Generalized Energy-Based Models

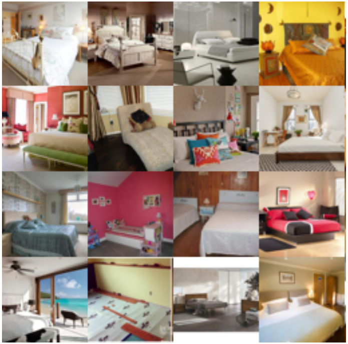
Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

Simons Institute, Berkeley, 2022

Training generative models

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

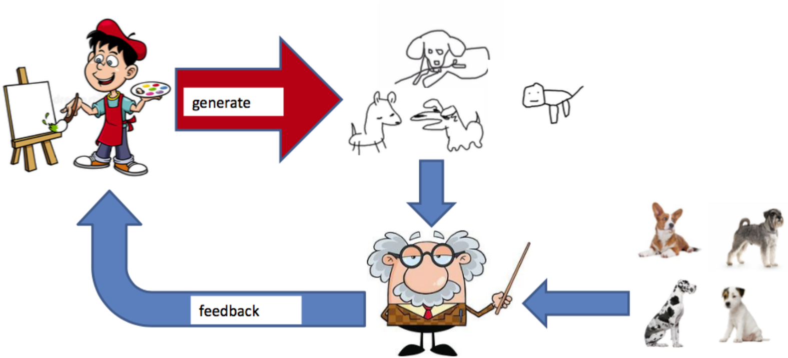


LSUN bedroom samples P

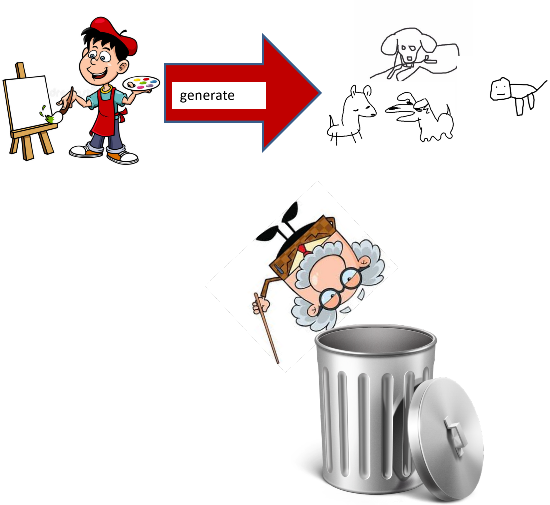
Generated Q, MMD GAN

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

Visual notation: GAN setting



Visual notation: GAN setting



Outline

Divergences $D(P, Q)$

- Integral probability metrics (MMD, Wasserstein)
- ϕ-divergences (f-divergences) and a variational lower bound (KL)

Generalized energy-based models:

- Energy-based model supported on a low-dimenional subspace
- Robust to mismatch in model/data support

Arbel, Zhou, G., Generalized Energy Based Models (ICLR 2021)

Outline

Divergences $D(P, Q)$

- Integral probability metrics (MMD, Wasserstein)
- ϕ-divergences (f-divergences) and a variational lower bound (KL)

Generalized energy-based models:

- Energy-based model supported on a low-dimenional subspace
- Robust to mismatch in model/data support

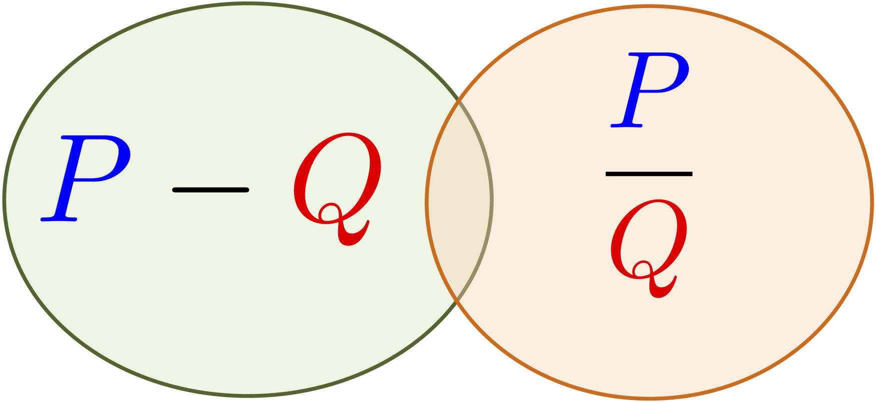
GEBM vs GAN:

- "Like a GAN" but incorporate critic into sample generation
- Perform better than using generator alone

Arbel, Zhou, G., Generalized Energy Based Models (ICLR 2021)

Divergence measures (critics)

Divergences



Divergences

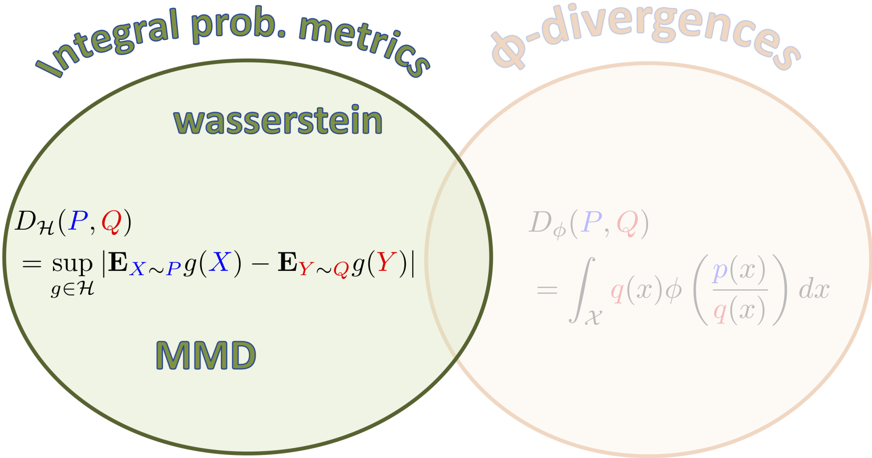
Integral prob. metrics

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

Φ-divergences

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

The Integral Probability Metrics



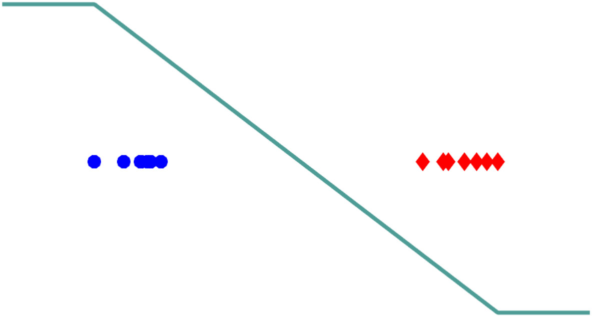
Wasserstein distance

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)

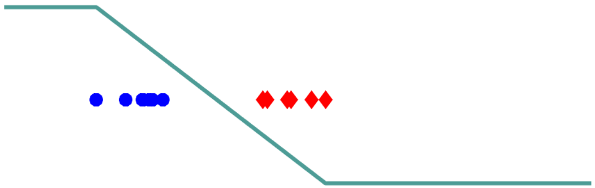
Wasserstein distance

A helpful critic witness:

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

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

W_1=0.65



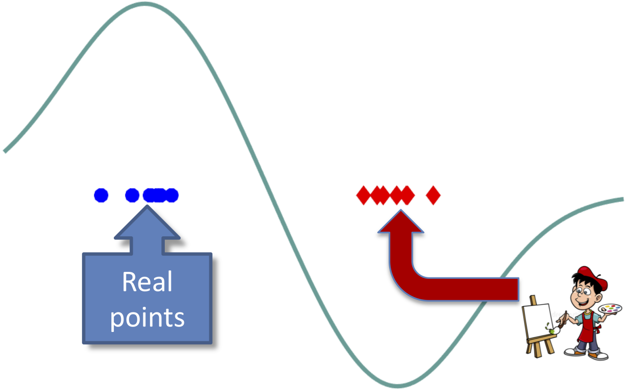
Santambrogio, Optimal Transport for Applied Mathematicians (2015, Section 5.4)
G Peyré, M Cuturi, Computational Optimal Transport (2019)
M. Cuturi, J. Solomon, NeurIPS tutorial (2017)

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

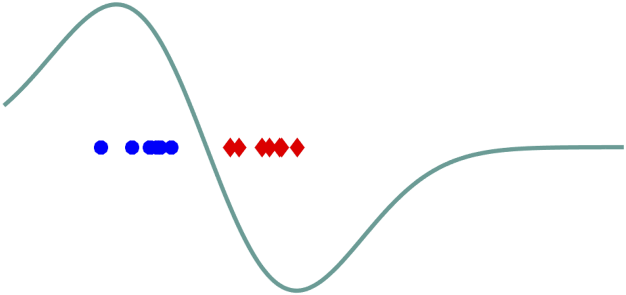


Maximum mean discrepancy

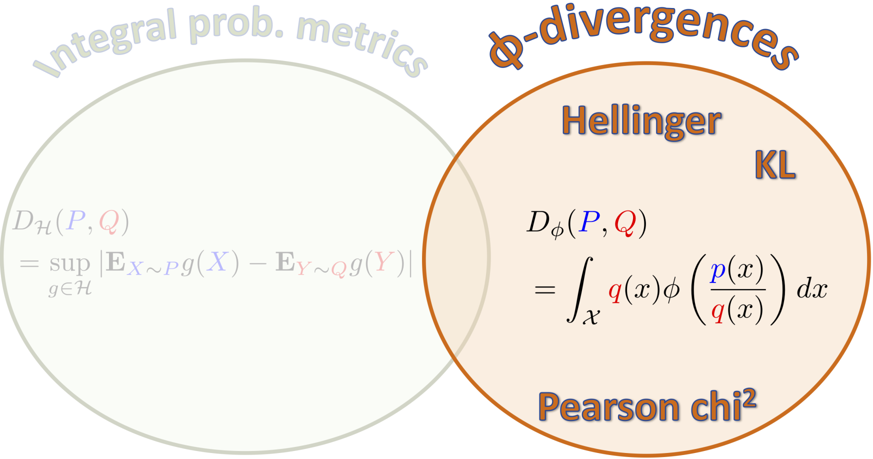
A helpful critic witness:

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

MMD=1.1



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.

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, $\phi(1) = 0$.

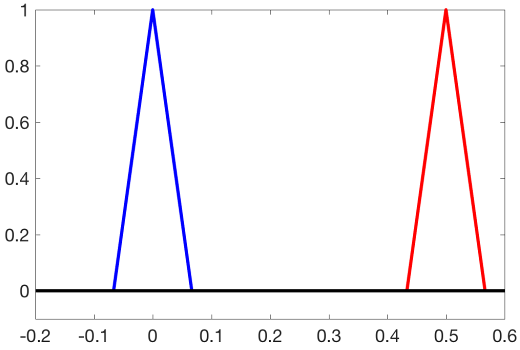
- Example: $\phi(u) = u\log(u)$ gives KL divergence.

Are ϕ-divergences good critics?

Simple example: disjoint support.

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

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

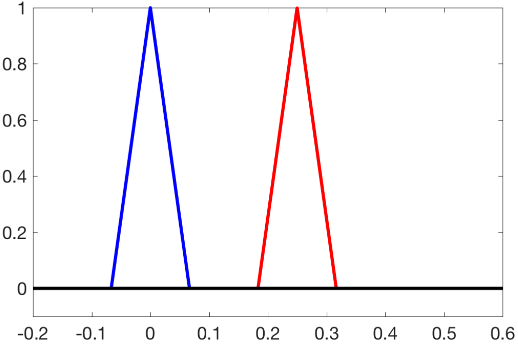


Are ϕ-divergences good critics?

Simple example: disjoint support.

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

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



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:

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

$\phi^*(v)$ is dual of $\phi(x)$.

A variational lower bound

A lower-bound ϕ-divergence approximation:

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

(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:

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

(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

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

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

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

Case of the KL

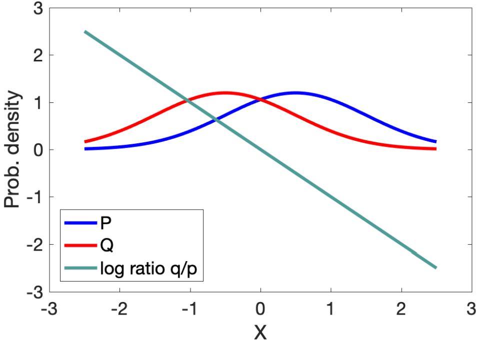
$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

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

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

$$
\begin{aligned}
D_{KL}(P, Q) &= \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz \\
&\geq \sup_{f \in \mathcal{H}} -\mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(-f(Y)\right) \\
&\approx \sup_{f \in \mathcal{H}} \left[-\frac{1}{n} \sum_{j=1}^{n} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} \exp(-f(y_i)) \right] + 1
\end{aligned}
$$

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

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

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

Case of the KL

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

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

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

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

$$
\begin{aligned}
D_{KL}(P, Q) &= \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz \\
&\geq \sup_{f \in \mathcal{H}} -\mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(-f(Y)\right) \\
&\approx \sup_{f \in \mathcal{H}} \left[-\frac{1}{n} \sum_{j=1}^{n} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} \exp(-f(y_i)) \right] + 1
\end{aligned}
$$

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

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

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

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

The KALE divergence

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

Empirical properties of KALE

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

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

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

Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (NeurIPS 2021, Section 2)

Empirical properties of KALE

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

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

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

Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (NeurIPS 2021, Section 2)

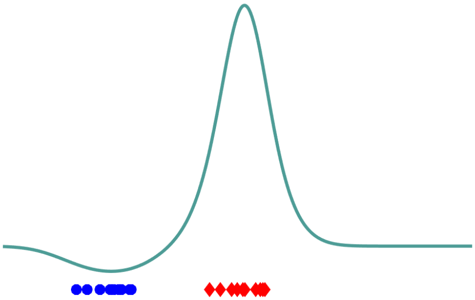
Empirical properties of KALE

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

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

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

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



Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (NeurIPS 2021, Section 2)

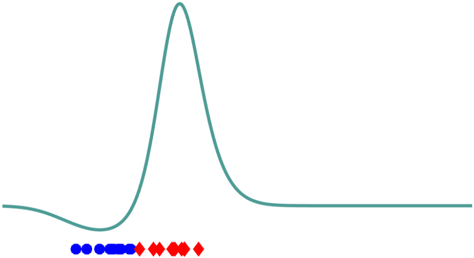
Empirical properties of KALE

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

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

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

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



Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (NeurIPS 2021, Section 2)

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. (ICLR 2021, 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 $0 \leq c \leq C_{\max}$.

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. (ICLR 2021, 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. (ICLR 2021, 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.

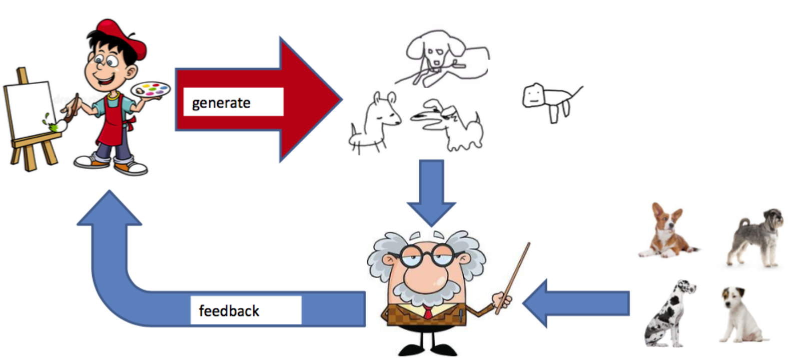
Partial proof idea:

$$\begin{aligned} KALE(P, Q; \mathcal{H}) &= -\int f \, dP - \int \exp(-f) \, dQ + 1 \\ &= \int f(x) \, dQ(x) - f(x') \, dP(x') \\ &\quad - \int \underbrace{(\exp(-f) + f - 1)}_{\geq 0} \, dQ \\ &\leq \int f(x) \, dQ(x) - f(x') \, dP(x') \leq L W_1(P, Q) \end{aligned}$$

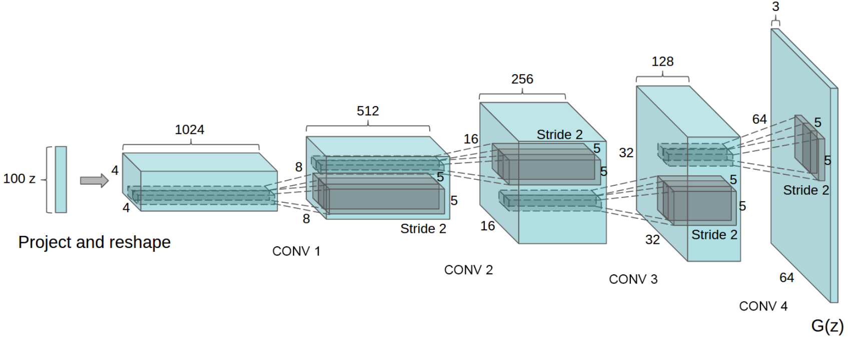
Liu, Bousquet, Chaudhuri. "Approximation and Convergence Properties of Generative Adversarial Learning" (NeurIPS 2017); Arbel, Liang, G. (ICLR 2021, Proposition 1)

Generalized Energy-Based Models

Visual notation: GAN setting



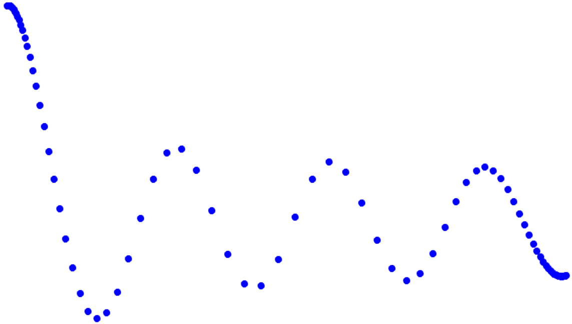
Reminder: the generator



Radford, Metz, Chintala, ICLR 2016

Generalized Energy-Based Models - the idea

Target distribution P



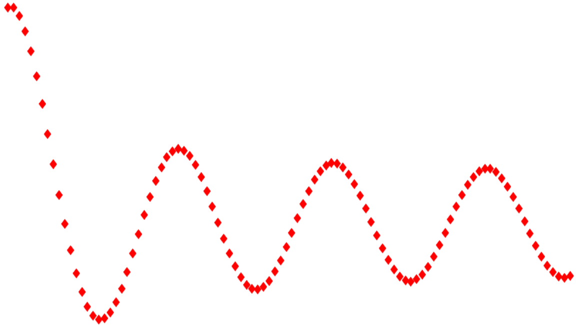
Arbel, Zhou, G. (ICLR 2021)

Generalized Energy-Based Models - the idea

GAN (generator)

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

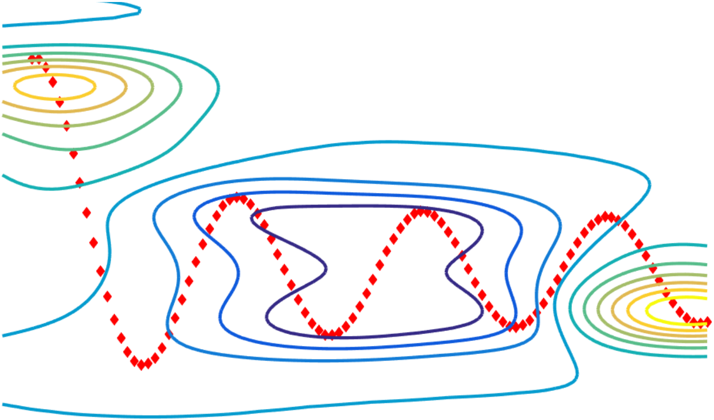
correct support but wrong mass



Arbel, Zhou, G. (ICLR 2021)

Generalized Energy-Based Models - the idea

Log energy function and Q_θ



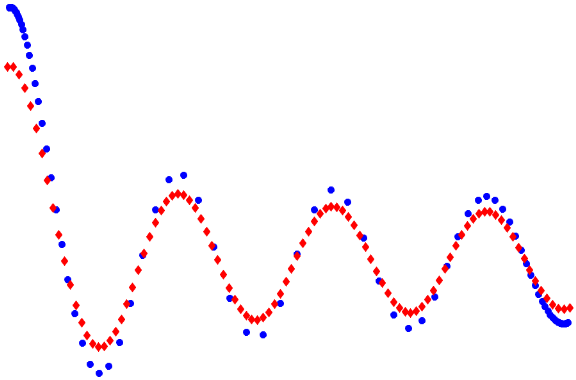
Key:

- Orange: increase mass
- Blue: reduce mass

Arbel, Zhou, G. (ICLR 2021)

Generalized Energy-Based Models - the idea

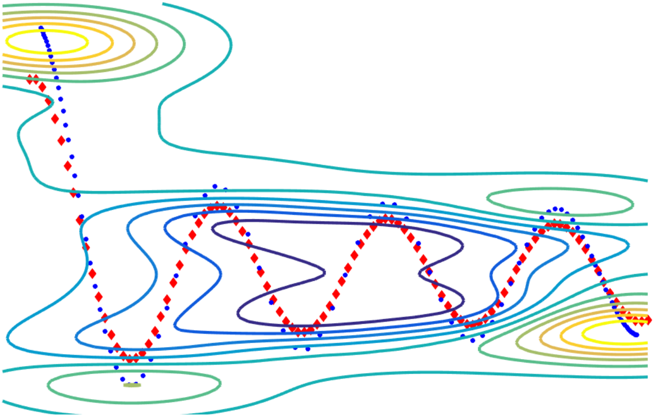
Target distribution P and GAN (generator) Q_θ, wrong support and wrong mass



Arbel, Zhou, G. (ICLR 2021)

Generalized Energy-Based Models - the idea

Log energy function, P, and Q_θ



Key:

- Orange: increase mass
- Blue: reduce mass

Arbel, Zhou, G. (ICLR 2021)

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 $\mathcal{X}$, standard energy-based model (special case)
- Sample from model via HMC on posterior of Z.

Arbel, Zhou, G. (ICLR 2021)

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 $\mathcal{X}$, standard energy-based model (special case)
- Sample from model via HMC on posterior of Z.

Arbel, Zhou, G. (ICLR 2021)

Generalized energy-based models

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

- Sample from generator with parameters θ

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

- Reweight the samples according to importance weights:

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

  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 $\mathcal{X}$, standard energy-based model (special case)
- Sample from model via HMC on posterior of Z.

Arbel, Zhou, G. (ICLR 2021)

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 $\mathcal{X}$, standard energy-based model (special case)
- Sample from model via HMC on posterior of Z.

Arbel, Zhou, G. (ICLR 2021)

How do we learn the energy E?

How do we learn the energy E?

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 (as long as E smooth)!

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

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

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

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

One last trick...(convexity of exponential)

$$-\log \int \exp(-E)\, dQ_\theta \geq -c - e^{-c} \int \exp(-E)\, dQ_\theta + 1$$

tight whenever $c = \log \int \exp(-E)\, dQ_\theta$.

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

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

One last trick...(convexity of exponential)

$$-\log \int \exp(-E)\, dQ_\theta \geq -c - e^{-c} \int \exp(-E)\, dQ_\theta + 1$$

tight whenever $c = \log \int \exp(-E)\, dQ_\theta$.

Generalized Log-Likelihood has the lower bound:

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

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

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

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

One last trick...(convexity of exponential)

$$-\log \int \exp(-E)\, dQ_\theta \geq -c - e^{-c} \int \exp(-E)\, dQ_\theta + 1$$

tight whenever $c = \log \int \exp(-E)\, dQ_\theta$.

Generalized Log-Likelihood has the lower bound:

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

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

This is the KALE! with function class $\mathcal{E} + \mathbb{R}$.

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

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

One last trick...(convexity of exponential)

$$-\log \int \exp(-E)\, dQ_\theta \geq -c - e^{-c} \int \exp(-E)\, dQ_\theta + 1$$

tight whenever $c = \log \int \exp(-E)\, dQ_\theta$.

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 $c^* = \log \int \exp(-E)\, dQ_\theta$.

Training the base measure (generator)

Recall the generator:

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

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

Training 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 $\mathcal{F}(P, Q; \mathcal{E} + \mathbb{R})$.

Training 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

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

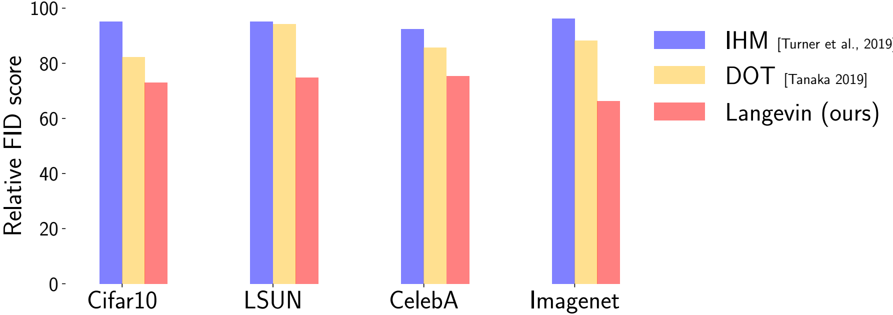
Experiments

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

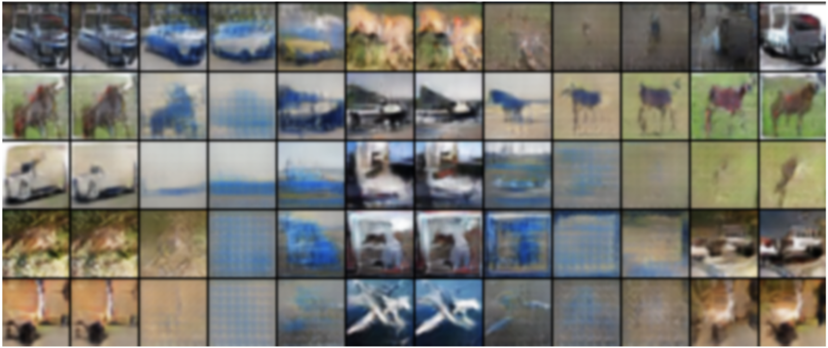
The relative FID score: $\frac{\mathrm{FID}(Q_{B_{\theta},E})}{\mathrm{FID}(B_{\theta})}$



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

Examples: moving between modes

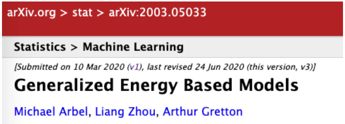
Tempered GEBM Cifar10 samples at different stages of sampling using KLA. Early samples → late samples.
Model run at *lower friction* (but still low temperature, $\beta = 100$) for mode exploration.



Summary

- Generalized energy based model:
  - End-to-end model incorporating generator and critic
  - Always better samples than generator alone.
- ICLR 2021

https://github.com/MichaelArbel/GeneralizedEBM



Summary

- **Generalized energy based model:**
  - End-to-end model incorporating generator and critic
  - Always better samples than generator alone.
- ICLR 2021

https://github.com/MichaelArbel/GeneralizedEBM

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

NeurIPS 2020:

arXiv.org > cs > arXiv:2003.06060

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

ICLR 2021:

arXiv.org > cs > arXiv:2012.00780

Computer Science > Machine Learning

[Submitted on 1 Dec 2020 (v1), last revised 5 Jun 2021 (this version, v4)]

Refining Deep Generative Models via Discriminator Gradient Flow

Abdul Fatir Ansari, Ming Liang Ang, Harold Soh

ICLR 2021:

arXiv.org > cs > arXiv:2010.00654

Computer Science > Machine Learning

[Submitted on 1 Oct 2020 (v1), last revised 9 Feb 2021 (this version, v2)]

VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models

Zhisheng Xiao, Karsten Kreis, Jan Kautz, Arash Vahdat

Questions?


THE KERNEL
BREWERY LONDON
BALTIC PORTER
7.3% ABV
THE KERNEL
EXPORT STOUT

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.