Generalized Energy-Based Models

Arthur Gretton

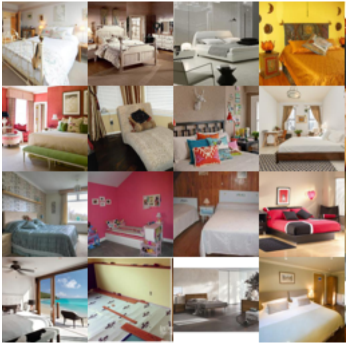
 

Gatsby Computational Neuroscience Unit,
University College London

LSE, 2020

Training generative models

- **Have:** One collection of samples X from unknown distribution P.
- **Goal:** generate samples Q that look like P

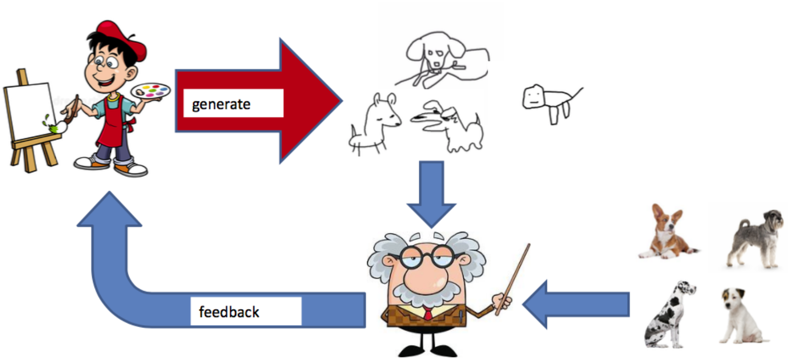


LSUN bedroom samples P

Generated Q, MMD GAN

Role of divergence $D(P, Q)$?

Reminder: generative adversarial network



Outline

- A quick overview of divergence measures (critics)

- Variational lower bound on ϕ-divergences (f-divergences)

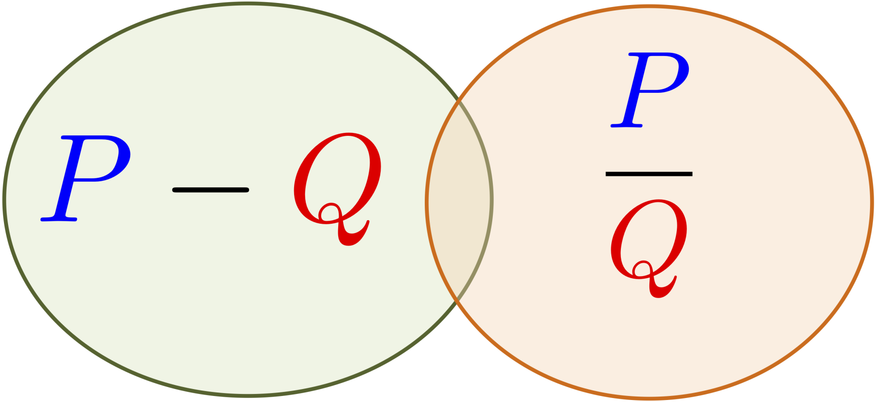
- **Generalized energy-based models**

Arbel, Zhou, G., Generalized Energy Based Models (arXiv 2020)

Key message: all else being equal, incorporating critic into model performs better than using generator alone.

Divergence measures (critics)

Divergences



Divergences

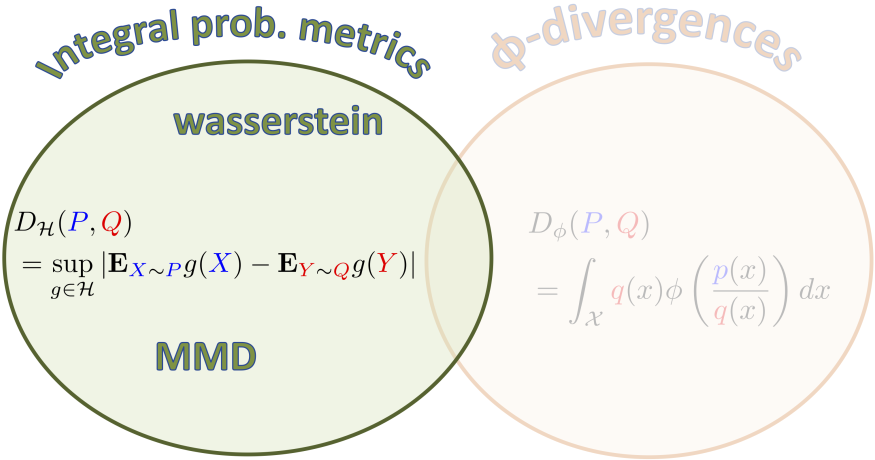
Integral prob. metrics

$$D_{\mathcal{H}}(P, Q) = \sup_{g \in \mathcal{H}} \left| \mathbf{E}_{X \sim P}\, g(X) - \mathbf{E}_{Y \sim Q}\, g(Y) \right|$$

ϕ-divergences

$$D_{\phi}(P, Q) = \int_{\mathcal{X}} q(x)\, \phi\left(\frac{p(x)}{q(x)}\right) dx$$

The Integral Probability Metrics



$$D_{\mathcal{H}}(P, Q) = \sup_{g \in \mathcal{H}} \left| \mathbf{E}_{X \sim P} g(X) - \mathbf{E}_{Y \sim Q} g(Y) \right|$$

$$D_{\phi}(P, Q) = \int_{\mathcal{X}} q(x) \phi\left(\frac{p(x)}{q(x)}\right) dx$$

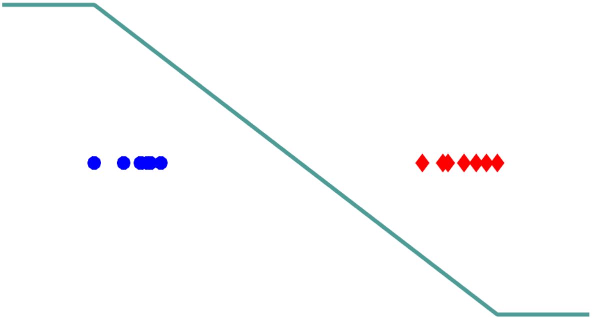
Wasserstein distance

A helpful critic witness:

$$W_1(P, Q) = \sup_{\|f\|_L \leq 1} E_P f(X) - E_Q f(Y).$$

$$\|f\|_L := \sup_{x \neq y} |f(x) - f(y)| / \|x - y\|$$

W_1=0.88



Santambrogio, Optimal Transport for Applied Mathematicians (2015, Section 5.4)
G Peyré, M Cuturi, Computational Optimal Transport (2019)
M. Cuturi, J. Solomon, NeurIPS tutorial (2017)

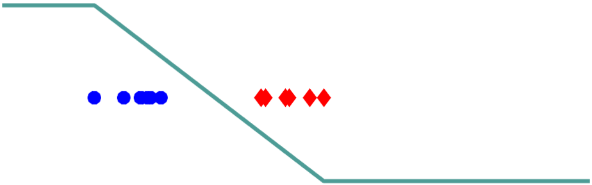
Wasserstein distance

A helpful critic witness:

$$W_1(P, Q) = \sup_{\|f\|_L \le 1} E_P f(X) - E_Q f(Y).$$

$$\|f\|_L := \sup_{x \neq y} |f(x) - f(y)| / \|x - y\|$$

W_1=0.65



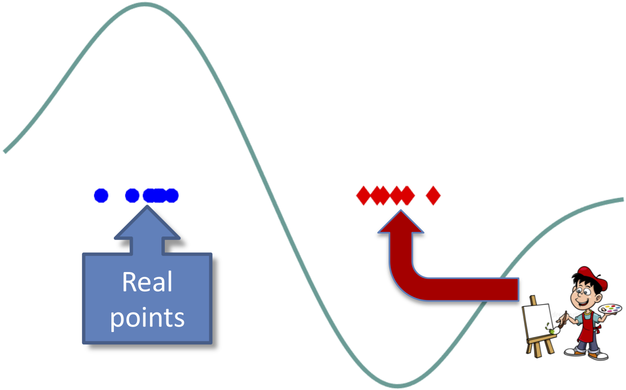
Santambrogio, Optimal Transport for Applied Mathematicians (2015, Section 5.4)
G Peyré, M Cuturi, Computational Optimal Transport (2019)
M. Cuturi, J. Solomon, NeurIPS tutorial (2017)

Maximum mean discrepancy

A helpful critic witness:

$$MMD(P, Q) = \sup_{\|f\|_{\mathcal{F}} \leq 1} E_P f(X) - E_Q f(Y).$$

MMD=1.8

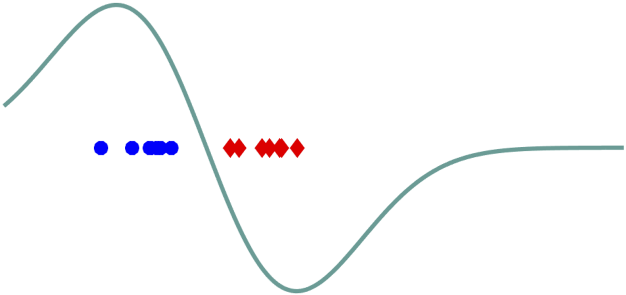


Maximum mean discrepancy

A helpful critic witness:

$MMD(P, Q) = \sup_{\|f\|_{\mathcal{F}} \leq 1} E_P f(X) - E_Q f(Y)$

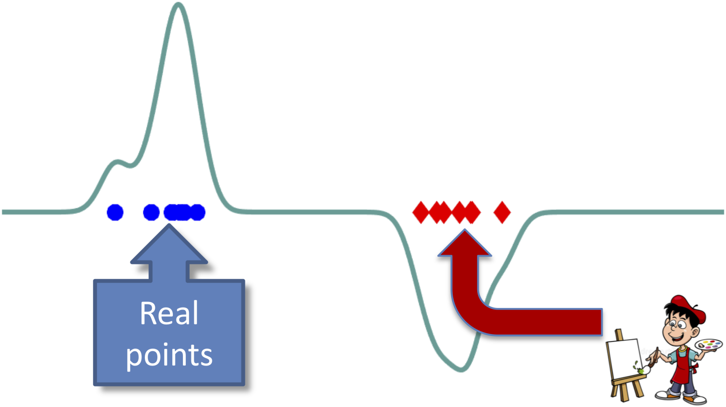
MMD=1.1



Maximum mean discrepancy

An unhelpful critic witness:
$MMD(P, Q)$ with a narrow kernel.

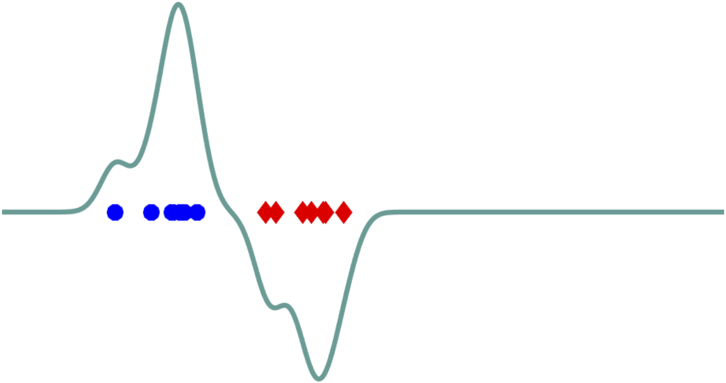
MMD=0.64



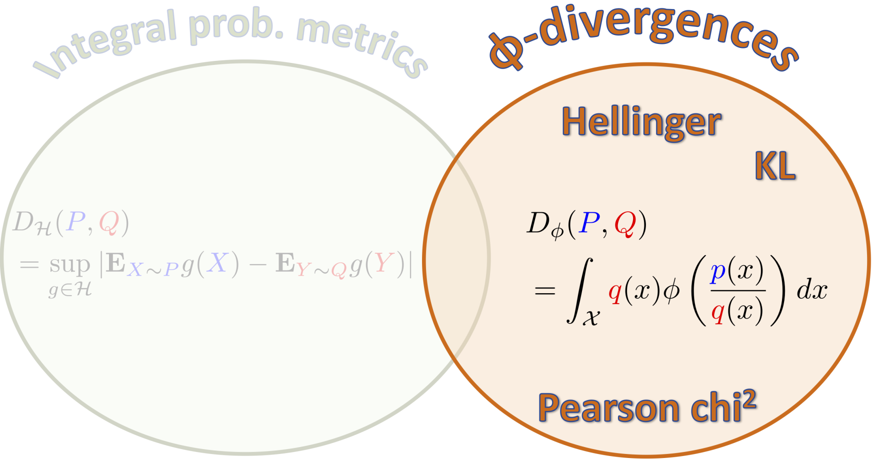
Maximum mean discrepancy

An unhelpful critic witness:
$MMD(P, Q)$ with a narrow kernel.

MMD=0.64



The ϕ-divergences



The ϕ-divergences

Define the ϕ-divergence(f-divergence):

$$D_\phi(P, Q) = \int \phi\left(\frac{p(z)}{q(z)}\right) q(z)dz$$

where ϕ is convex, lower-semicontinuous, $\phi(1) = 0$.

- **Example:** $\phi(u) = u\log(u)$ gives KL divergence,

$$\begin{aligned} D_{KL}(P, Q) &= \int \log\left(\frac{p(z)}{q(z)}\right) p(z)dz \\ &= \int \left(\frac{p(z)}{q(z)}\right) \log\left(\frac{p(z)}{q(z)}\right) q(z)dz \end{aligned}$$

The ϕ-divergences

Define the ϕ-divergence(f-divergence):

$$D_\phi(P, Q) = \int \phi\left(\frac{p(z)}{q(z)}\right) q(z)dz$$

where ϕ is convex, lower-semicontinuous, $\phi(1) = 0$.

- Example: $\phi(u) = u\log(u)$ gives KL divergence,

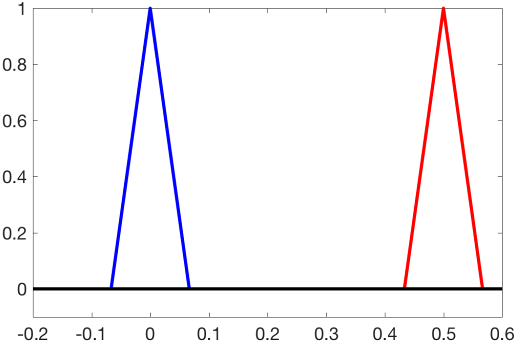
$$\begin{aligned} D_{KL}(P, Q) &= \int \log\left(\frac{p(z)}{q(z)}\right) p(z)dz \\ &= \int \left(\frac{p(z)}{q(z)}\right)\log\left(\frac{p(z)}{q(z)}\right) q(z)dz \end{aligned}$$

Are ϕ-divergences good critics?

Simple example: disjoint support.

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

$$D_{KL}(P, Q) = \infty \qquad D_{JS}(P, Q) = \log 2$$

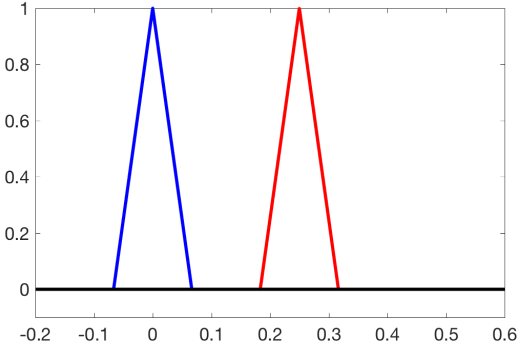


Are ϕ-divergences good critics?

Simple example: disjoint support.

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

$$D_{KL}(Q, P) = \infty \qquad D_{JS}(P, Q) = \log 2$$



A variational lower bound

A lower-bound ϕ-divergence approximation:

$$D_\phi(P, Q) = \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

A variational lower bound

A lower-bound ϕ-divergence approximation:

$$
\begin{aligned}
D_\phi(P, Q) &= \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz \\
&= \int q(z) \underbrace{\sup_{f_z}\left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right)}_{\phi\left(\frac{p(z)}{q(z)}\right)}
\end{aligned}
$$

$\phi^*(u)$ is dual of $\phi(u)$.

A variational lower bound

A lower-bound ϕ-divergence approximation:

$$
\begin{aligned}
D_\phi(P, Q) &= \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz \\
&= \int q(z) \sup_{f_z} \left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right) \\
&\geq \sup_{f \in \mathcal{H}} \mathbf{E}_P f(X) - \mathbf{E}_Q \phi^*\left(f(Y)\right)
\end{aligned}
$$

(restrict the function class)

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

A variational lower bound

A lower-bound ϕ-divergence approximation:

$$
\begin{aligned}
D_\phi(P, Q) &= \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz \\
&= \int q(z) \sup_{f_z}\left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right) \\
&\geq \sup_{f \in \mathcal{H}} \mathbf{E}_P f(X) - \mathbf{E}_Q \phi^*\left(f(Y)\right)
\end{aligned}
$$

(restrict the function class)

Bound tight when:

$$
f^\diamond(z) = \partial\phi\left(\frac{p(z)}{q(z)}\right)
$$

if ratio defined.

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

$$\geq \sup_{f \in \mathcal{H}} -\mathbf{E}_P f(X) + 1 - \mathbf{E}_Q \underbrace{\exp\left(-f(Y)\right)}_{\phi^*(-f(Y)+1)}$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

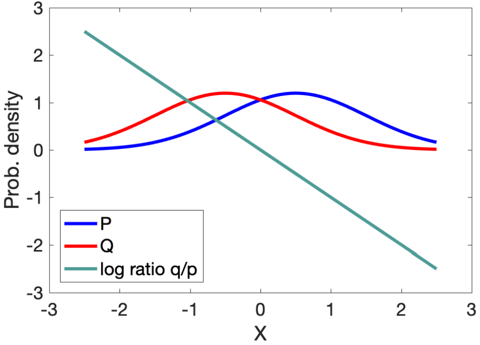
$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

$$\geq \sup_{f \in \mathcal{H}} -\mathbf{E}_P f(X) + 1 - \mathbf{E}_Q \exp\left(-f(Y)\right)$$

Bound tight when:

$$f^\diamond(z) = -\log \frac{p(z)}{q(z)}$$

if ratio defined.



Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

$$\geq \sup_{f \in \mathcal{H}} -\mathbf{E}_P f(X) + 1 - \mathbf{E}_Q \exp\left(-f(Y)\right)$$

$$\approx \sup_{f \in \mathcal{H}} \left[-\frac{1}{n}\sum_{j=1}^{n} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} \exp(-f(y_i)) \right] + 1$$

$$x_i \overset{\text{i.i.d.}}{\sim} P$$

$$y_i \overset{\text{i.i.d.}}{\sim} Q$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

$$\geq \sup_{f \in \mathcal{H}} -\mathbf{E}_P f(X) + 1 - \mathbf{E}_Q \exp(-f(Y))$$

$$\approx \sup_{f \in \mathcal{H}} \left[-\frac{1}{n}\sum_{j=1}^{n} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} \exp(-f(y_i)) \right] + 1$$

This is a

KL

Approximate

Lower-bound

Estimator.

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\, dz$$

$$\geq \sup_{f \in \mathcal{H}} -\mathbf{E}_P f(X) + 1 - \mathbf{E}_Q \exp(-f(Y))$$

$$\approx \sup_{f \in \mathcal{H}} \left[-\frac{1}{n} \sum_{j=1}^{n} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} \exp(-f(y_i)) \right] + 1$$

This is a

K

A

L

E

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

$$\geq \sup_{f\in\mathcal{H}} -\mathbf{E}_P f(X) + 1 - \mathbf{E}_Q \exp\left(-f(Y)\right)$$

$$\approx \sup_{f\in\mathcal{H}} \left[-\frac{1}{n}\sum_{j=1}^{n} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} \exp(-f(y_i))\right] + 1$$

The KALE divergence

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Topological properties of KALE (1)

Key requirements on $\mathcal{H}$ and $\mathcal{X}$:

- Compact domain $\mathcal{X}$,
- $\mathcal{H}$ dense in the space $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$ wrt $\|\cdot\|_\infty$.
- If $f \in \mathcal{H}$ then $-f \in \mathcal{H}$ and $cf \in \mathcal{H}$ for $0 \leq c \leq C_{\max}$.

Theorem: $KALE(P, Q; \mathcal{H}) \geq 0$ and $KALE(P, Q; \mathcal{H}) = 0$ iff $P = Q$.

Zhang, Liu, Zhou, Xu, and He. "On the Discrimination-Generalization Tradeoff in GANs" (ICLR 2018, Corollary 2.4; Theorem B.1)
Arbel, Liang, G. (arXiv 2020, Proposition 1)

Topological properties of KALE (1)

Key requirements on $\mathcal{H}$ and $\mathcal{X}$:

- Compact domain $\mathcal{X}$,
- $\mathcal{H}$ dense in the space $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$ wrt $\|\cdot\|_\infty$.
- If $f \in \mathcal{H}$ then $-f \in \mathcal{H}$ and $cf \in \mathcal{H}$ for $0 \leq c \leq C_{\text{max}}$.

Theorem: $KALE(P, Q; \mathcal{H}) \geq 0$ and $KALE(P, Q; \mathcal{H}) = 0$ iff $P = Q$.

$\mathcal{H}$ dense in $C(\mathcal{X})$ for $\mathcal{X} \subset \mathbb{R}^d$ when:

$$\mathcal{H} = \operatorname{span}\{\sigma(w \top x + b) : [w, b] \in \Theta\}$$

$\sigma(u) = \max\{u, 0\}^\alpha$, $\alpha \in \mathbb{N}$, and $\{\lambda\theta : \lambda \geq 0, \theta \in \Theta\} = \mathbb{R}^{d+1}$.

Zhang, Liu, Zhou, Xu, and He. "On the Discrimination-Generalization Tradeoff in GANs"

(ICLR 2018, Corollary 2.4; Theorem B.1)

Arbel, Liang, G. (arXiv 2020, Proposition 1)

Topological properties of KALE (2)

Additional requirement: all functions in $\mathcal{H}$ Lipschitz in their inputs with constant L

Theorem: $KALE(P, Q^n; \mathcal{H}) \to 0$ iff $Q^n \to P$ under the weak topology.

Liu, Bousquet, Chaudhuri. "Approximation and Convergence Properties of Generative Adversarial Learning" (NeurIPS 2017); Arbel, Liang, G. (arXiv 2020, Proposition 1)

Topological properties of KALE (2)

Additional requirement: all functions in $\mathcal{H}$ Lipschitz in their inputs with constant L

Theorem: $KALE(P, Q^n; \mathcal{H}) \to 0$ iff $Q^n \to P$ under the weak topology.

Partial proof idea:

$$
\begin{aligned}
KALE(P, Q; \mathcal{H}) &= -\int f \, dP - \int \exp(-f) \, dQ + 1 \\
&= \int f(x) \, dQ(x) - f(x') \, dP(x') \\
&\quad - \int \underbrace{(\exp(-f) + f - 1)}_{\geq 0} \, dQ \\
&\leq \int f(x) \, dQ(x) - f(x') \, dP(x') \leq LW_1(P, Q)
\end{aligned}
$$

Liu, Bousquet, Chaudhuri. "Approximation and Convergence Properties of Generative Adversarial Learning" (NeurIPS 2017); Arbel, Liang, G. (arXiv 2020, Proposition 1)

Empirical properties of KALE

$$KALE(P, Q; \mathcal{H}) = \sup_{f \in \mathcal{H}} -E_P f(X) - E_Q \exp(-f(Y)) + 1$$

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

$\|w\|^2_{\mathcal{H}}$ penalized :

Empirical properties of KALE

$$KALE(P, Q; \mathcal{H}) = \sup_{f \in \mathcal{H}} -E_P f(X) - E_Q \exp(-f(Y)) + 1$$

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

$$\|w\|_{\mathcal{H}}^2 \quad \text{penalized : KALE smoothie}$$

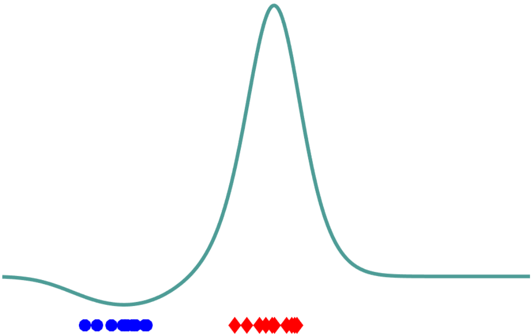
Empirical properties of KALE

$$KALE(P, Q; \mathcal{H}) = \sup_{f \in \mathcal{H}} -E_P f(X) - E_Q \exp(-f(Y)) + 1$$

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

$\|w\|^2_{\mathcal{H}}$ penalized : KALE smoothie

$$KALE(Q, P; \mathcal{H}) = 0.18$$



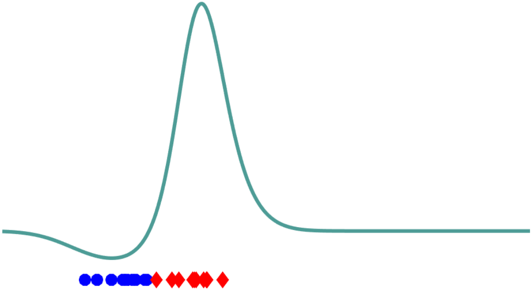
Empirical properties of KALE

$$KALE(P, Q; \mathcal{H}) = \sup_{f \in \mathcal{H}} -E_P f(X) - E_Q \exp(-f(Y)) + 1$$

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

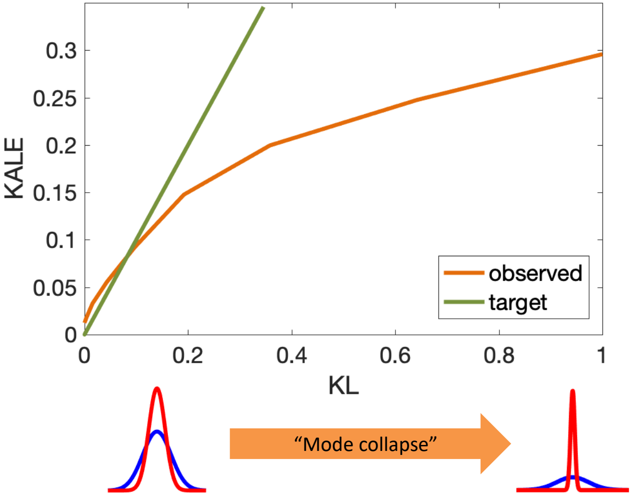
$$\|w\|^2_{\mathcal{H}} \quad \text{penalized : KALE smoothie}$$

$$KALE(Q, P; \mathcal{H}) = 0.12$$



The KALE smoothie and “mode collapse”

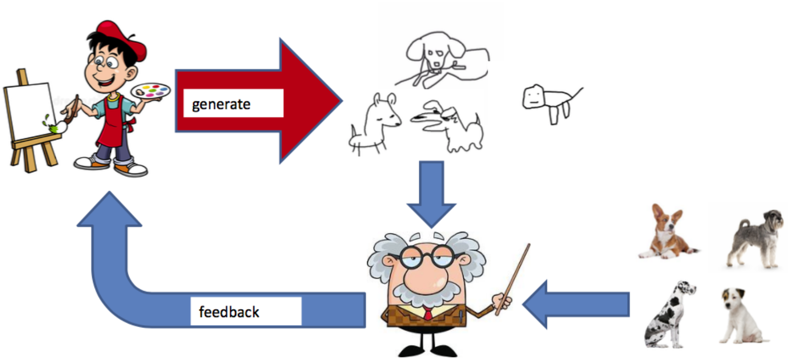
- Two Gaussians with same means, different variance



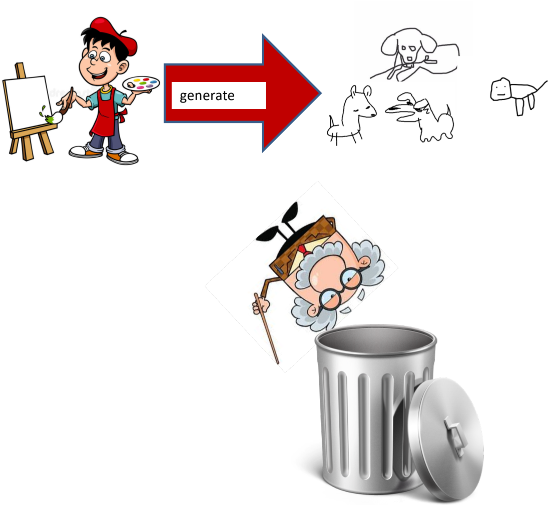
Example thanks to M. Arbel and M. Rosca

Generalized Energy-Based Models

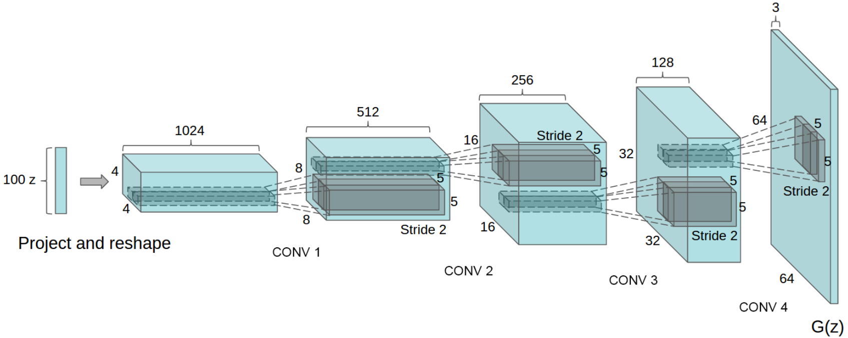
Visual notation: GAN setting



Visual notation: GAN setting



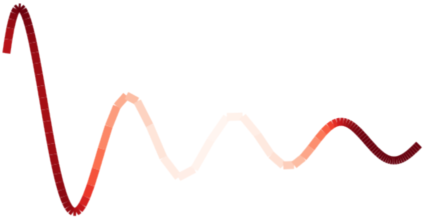
Reminder: the generator



Radford, Metz, Chintala, ICLR 2016

Generalized energy-based models: illustration

Target distribution P



$$z \sim Unif[0,1]$$

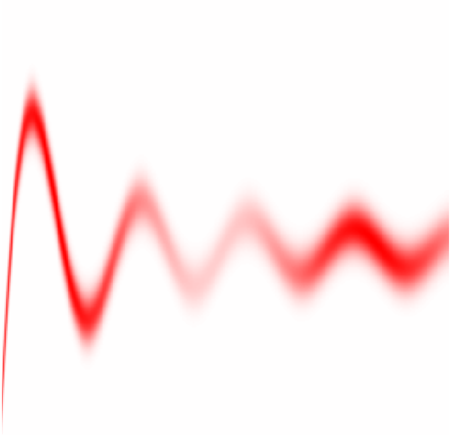
$$\tilde{z} = \tau(z)$$

$$X = G_{\theta^\star}(\tilde{z}), \quad X_1 = \tilde{z}$$

Example thanks to M. Arbel

Generalized energy-based models: illustration

EBM approximation to target:



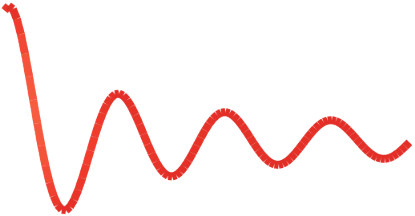
$$p(X) \propto \exp(-E(X))$$

$$E(X) = \frac{1}{2\sigma^2} \| G_\theta(X_1) - X \|^2 + A_\theta(X_1)$$

Example thanks to M. Arbel

Generalized energy-based models: illustration

GAN (generator) distribution Q_θ

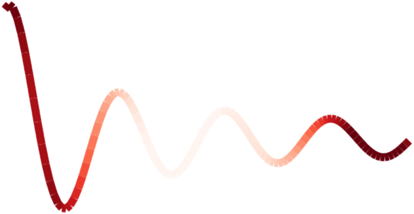


Generator	Critic
$z \sim unif[0,1]$	$MLP(X)$
$X = B_\theta(z)$	

Example thanks to M. Arbel

Generalized energy-based models: illustration

Mass of GEBM corrected by critic



Generator

$z \sim unif[0, 1]$

$X = B_\theta(z)$

Re-weight using importance weights defined by energy:

$$w(x) \propto \exp(-E(x))$$

Example thanks to M. Arbel

Generalized energy-based models

Define a model $Q_{B_\theta, E}$ as follows:

- Sample from generator with parameters θ

$$X \sim Q_\theta \quad \Longleftrightarrow \quad X = B_\theta(Z), \quad Z \sim \eta$$

- Reweight the samples according to importance weights:

$$f_{Q,E}(x) = \frac{\exp(-E(x))}{Z_{Q_\theta, E}}, \qquad Z_{Q,E} = \int \exp(-E(x))\, dQ_\theta(x),$$

where $E \in \mathcal{E}$, the energy function class.

$f_{Q,E}(x)$ is Radon-Nikodym derivative of $Q_{B_\theta, E}$ wrt Q_θ.

- When Q_θ has density wrt Lebesgue on $\mathcal{X}$, this is a standard energy-based model.

Generalized Energy-Based Models

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\,dP = -\int E\,dP - \log Z_{Q,E}$$

- When $KL(P, Q_\theta)$ well defined, above is Donsker-Varadhan lower bound on KL
  - tight when $E(z) = -\log(p(z)/q(z))$.
- However, Generalized Log-Likelihood still defined when P and Q_θ mutually singular!

arXiv.org > stat > arXiv:2003.05033

Statistics > Machine Learning

[Submitted on 10 Mar 2020 (v1), last revised 24 Jun 2020 (this version, v3)]

Generalized Energy Based Models

Michael Arbel, Liang Zhou, Arthur Gretton

arXiv.org > cs > arXiv:2003.06060

Search...

Help | Advanc

Computer Science > Machine Learning

[Submitted on 12 Mar 2020 (v1), last revised 24 Mar 2020 (this version, v2)]

Your GAN is Secretly an Energy-based Model and You Should use Discriminator Driven Latent Sampling

Tong Che, Ruixiang Zhang, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, Yoshua Bengio

https://github.com/MichaelArbel/GeneralizedEBM

Learning the energy function: amortized approach

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log Z_{Q,E}$$

Learning the energy function: amortized approach

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log Z_{Q,E}$$

Don't do this: minibatch estimate of $\log Z_{Q,E}$ (large variance)

$$\widehat{\log(Z_{Q,E})} = \log\left(\frac{1}{n}\sum_{i=1}^{n} \exp\left(-E[B_{\theta}(z_i)]\right)\right) \qquad z_i \overset{\text{i.i.d.}}{\sim} \eta$$

Learning the energy function: amortized approach

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log Z_{Q,E}$$

Instead, do this: from convexity of exponential,

$$-\log(Z_{Q,E}) \geq -c - \exp(-c) Z_{Q,E} + 1$$

tight whenever $c = \log(Z_{Q,E})$.

Learning the energy function: amortized approach

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log Z_{Q,E}$$

Instead, do this: from convexity of exponential,

$$-\log(Z_{Q,E}) \geq -c - \exp(-c) Z_{Q,E} + 1$$

tight whenever $c = \log(Z_{Q,E})$.

Generalized Log-Likelihood has the lower bound:

$$\mathcal{L}_{P,Q}(E) \geq -\int (E + c)\, dP - \int \exp(-(E + c))\, dQ_\theta + 1$$

$$:= \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})$$

Learning the energy function: amortized approach

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log Z_{Q,E}$$

Instead, do this: from convexity of exponential,

$$-\log(Z_{Q,E}) \geq -c - \exp(-c) Z_{Q,E} + 1$$

tight whenever $c = \log(Z_{Q,E})$.

Generalized Log-Likelihood has the lower bound:

$$\begin{aligned}\mathcal{L}_{P,Q}(E) &\geq -\int (E + c)\, dP - \int \exp(-(E + c))\, dQ_\theta + 1\\ &:= \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})\end{aligned}$$

Jointly maximizing yields the maximum likelihood energy E^* and corresponding $c^* = \log(Z_{Q,E^*})$.

Learning the base measure (generator)

Recall the generator:

$$X = B_\theta(Z), \quad Z \sim \eta$$

Define: $\mathcal{K}(\theta) := \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})$

Learning the base measure (generator)

Recall the generator:

$$X = B_\theta(Z), \quad Z \sim \eta$$

Define: $\mathcal{K}(\theta) := \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})$

Theorem: $\mathcal{K}$ is lipschitz and differentiable for almost all $\theta \in \Theta$ with:

$$\nabla \mathcal{K}(\theta) = Z_{Q,E^*}^{-1} \int \nabla_x E^*(B_\theta(z)) \nabla_\theta B_\theta(z) \exp(-E^*(B_\theta(z)))\eta(z)\,dz.$$

where E^* achieves supremum in $\mathcal{F}(P, Q; \mathcal{E} + \mathbb{R})$.

Learning the base measure (generator)

Recall the generator:

$$X = B_\theta(Z), \quad Z \sim \eta$$

Define: $\mathcal{K}(\theta) := \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})$

Theorem: $\mathcal{K}$ is lipschitz and differentiable for almost all $\theta \in \Theta$ with:

$$\nabla \mathcal{K}(\theta) = Z_{Q,E^*}^{-1} \int \nabla_x E^*(B_\theta(z)) \nabla_\theta B_\theta(z) \exp(-E^*(B_\theta(z)))\eta(z)dz.$$

where E^* achieves supremum in $\mathcal{F}(P, Q; \mathcal{E} + \mathbb{R})$.

Assumptions:

- Functions in $\mathcal{E}$ parametrized by $\psi \in \Psi$, where Ψ compact,
  - jointly continous w.r.t. (ψ, x), L-lipschitz and L-smooth w.r.t. x.
- $(\theta, z) \mapsto B_\theta(z)$ jointly continuous wrt (θ, z), $z \mapsto B_\theta(z)$ uniformly Lipschitz w.r.t. z, lipschitz and smooth wrt θ (see paper: constants depend on z)

Sampling from the model

Consider end-to-end model $Q_{B_\theta, E}$, where recall that
$X = B_\theta(Z), \quad Z \sim \eta,$

$$f_{B,E}(x) := \frac{\exp(-E(x))}{Z_{Q,E}}$$

Sampling from the model

Consider end-to-end model $Q_{B_\theta,E}$, where recall that
$X = B_\theta(Z), \quad Z \sim \eta,$

$$f_{B,E}(x) := \frac{\exp(-E(x))}{Z_{Q,E}}$$

For a test function g,

$$\int g(x)\, dQ_{B,E}(x) = \int g(B(z)) f_{B,E}(B(z)) \eta(z)\, dz$$

Posterior latent distribution therefore

$$\nu_{B,E}(z) = \eta(z) f_{B,E}(B(z))$$

Sampling from the model

Consider end-to-end model $Q_{B_\theta,E}$, where recall that
$X = B_\theta(Z), \quad Z \sim \eta,$

$$f_{B,E}(x) := \frac{\exp(-E(x))}{Z_{Q,E}}$$

For a test function g,

$$\int g(x)\, dQ_{B,E}(x) = \int g(B(z)) f_{B,E}(B(z))\eta(z)\, dz$$

Posterior latent distribution therefore

$$\nu_{B,E}(z) = \eta(z) f_{B,E}(B(z))$$

Sample $z \sim \nu_{B,E}$ via Langevin diffusion-derived algorithms (MALA, ULA, HMC,...) to exploit gradient information.

Generate new samples in $\mathcal{X}$ via

$$X \sim Q_{B,E} \quad \Longleftrightarrow \quad Z \sim \nu_{B,E}, \quad X = B_\theta(Z).$$

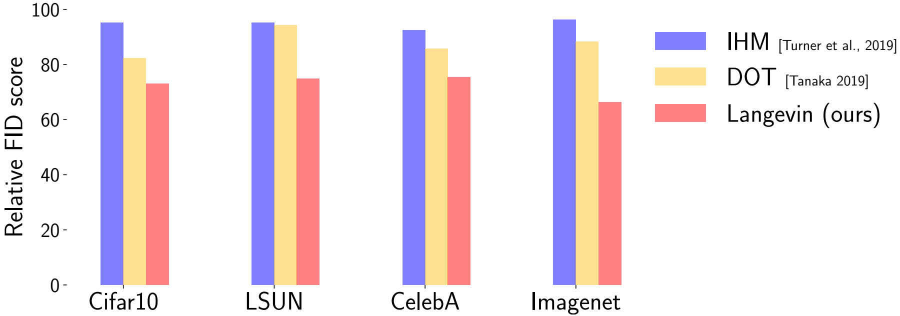
Experiments

Examples: sampling at modes

Tempered GEBM Cifar10 samples at different stages of sampling using a Kinetic Langevin Algorithm (KLA). Early samples $\rightarrow$ late samples. Model run at low temperature ($\beta = 100$) for better quality samples.

Sampling at modes: results

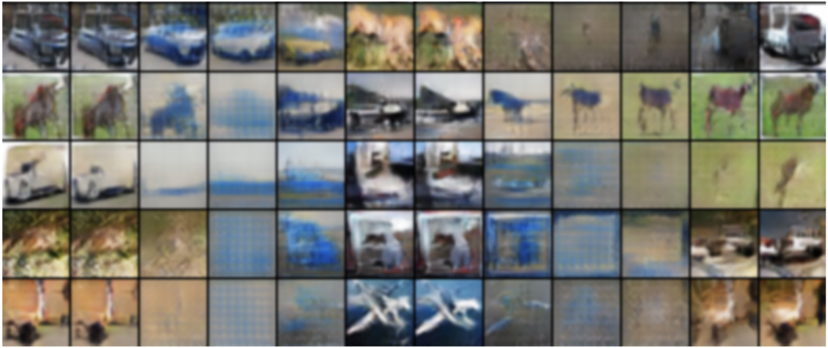
The relative FID score: $\frac{\mathrm{FID}(Q_{B_{\theta},E})}{\mathrm{FID}(B_{\theta})}$



For a given generator B_{θ} and energy E, samples always better (FID score) than generator alone.

Examples: moving between modes

Tempered GEBM Cifar10 samples at different stages of sampling using KLA. Early samples $\rightarrow$ late samples.
Model run at lower friction (but still low temperature, $\beta = 100$) for mode exploration.



Summary

- Generalized energy based model:
  - End-to-end model incorporating generator and critic
  - Always better samples than generator alone.

Demystifying MMD GANs, ICLR 2018:
https://github.com/mbinkowski/MMD-GAN

Gradient regularised MMD, NeurIPS 2018:
https://github.com/MichaelArbel/Scaled-MMD-GAN

Generalized Energy-Based Models, arXiv 2020:
https://github.com/MichaelArbel/GeneralizedEBM

Questions?

Post-credit scene: MMD flow

From NeurIPS 2019:

Maximum Mean Discrepancy Gradient Flow

Michael Arbel
Gatsby Computational Neuroscience Unit
University College London
michael.n.arbel@gmail.com

Anna Korba
Gatsby Computational Neuroscience Unit
University College London
a.korba@ucl.ac.uk

Adil Salim
Visual Computing Center
KAUST
adil.salim@kaust.edu.sa

Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London
arthur.gretton@gmail.com

Sanity check: reduction to EBM case

Base measure B_θ is real NVP with closed-form density.