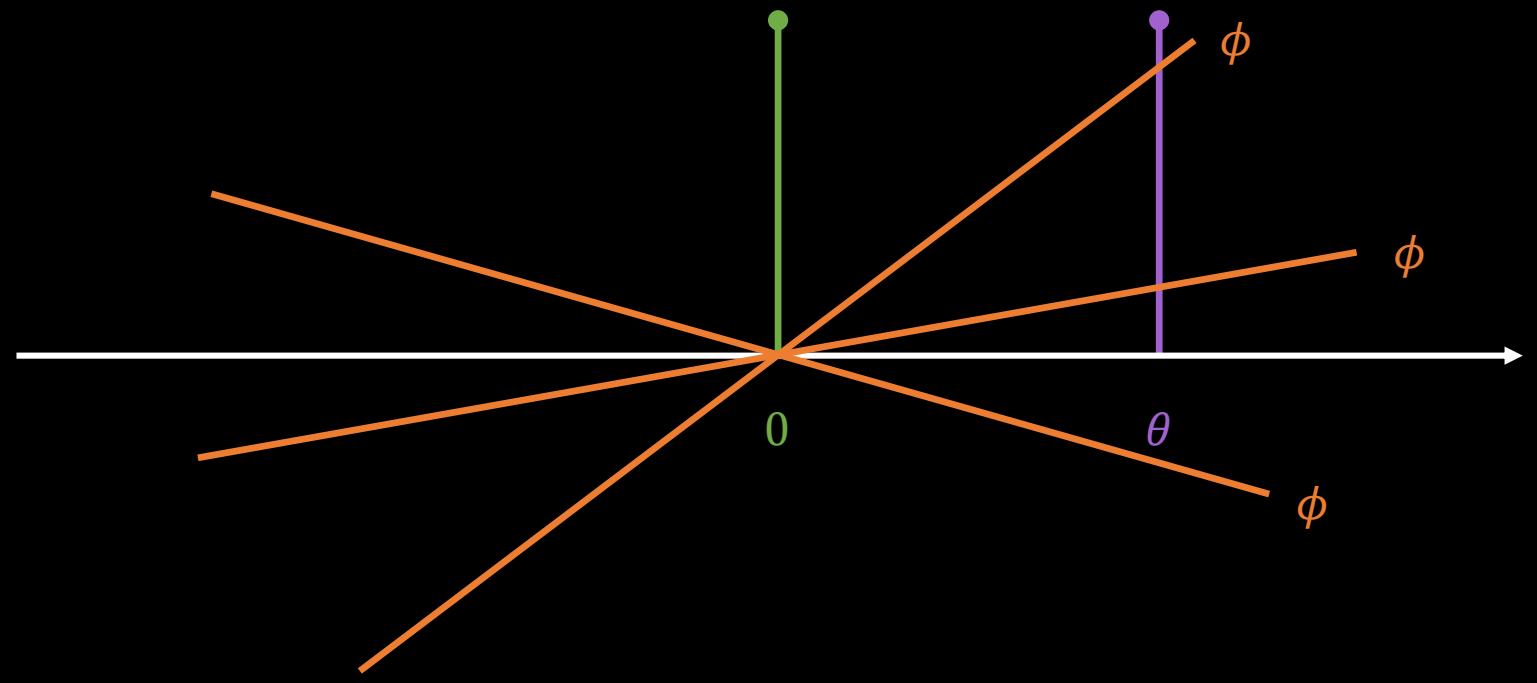


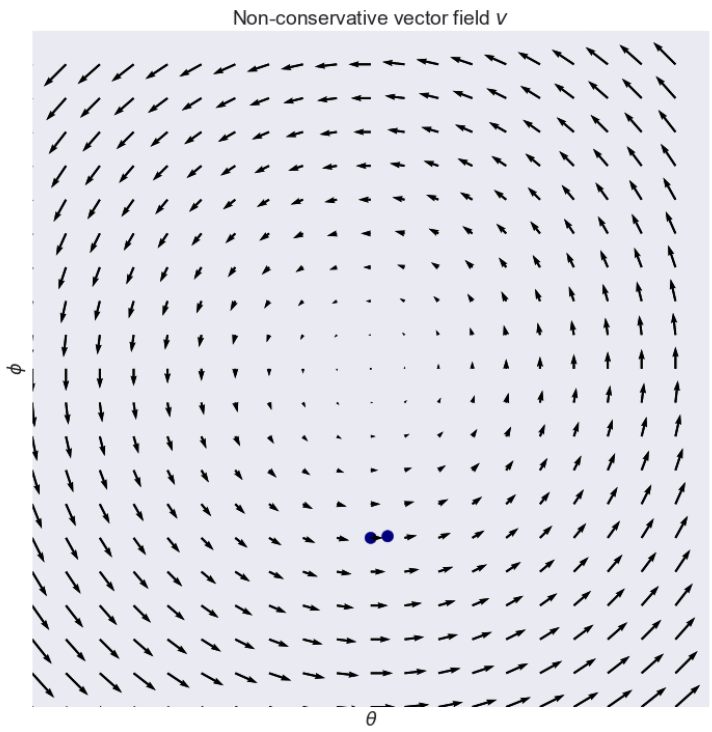
- Solving for saddle points is different from gradient descent

- Minimization problems yield *conservative* vector fields
- Min-max saddle point problems may yield *non-conservative* vector fields

- Our solution:
  consensus optimization ==
  adding gradient vector field

Dirac GAN


ϕ
ϕ
0
θ
ϕ


Non-conservative vector field v
ϕ
θ

Algorithmic View (cont'd)

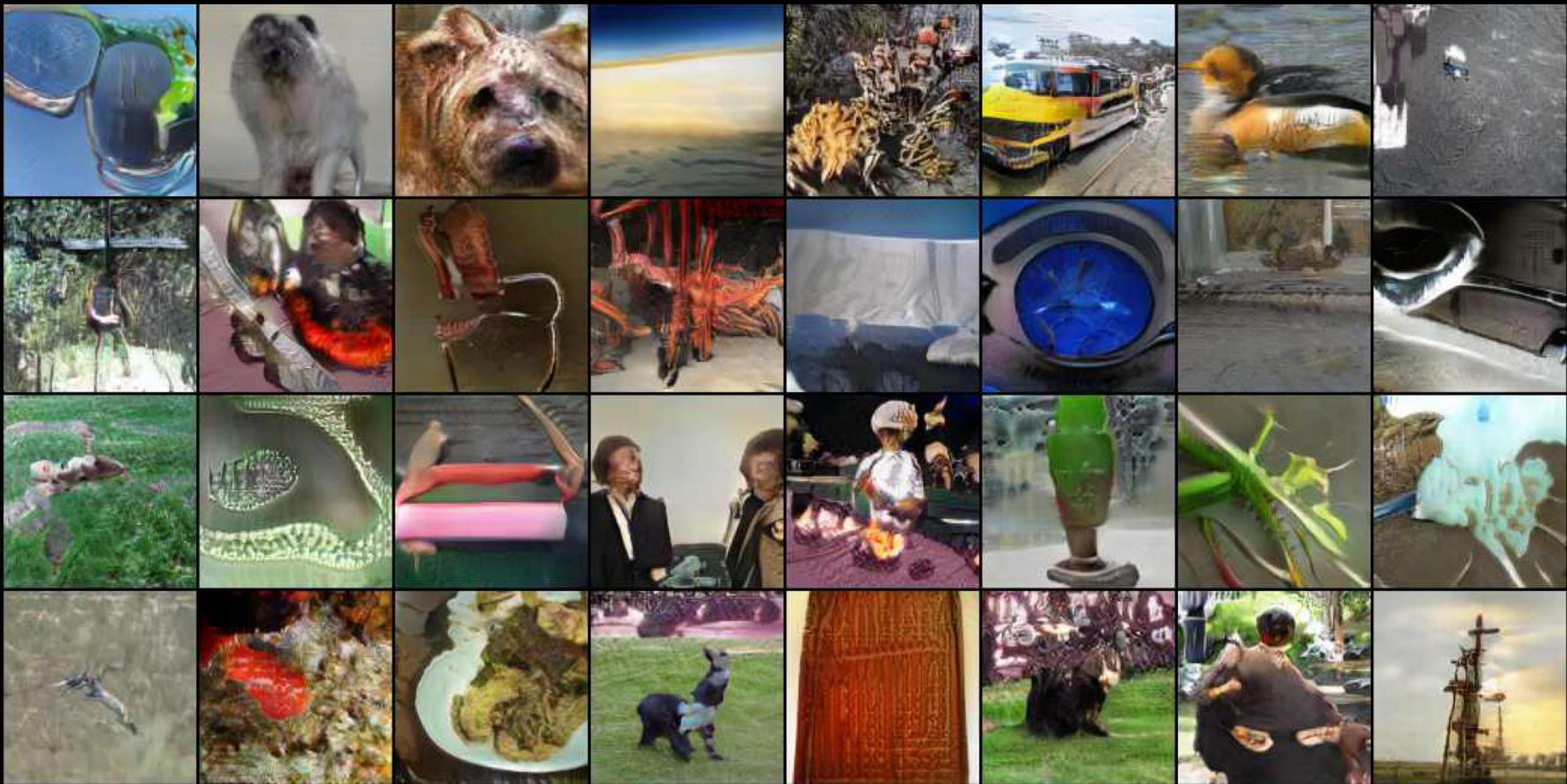
- Theory of GAN convergence, unification with divergence view [Mescheder et al., 2018]

Method	Local convergence (a.c. case)	Local convergence (general case)
unregularized (Goodfellow et al., 2014)	✓	✗
WGAN (Arjovsky et al., 2017)	✗	✗
WGAN-GP (Gulrajani et al., 2017)	✗	✗
DRAGAN (Kodali et al., 2017)	✓	✗
Instance noise (Sønderby et al., 2016)	✓	✓
ConOpt (Mescheder et al., 2017)	✓	✓
Gradient penalties (Roth et al., 2017)	✓	✓
Gradient penalty on real data only	✓	✓
Gradient penalty on fake data only	✓	✓

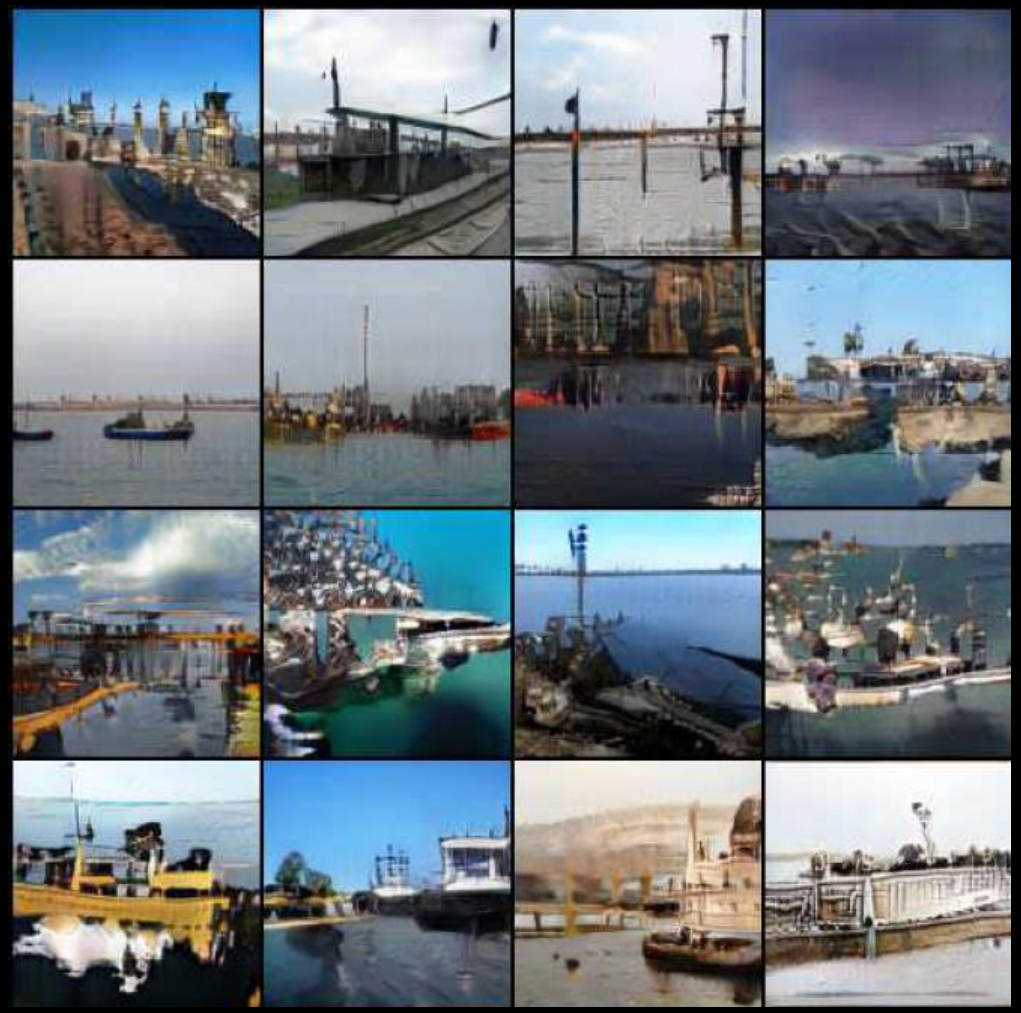
Simple Gradient Penalties

- $$\min_{\theta} \max_{\omega} \left(\mathbb{E}_{x \sim P_\theta}[T_\omega(x)] - \mathbb{E}_{x \sim Q}[f^*(T_\omega(x))] \right) - \frac{\gamma}{2} \mathbb{E}_{x \sim Q}[\, \|\nabla T_\omega(x)\|^2] - \frac{\gamma}{2} \mathbb{E}_{x \sim P_\theta}[\, \|\nabla T_\omega(x)\|^2]$$

ImageNet 1000 classes, 128x128, unconditional generation, ResNet 55 [Mescheder et al., ICML 2018], arXiv:1801.04406



ImageNet conditional generation [Mescheder et al., ICML 2018], arXiv:1801.04406

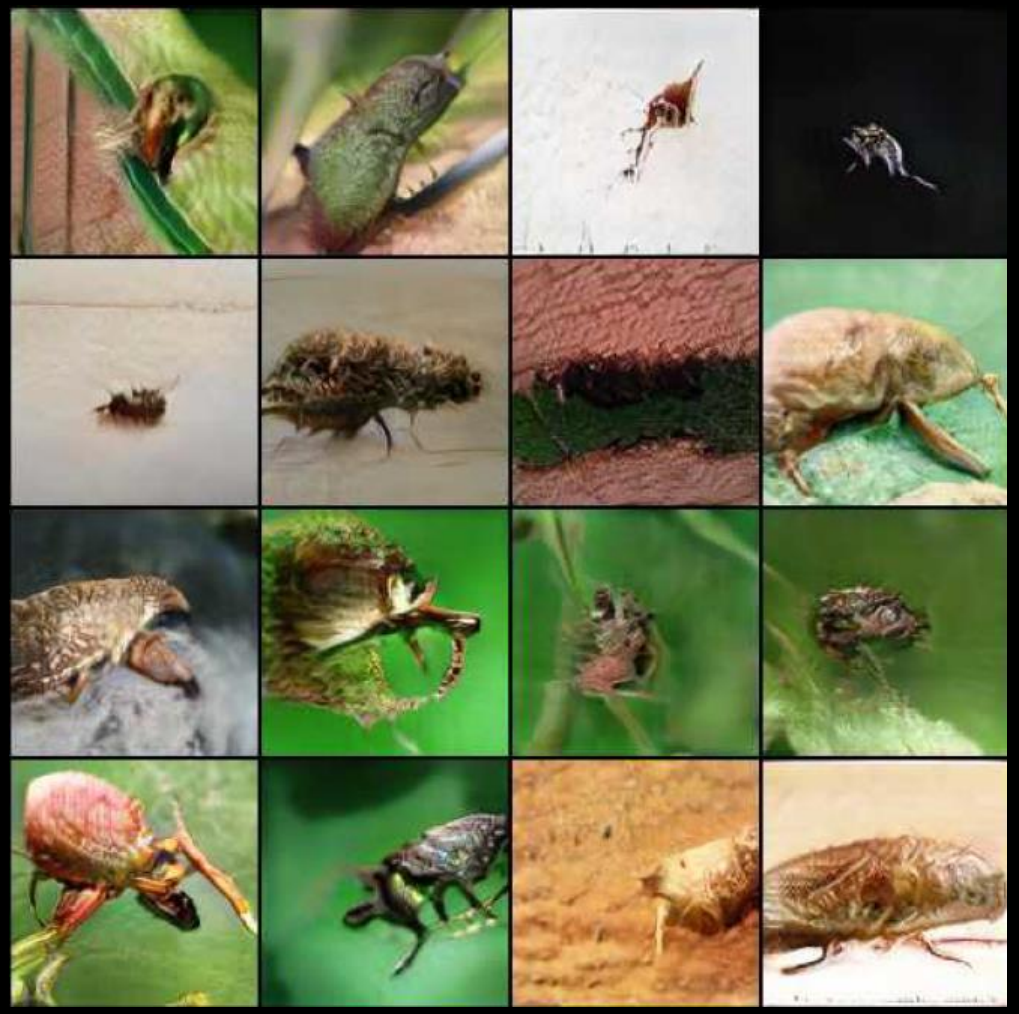


dock

pizza

ImageNet conditional generation [Mescheder et al., ICML 2018], arXiv:1801.04406

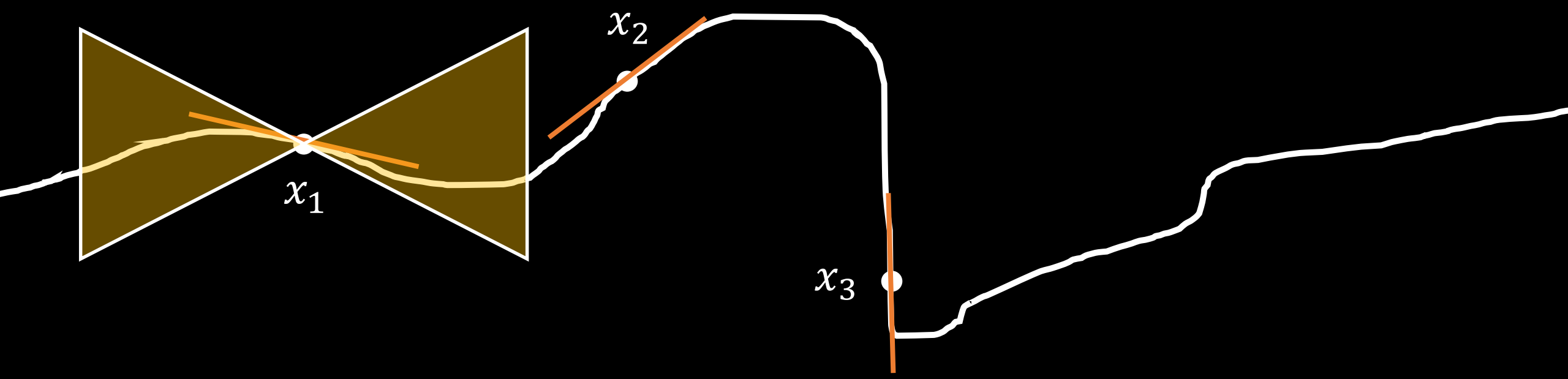
papillon



weevil

Spectral Normalization

Lipschitz norm of a function



$$\|f\|_{\mathrm{Lip}} = \sup_{x,x'} \frac{|f(x) - f(x')|}{|x - x'|}$$

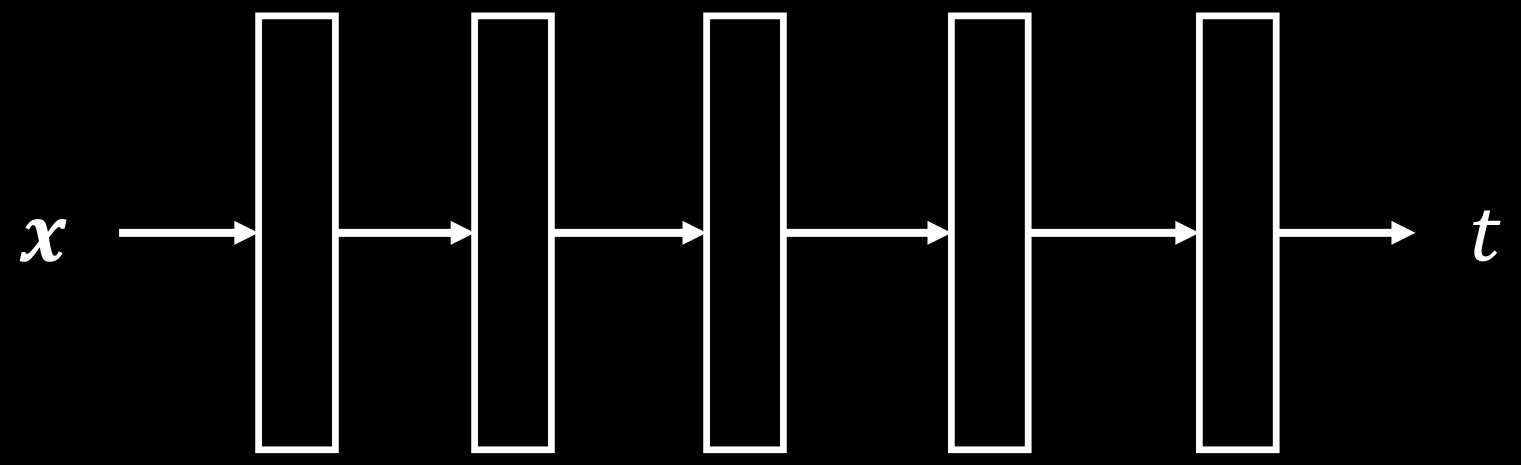
$$\|f\|_{\mathrm{Lip}} = \sup_{x,x'} \frac{\|f(x) - f(x')\|}{\|x - x'\|}$$

Spectral Normalization

- [Miyato et al., ICLR 2018]
- Idea:
  1. Limit discriminator function class to functions with bounded Lipschitz norm
  2. Bound global Lipschitz norm by product of Lipschitz norm per layer
  3. Compute Lipschitz norm per layer efficiently using power method

Spectral Normalization

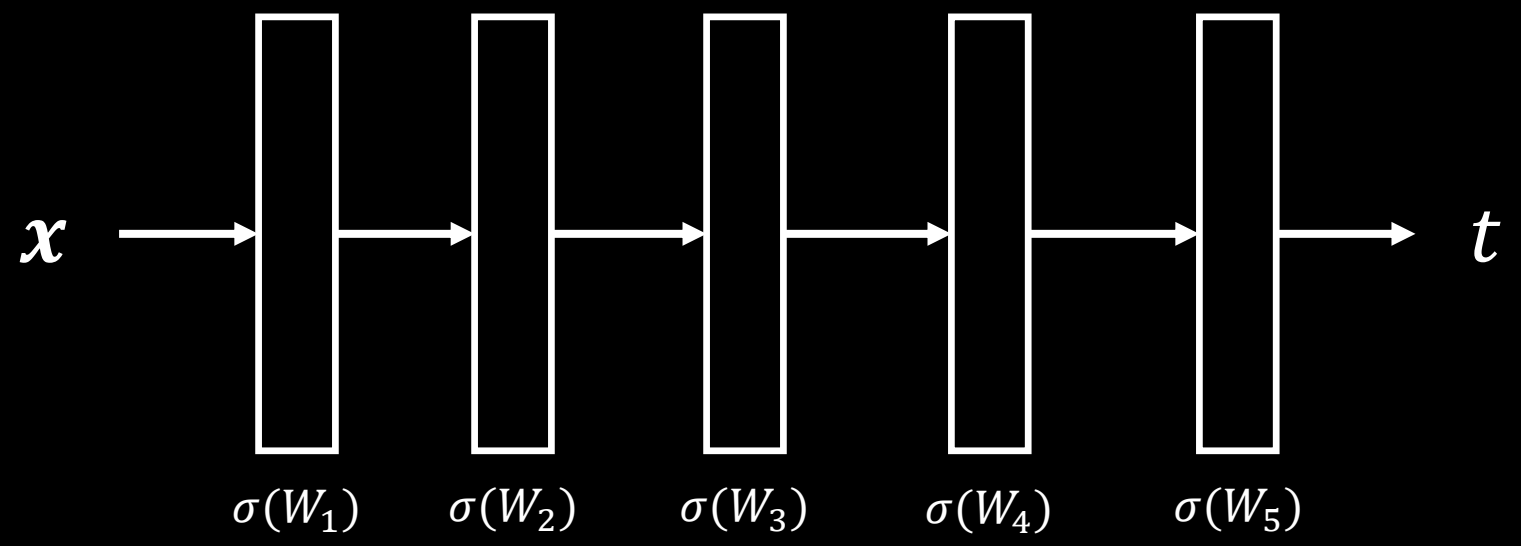
Limit discriminator function class to functions with bounded Lipschitz norm



$$\|D\|_{\mathrm{Lip}} = \sup_{x,x'} \frac{\|D(x) - D(x')\|}{\|x - x'\|}$$

Spectral Normalization

Bound global Lipschitz norm by product of Lipschitz norm per layer



$$\|D\|_{\mathrm{Lip}} \leq \prod_{i=1}^{L} \sigma(W_i) \qquad \sigma(W_i) = \max_{\|\boldsymbol{h}\|_2 \leq 1} \|W_i \boldsymbol{h}\|_2$$

Spectral Normalization

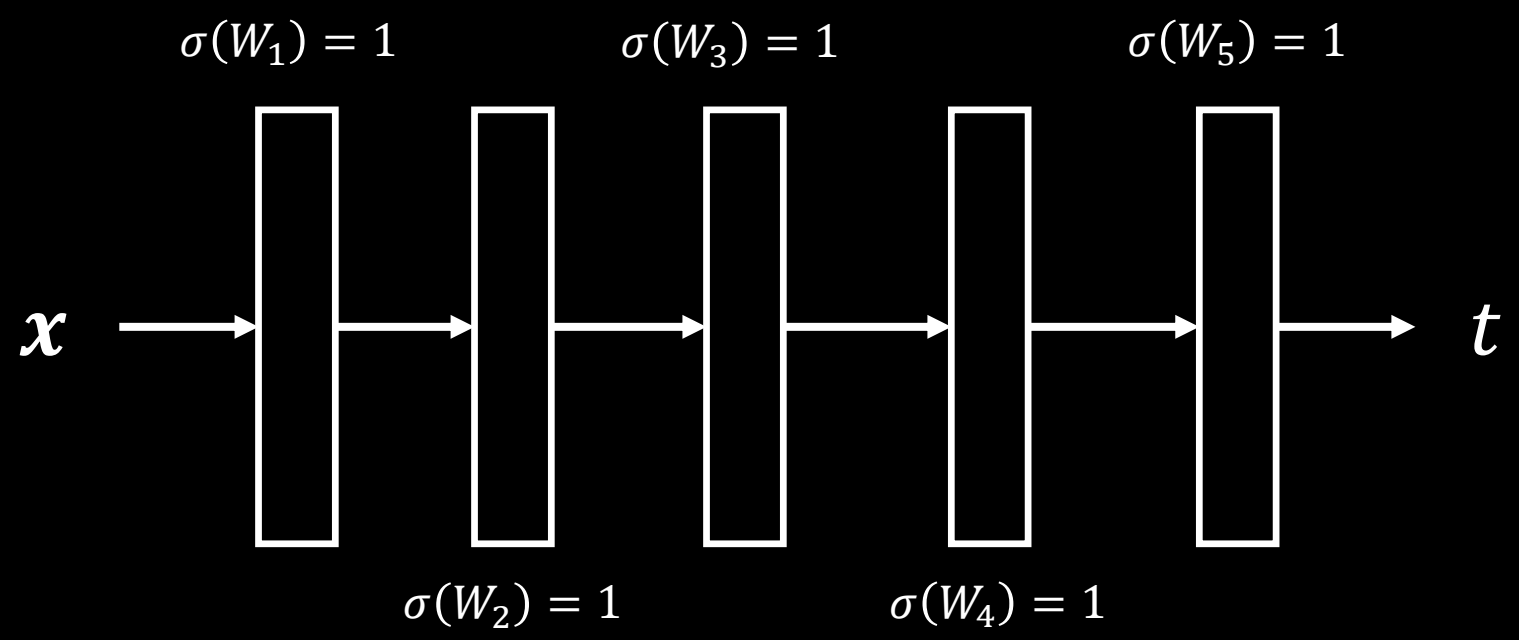
Compute Lipschitz norm per layer efficiently using power method

$$\sigma(W_i) = \max_{\|\boldsymbol{h}\|_2 \leq 1} \|W_i \boldsymbol{h}\|_2$$

1. Exact computations requires SVD of matrix and backprop through SVD (easily possible in modern frameworks, but expensive)
2. Miyato et al. apply one iteration of an iterative method (power method), amortizing the cost of finding the largest singular value
3. Then normalize each weight matrix to set $\sigma(W_i) = 1$

Spectral Normalization

Application to GANs: force spectral norm to be one for each layer



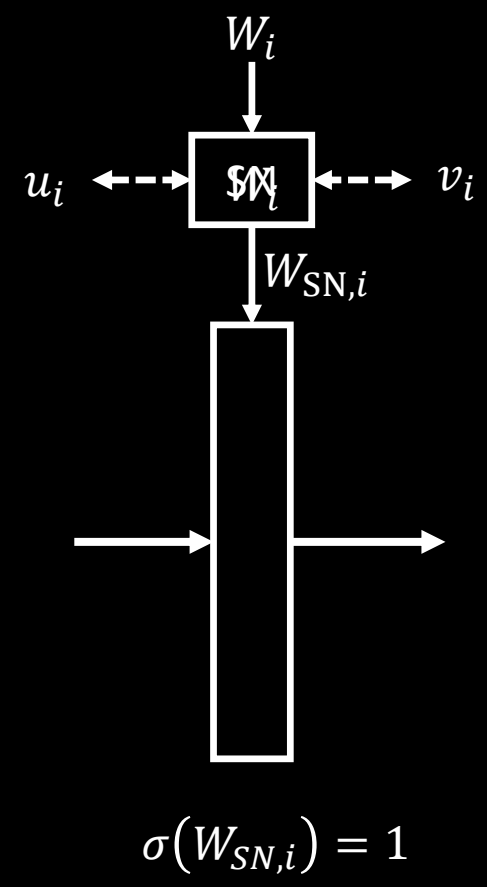
Spectral Normalization: Power Method

- How to enforce the spectral normalization constraint?

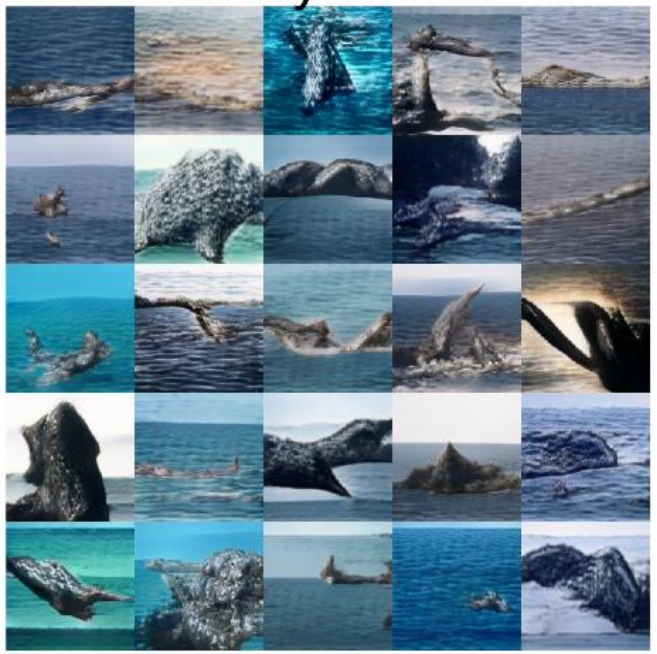
- Spectral normalization layer:

$$1.\ v_i \leftarrow W_i{}^T u_i / \|W_i{}^T u_i\|_2$$
$$2.\ u_i \leftarrow W_i v_i / \|W_i v_i\|_2$$
$$3.\ W_{\mathrm{SN},i} = W_i / (u_i{}^T W_i v_i)$$

- $(u_i{}^T W_i v_i)$ estimates the spectral norm
- Clever: *amortization* of SVD computation
- No backprop to u_i or v_i



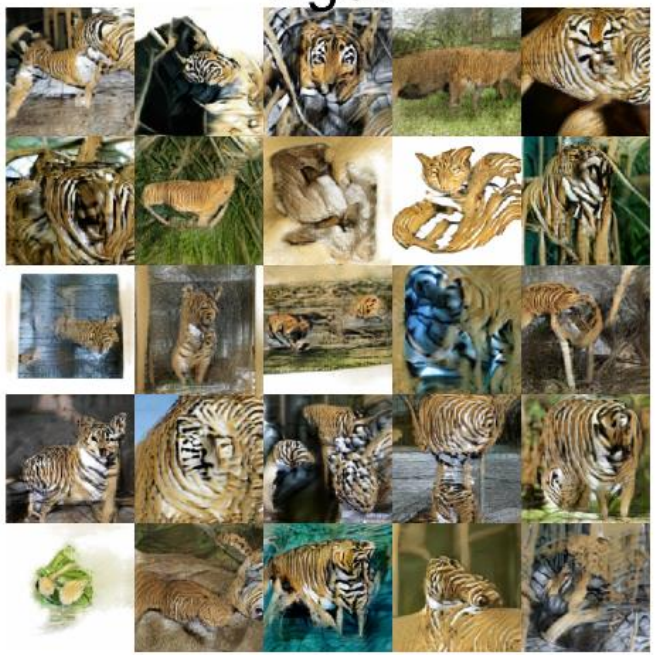
Gray whale



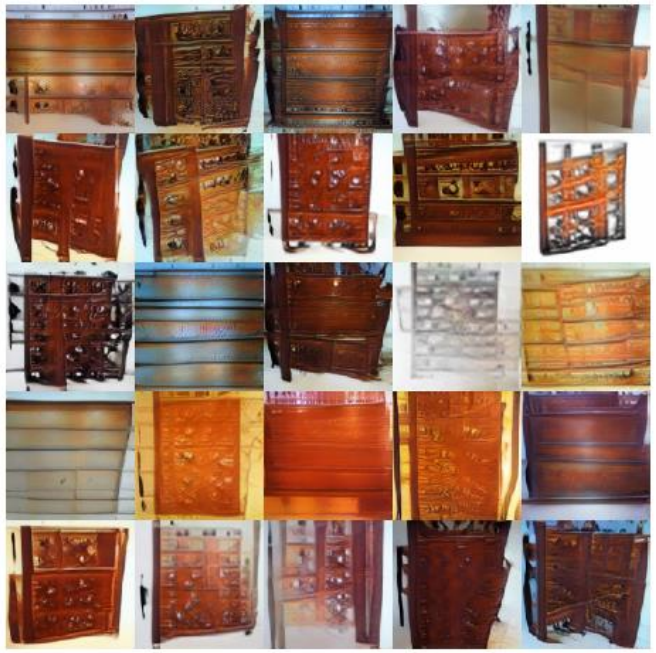
Welsh springer spaniel

Persian cat

Tiger



Chiffonier

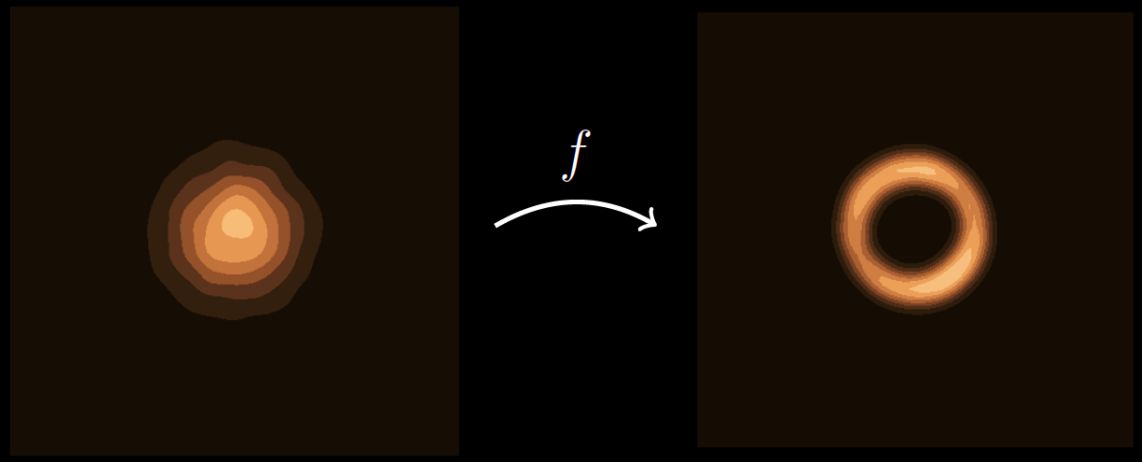


Fire truck

Implicit Models more Generally

Implicit models as building blocks

- Adversarial Autoencoders [Makhzani et al,. 2015]
- Hierarchical implicit models [Tran et al., NIPS 2017]
- Adversarial variational Bayes [Mescheder et al., ICML 2017]



Adversarial Variational Bayes:

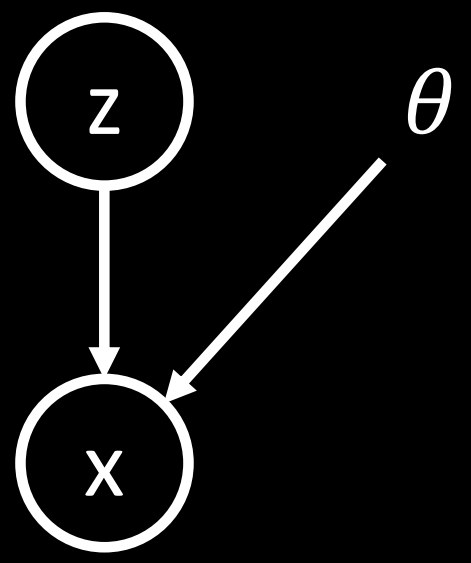
Unifying Variational Autoencoders and Generative Adversarial Networks

Lars Mescheder, Sebastian Nowozin, Andreas Geiger

ICML 2017

arXiv:1701.04722

VAE: Model



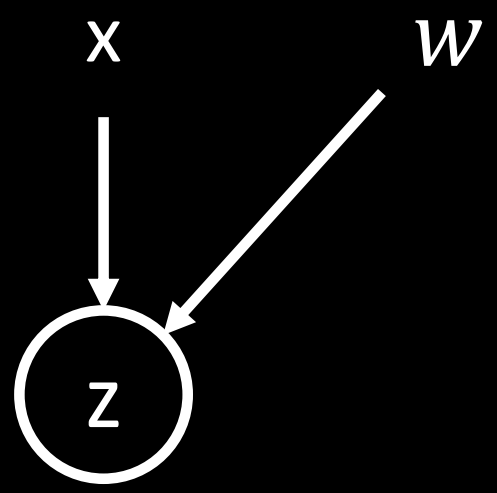
$$p(x|\theta) = \int \mathrm{p(x|z,\theta)}p(z)\mathrm{dz}$$

- $p(z)$ is a multivariate standard Normal
- $p(x|z,\theta)$ is a neural network outputting a simple distribution (e.g. diagonal Normal)

Inference networks

x_1 x_2 x_3

$q_1(z)$ $q_2(z)$ $q_3(z)$



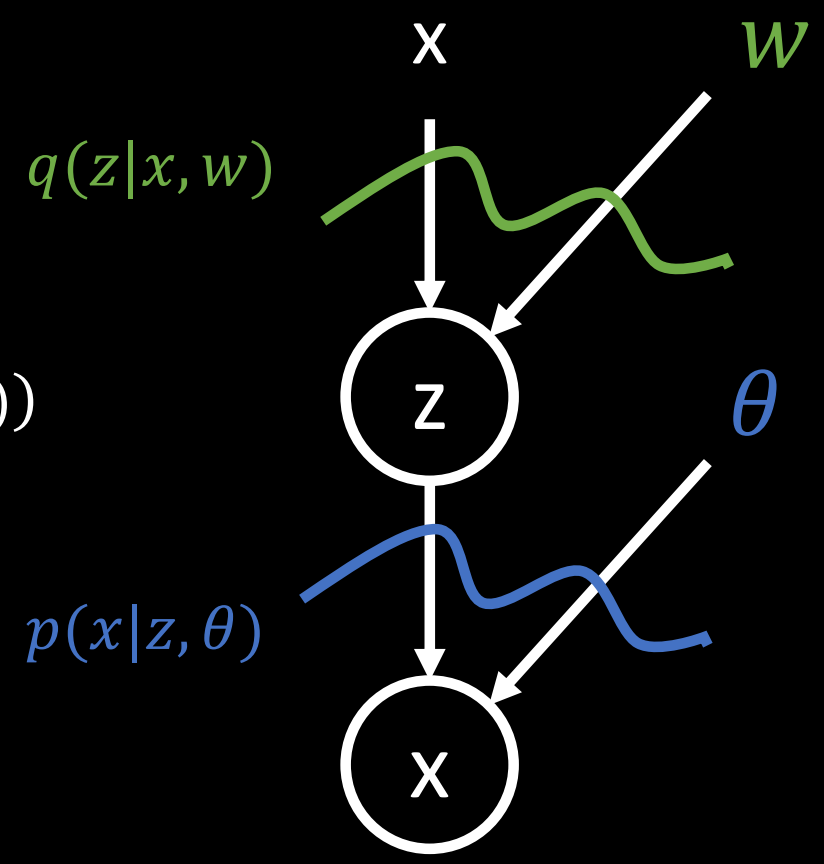
VAE: Maximum Likelihood Training

- Maximize the data log-likelihood, inference network variational approximation

$$
\begin{aligned}
\log p(x|\theta) &= \log \int \mathrm{p}(\mathrm{x}|\mathrm{z},\theta)p(z)\mathrm{d}z \\
&= \log \int \mathrm{p}(\mathrm{x}|\mathrm{z},\theta)\frac{q(z|x,w)}{q(z|x,w)}p(z)\,\mathrm{d}z \\
&= \log \int \mathrm{p}(\mathrm{x}|\mathrm{z},\theta)\frac{p(z)}{q(z|x,w)}q(z|x,w)\,\mathrm{d}z \\
&= \log \mathbb{E}_{z\sim q(z|x,w)}\left[\mathrm{p}(\mathrm{x}|\mathrm{z},\theta)\frac{p(z)}{q(z|x,w)}\right] \\
&\geq \mathbb{E}_{z\sim q(z|x,w)}\left[\log \mathrm{p}(\mathrm{x}|\mathrm{z},\theta)\frac{p(z)}{q(z|x,w)}\right] \\
&= \mathbb{E}_{z\sim q(z|x,w)}[\log \mathrm{p}(\mathrm{x}|\mathrm{z},\theta)] - D_{\mathrm{KL}}(q(z|x,w)\parallel p(z))
\end{aligned}
$$

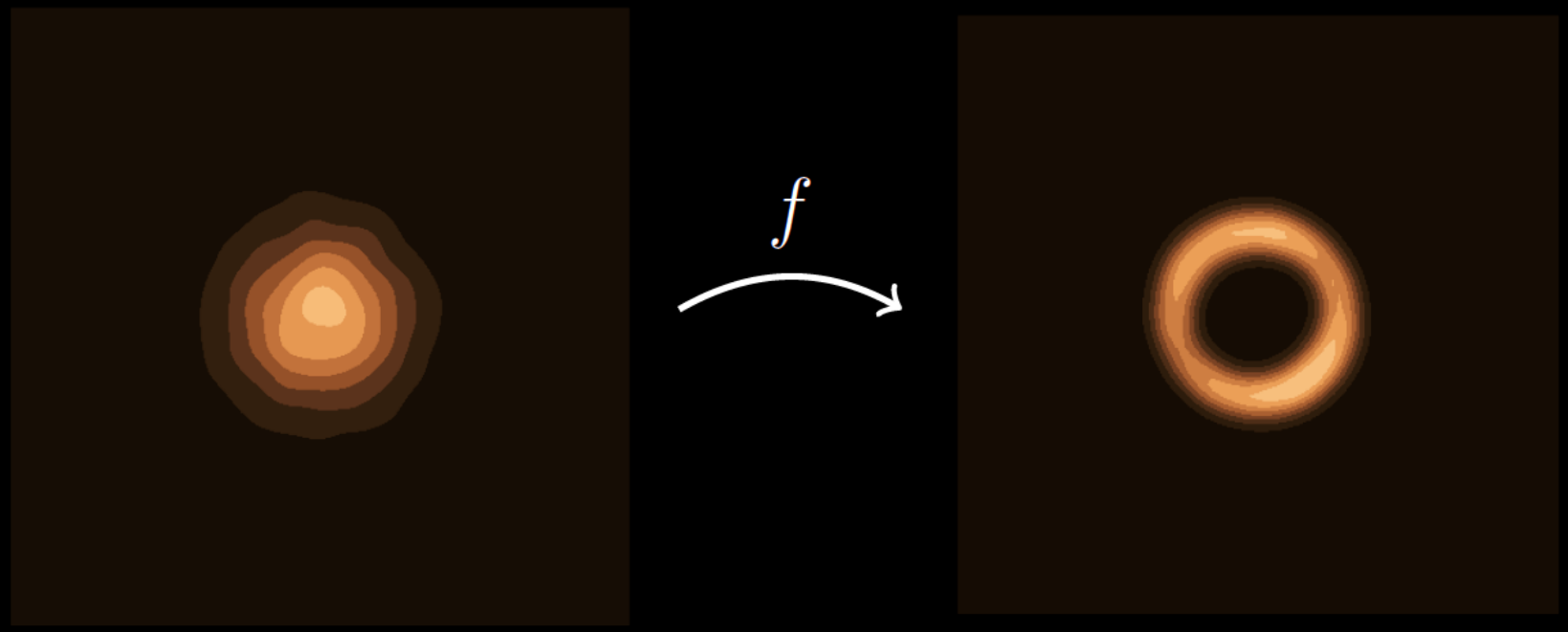
Autoencoder viewpoint

$$\max_{w,\theta} \; \mathbb{E}_{z\sim q(z|x,w)}[\log \mathrm{p}(\mathrm{x}|\mathrm{z},\theta)] - D_{\mathrm{KL}}(q(z|x,w) \parallel p(z))$$



High-level idea: 1/2

- What do we currently require from $q(z|x,w)$?
  - Sampling: $z \sim q(z|x,w)$
  - ~~Log-density computation: $\log q(z|x,w)$~~



High-level idea: 2/2

$$\mathbb{E}_{z\sim q(z|x,w)}[\log \mathrm{p}(\mathrm{x}|\mathrm{z},\theta)] - D_{\mathrm{KL}}\big(q(z|x,w) \parallel p(z)\big)$$
$$= \mathbb{E}_{z\sim q(z|x,w)}[\log \mathrm{p}(\mathrm{x}|\mathrm{z},\theta) + \log p(z) - \log q(z|x,w)]$$

- Introduce a real-valued discriminator function $T(x,z)$ such that

$$T(x,z) \approx -\log p(z) + \log q(z|x,w)$$

Variational Approximation

$$\max_{T \in \mathcal{T}} \mathbb{E}_{x \sim p_D}\left[\mathbb{E}_{z \sim q(z|x,w)}[\log \sigma(T(x,z))] + \mathbb{E}_{z \sim p(z)}[\log(1 - \sigma(T(x,z)))]\right]$$

Proposition:

For $q(z|x,w)$ fixed the optimal discriminator T^* is given by

$$T^*(x,z) = -\log p(z) + \log q(z|x,w)\,.$$

Rewrite

$$\mathbb{E}_{z \sim q(z|x,w)}[\log \mathrm{p}(\mathrm{x}|\mathrm{z},\theta) + \log p(z) - \log q(z|x,w)]$$
$$= \mathbb{E}_{z \sim q(z|x,w)}\,[\log \mathrm{p}(\mathrm{x}|\mathrm{z},\theta) - T^*(x,z)\,]$$

Reparametrization Trick

- Learning objective

$$\mathbb{E}_{z\sim q(z|x,w)}[\log \mathrm{p}(\mathrm{x}|\mathrm{z},\theta) - T^*(x,z)\ ]$$

- Reparametrize [Kingma and Welling, 2013]

$$z\sim q(z|x,w) \Leftrightarrow \varepsilon\sim\mathcal{N}, z(x,w,\varepsilon)$$

- Reparametrized learning objective

$$\mathbb{E}_{\varepsilon}[\log \mathrm{p}(\mathrm{x}|\mathrm{z}(\mathrm{x},\mathrm{w},\varepsilon),\theta) - T^*(x,z(x,w,\varepsilon))\ ]$$

Variational Approximation (Discriminator)

$$\max_{T \in \mathcal{T}} \mathbb{E}_{x \sim p_D}\left[\mathbb{E}_{z \sim q(z|x,w)}[\log \sigma(T(x,z))] + \mathbb{E}_{z \sim p(z)}[\log(1 - \sigma(T(x,z)))]\right]$$

Variational adversary

Data distribution (fixed)

Variational distribution (fixed)

Adversarial Variational Bayes

$$\max_{\theta,w} \mathbb{E}_{\varepsilon}[\log \mathrm{p}(\mathrm{x}|\mathrm{z}(\mathrm{x},\mathrm{w},\varepsilon),\theta) - T(x, z(x, w, \varepsilon), \psi)\ ]$$

$$\max_{\psi} \mathbb{E}_{x\sim p_D}\left[\mathbb{E}_{z\sim q(z|x,w)}[\log \sigma(T(x, z, \psi))] + \mathbb{E}_{z\sim p(z)}[\log(1 - \sigma(T(x, z, \psi)))]\right]$$

- Parameter-free expectation form → unbiased estimation
- GAN-type algorithm

Algorithm 1 Adversarial Variational Bayes (AVB)

$i = 0$

while not converged **do**

Sample m examples $\{x^{(1)}, \ldots, x^{(m)}\}$ from data distribution $p_{\mathcal{D}}(x)$.

Sample m examples $\{z^{(1)}, \ldots, z^{(m)}\}$ from prior distribution $p(z)$.

Sample m noise examples $\{\epsilon^{(1)}, \ldots, \epsilon^{(m)}\}$ from $\mathcal{N}(0, 1)$.

Compute θ-gradient (eq. 3.9):

$$g_\theta = \nabla_\theta \frac{1}{m} \sum_{i=1}^{m} \log p_\theta \left(x^{(i)} \mid z_\phi \left(x^{(i)}, \epsilon^{(i)}\right)\right)$$

Compute ϕ-gradient (eq. 3.9):

$$g_\phi = \nabla_\phi \frac{1}{m} \sum_{i=1}^{m} \Big[-T_\psi \left(x^{(i)}, z_\phi(x^{(i)}, \epsilon^{(i)})\right) + \log p_\theta \left(x^{(i)} \mid z_\phi(x^{(i)}, \epsilon^{(i)})\right)\Big].$$

Compute ψ-gradient (eq. 3.3) :

$$g_\psi = \nabla_\psi \frac{1}{m} \sum_{i=1}^{m} \Big[\log \left(\sigma(T_\psi(x^{(i)}, z_\phi(x^{(i)}, \epsilon^{(i)}))\right) + \log \left(1 - \sigma(T_\psi(x^{(i)}, z^{(i)})\right)\Big].$$

Perform SGD-updates for θ, ϕ and ψ:

$$\theta = \theta + h_i \, g_\theta, \quad \phi = \phi + h_i \, g_\phi, \quad \psi = \psi - h_i \, g_\psi.$$

$i = i + 1$

end while

Binarized MNIST

- 28x28 binary images
- 50,000 training images
- 10,000 test images
- Train VAE model

- Report test ELBO
- Report Annealed importance sampling (AIS) estimates of test log-likelihood

Binarized MNIST density estimation

	ELBO	AIS	
AVB (8-dim)	$\approx -83.6 \pm 0.4$	-90.8 ± 1.0	
AVB + AC (8-dim)	$\approx -96.3 \pm 0.4$	-89.7 ± 1.0	
AVB + AC (32-dim)	$\approx -79.5 \pm 0.3$	-80.3 ± 0.6	
VAE (8-dim)	-98.0 ± 0.5	-91.0 ± 0.9	
VAE (32-dim)	-87.2 ± 0.3	-82.1 ± 0.6	
VAE + HF (T=2)	-79.5	–	(Tomczak & Welling, 2016)
VAE + NF (T=80)	-85.1	–	(Rezende & Mohamed, 2015)
VAE + NICE (T=80)	-88.3	–	(Dinh et al., 2014)
VAE + HVI (T=16)	-88.3	–	(Salimans et al., 2015)
convVAE + HVI (T=16)	-84.1	–	(Salimans et al., 2015)
VAE + VGP (2hl)	-81.3	–	(Tran et al., 2015)
DRAW + VGP	-79.9	–	(Tran et al., 2015)
VAE + IAF	-80.8	–	(Kingma et al., 2016)
Auxiliary VAE (L=2)	-83.0	–	(Maaløe et al., 2016)

Dataset samples

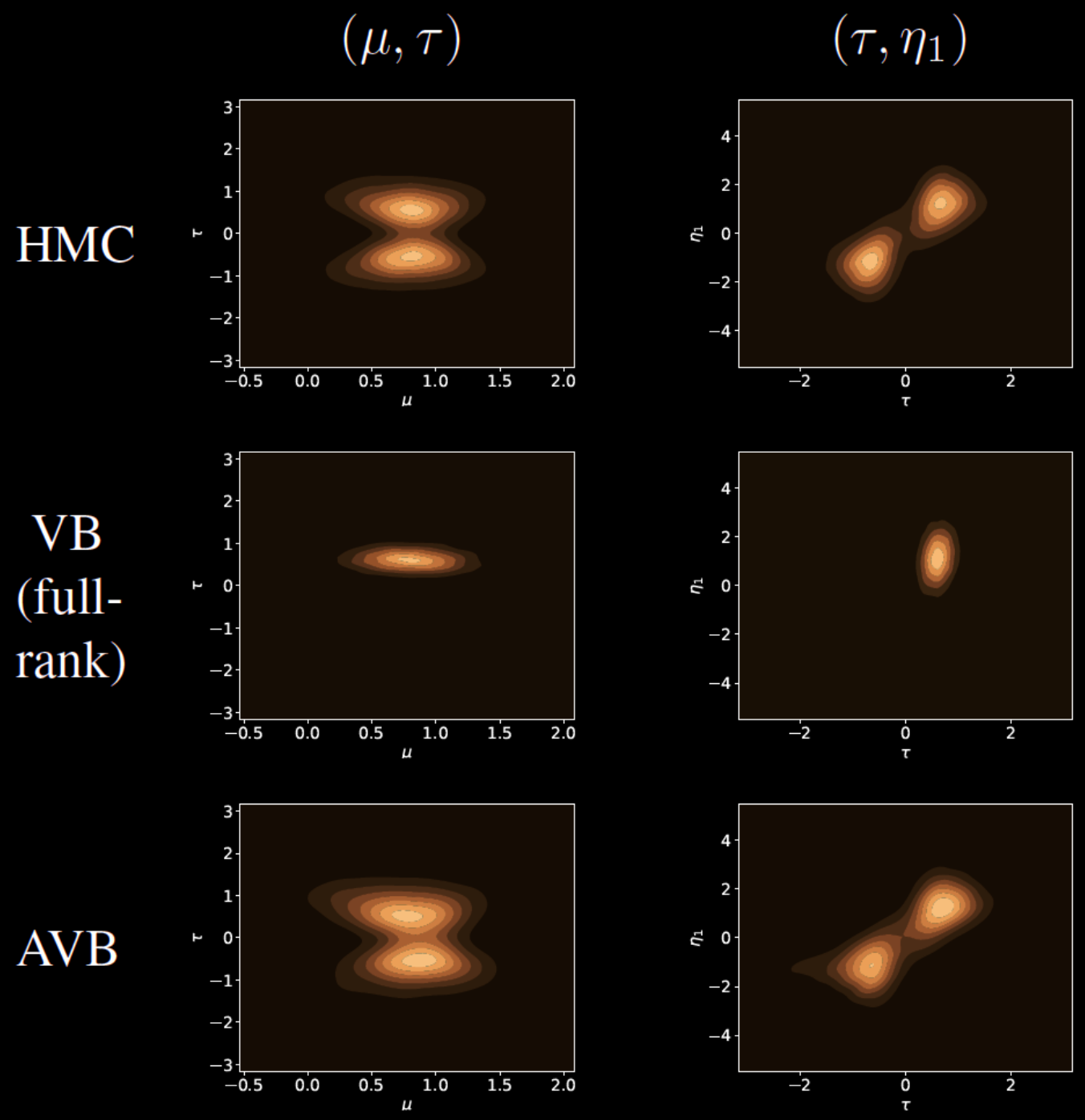
Model samples

CelebA face images

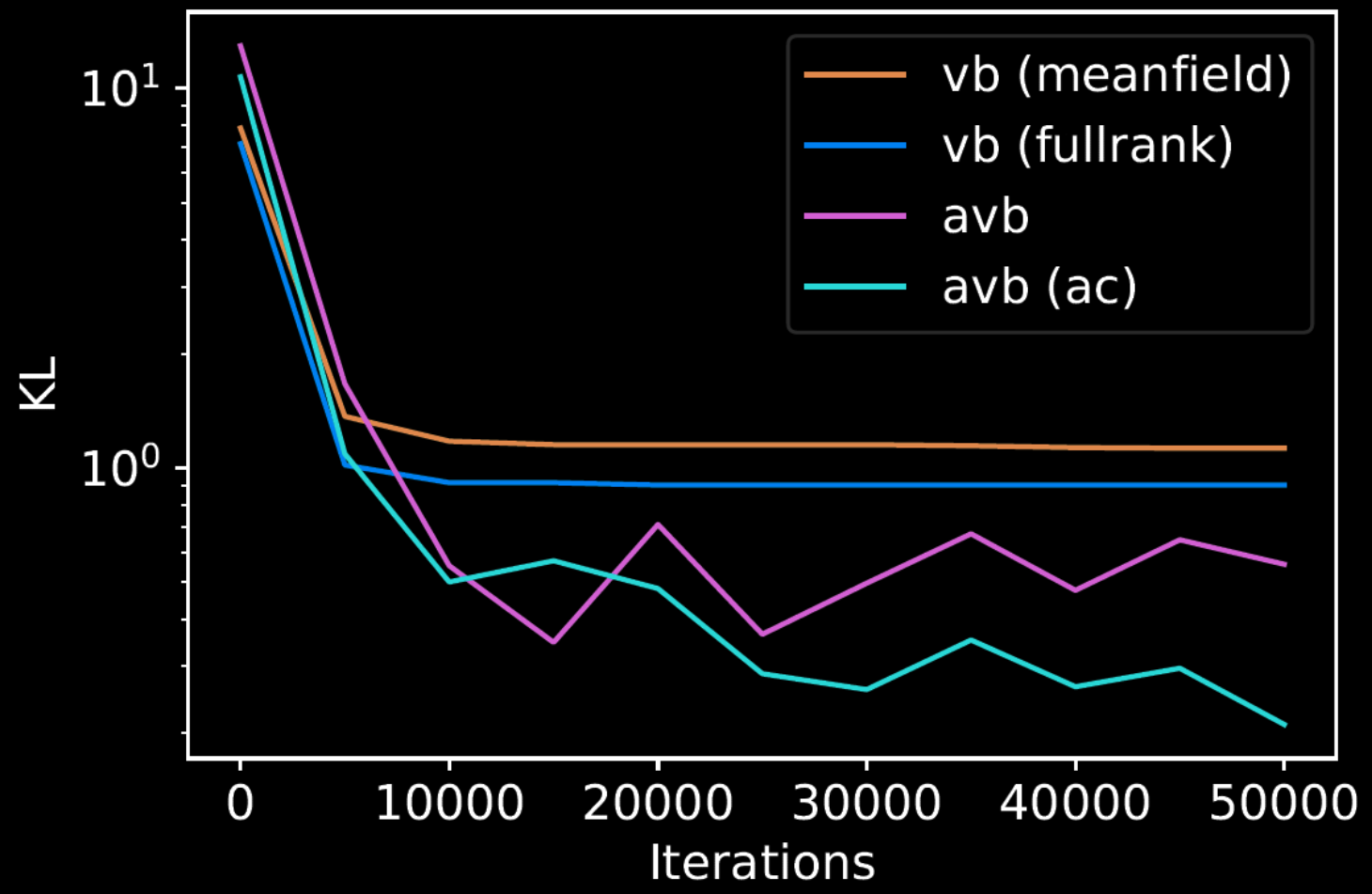
VB for Parameter Inference

- Stan [Stan Development Team, 2016]
- *Eight schools* model [Gelman et al., 2014]
- Eight parameters

- Ground truth posterior:
  MCMC with Hamiltonian Monte Carlo, 500k iterations (PyStan)
- Estimate KL divergence to true posterior
  - ITE package [Szabo, 2013]


(μ, τ)
(τ, η_1)
HMC
VB (full-rank)
AVB
μ
τ
η_1

KL to true posterior



Conclusions

- AVB: likelihood-free variational families
- State-of-the-art performance in competitive VAE field
- Parameter Variational Bayes in large variational families

- Our TensorFlow implementation: https://github.com/LMescheder/AdversarialVariationalBayes
  - Third party implementation from Ben Poole: https://gist.github.com/poolio/b71eb943d6537d01f46e7b20e9225149

Open Research Problems

Quantitative Evaluation Metrics

- Likelihood not applicable
- Not sound: Inception scores
- Other candidates: IPMs, Fréchet Inception Distance (FID)
- Tournaments as evaluation? cf [Roth et al., NIPS 2017]

GANs for discrete data

- Discrete latents (noise source) not a problem in GANs
- Discrete observations are problematic (but no problem in other methods (VAEs, RNNs))
- What makes discrete observables difficult to handle in GANs?
  - Answer: generation of fake data
  - Likelihood-based methods never generate fake data
- Early attempts for discrete GANs:
  - [Hjelm et al., ICLR 2018, “Boundary-seeking GANs”]: importance weights
  - [Guljarani et al., NIPS 2017, WGAN-GP]: continuous one-hot encoding
  - More work to be done

Estimation Uncertainty

- Estimation uncertainty goes away as you have more data
- Likelihood-based probabilistic models =>
  Bayesian posterior over models
- GANs do not have a likelihood nor a well-defined posterior

- Early attempts:
  - [Saatchi and Wilson, “Bayesian GAN”, NIPS 2017]: pseudo-posterior defined by a pair of conditional distributions.
    Interesting, but raises obvious questions: coherence of pseudo-conditionals?
    Cf. [Heckerman, “Dependency Networks”, JMLR 2000]

Practical Bounds on $\|f\|_L$

- Spectral Normalization is sound and efficient, but
  - How loose is the bound?
- Are there multi-layer extensions?
- More powerful regularization/capacity-control via "variational regularization"?

New Divergences

- Generalization of IPMs
  - μ-Fisher IPM [Mroueh and Sercu, 2017]: measure-dependent function norms
  - μ-Sobolev IPM [Mroueh et al., 2017]: measure-dependent gradient norms
  - Implementable
- Are there interesting classes of divergences left?
- Interesting: efficient implementable, consistent in the general case, sample-efficient

Theory

- Generalization
  - [Zhang et al., arXiv:1711.02771]: generalization bounds, showing that discriminator capacity needs to be limited and the rest does not matter
  - [Arora et al., ICLR 2018]: empirical study using “birthday paradox test”, demonstrating GAN failures even when images look good
  - [Arpra et al., 2017]: study of neural network distance (generalization) versus study of divergences
- Questions
  - How to test for generalization empirically?
  - Useful theory: enable choice of capacity and model class

Thanks!

Feedback and comments:
Sebastian.Nowozin@microsoft.com