
Arthur Gretton

Gatsby Computational Neuroscience Unit, University College London

OAMLS, 2022
Training generative models

- Have: One collection of samples $\mathbf{X}$ from unknown distribution $\mathcal{P}$.
- Goal: generate samples $\mathbf{Q}$ that look like $\mathcal{P}$.

LSUN bedroom samples $\mathcal{P}$

Generated $\mathbf{Q}$, MMD GAN

Role of divergence $D(\mathcal{P}, \mathbf{Q})$?
Outline

- $\phi$-divergences ($f$-divergences) and a variational lower bound (KL)

- Generalized energy-based models
  - “Like a GAN” but incorporate critic into sample generation
  - Perform better than using generator alone

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

Integral prob. metrics

\[ D_H(P, Q) = \sup_{g \in \mathcal{H}} |\mathbb{E}_{X \sim P} g(X) - \mathbb{E}_{Y \sim Q} g(Y)| \]

\[ D_\phi(P, Q) = \int_X q(x) \phi \left( \frac{p(x)}{q(x)} \right) \, dx \]

ϕ-divergences
The Integral Probability Metrics

\[ D_H(P, Q) = \sup_{g \in \mathcal{H}} |\mathbf{E}_{X \sim P} g(X) - \mathbf{E}_{Y \sim Q} g(Y)| \]

\[ D_\phi(P, Q) = \int_X q(x) \phi \left( \frac{p(x)}{q(x)} \right) dx \]
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:

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

\[ \text{MMD} = 1.1 \]
The $\phi$-divergences

$\mathcal{H}(P, Q) = \sup_{g \in \mathcal{H}} |E_{X \sim P} g(X) - E_{Y \sim Q} g(Y)|$

$D_{\phi}(P, Q) = \int_X q(x) \phi \left( \frac{p(x)}{q(x)} \right) dx$
The $\phi$-divergences

Define the $\phi$-divergence ($f$-divergence):

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

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

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

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

$$= \int \left( \frac{p(z)}{q(z)} \right) \log \left( \frac{p(z)}{q(z)} \right) q(z) \, dz$$
The $\phi$-divergences

Define the $\phi$-divergence (f-divergence):

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

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

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

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

$$= \int \left( \frac{p(z)}{q(z)} \right) \log \left( \frac{p(z)}{q(z)} \right) q(z) dz$$
Are $\phi$-divergences good critics?

**Simple example:** disjoint support.

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

$$D_{KL}(P, Q) = \infty \quad D_{JS}(P, Q) = \log 2$$
Are $\phi$-divergences good critics?

Simple example: disjoint support.

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

\[
D_{KL}(P, Q) = \infty \quad D_{JS}(P, Q) = \log 2
\]
\textbf{\(\phi\)-divergences in practice}

**Background:** the conjugate (Fenchel) dual

\[
\phi^*(v) = \sup_{u \in \mathbb{R}} \{uv - \phi(u)\}.
\]

- \(\phi^*(v)\) is negative intercept of tangent to \(\phi\) with slope \(v\)
**ϕ-divergences in practice**

**Background:** the conjugate (Fenchel) dual

\[
\phi^*(v) = \sup_{u \in \mathbb{R}} \{uv - \phi(u)\}.
\]

- For a convex l.s.c. \( \phi \) we have

\[
\phi^{**}(x) = \phi(x) = \sup_{v \in \mathbb{R}} \{xv - \phi^*(v)\}
\]
\textbf{\(\phi\)-divergences in practice}

**Background:** the conjugate (Fenchel) dual

\[
\phi^*(v) = \sup_{u \in \mathbb{R}} \{uv - \phi(u)\}.
\]

- For a convex l.s.c. \(\phi\) we have

\[
\phi^{**}(x) = \phi(x) = \sup_{v \in \mathbb{R}} \{xv - \phi^*(v)\}
\]

- KL divergence:

\[
\phi(x) = x \log(x) \quad \phi^*(v) = \exp(v - 1)
\]
A variational lower bound

A lower-bound $\phi$-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 $\phi$-divergence approximation:

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

$$= \int q(z) \sup_{f_z} \left( \frac{p(z)}{q(z)} f_z - \phi^*(f_z) \right)$$

$\phi^*(\nu)$ is dual of $\phi(x)$. 
**A variational lower bound**

A lower-bound $\phi$-divergence approximation:

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

$$ = \int q(z) \sup_{f_z} \left( \frac{p(z)}{q(z)} f_z - \phi^*(f_z) \right) $$

$$ \geq \sup_{f \in \mathcal{H}} \mathbb{E}_P f(X) - \mathbb{E}_Q \phi^*(f(Y)) $$

(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 $\phi$-divergence approximation:

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

$$= \int q(z) \sup_{f_z} \left( \frac{p(z)}{q(z)} f_z - \phi^*(f_z) \right)$$

$$\geq \sup_{f \in \mathcal{H}} E_P f(X) - E_Q \phi^* (f(Y))$$

(restrict the function class)

Bound tight when:

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

if ratio defined.

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

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

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

\[ \phi^*(-f(Y)+1) \]

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

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

\[ \geq \sup_{f \in \mathcal{H}} -E_P f(X) + 1 - E_Q \exp(-f(Y)) \]

Bound tight when:

\[ f^\circ(z) = -\log \frac{p(z)}{q(z)} \]

if ratio defined.

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

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

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

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

\[ x_i \overset{\text{i.i.d.}}{\sim} P \]

\[ y_i \overset{\text{i.i.d.}}{\sim} Q \]

Case of the KL

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

\[ \geq \sup_{f \in \mathcal{H}} -E_P f(X) + 1 - E_Q \exp(-f(Y)) \]

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

This is a KL Approximate Lower-bound Estimator.

Case of the KL

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

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

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

This is a

KA

NE

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

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

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

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

The KALE divergence

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

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

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

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

Empirical properties of KALE

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

\[ f = \langle w, \phi(x) \rangle_{\mathcal{H}} \quad \mathcal{H} \text{ an RKHS} \]

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

**Empirical properties of KALE**

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

\[
f = \langle w, \phi(x) \rangle_{\mathcal{H}} \quad \mathcal{H} \text{ an RKHS}
\]

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

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

Empirical properties of KALE

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

\[ f = \langle w, \phi(x) \rangle_{\mathcal{H}} \quad \mathcal{H} \text{ an RKHS} \]

\[ ||w||_{\mathcal{H}}^2 \text{ penalized: KALE smoothie} \]

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

The KALE smoothie and “mode collapse”

- Two Gaussians with same means, different variance

Example thanks to M. Arbel and M. Rosca
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 $P = Q$.

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 $0 \leq c \leq C_{\text{max}}$.

**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:

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

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

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

Partial proof idea:

$$KALE(P, Q; \mathcal{H}) = -\int f \, dP - \int \exp(-f) \, dQ + 1$$

$$= \int f(x) \, dQ(x) - f(x') \, dP(x')$$

$$- \int (\exp(-f) + f - 1) \, dQ$$

$$\geq 0$$

$$\leq \int f(x) \, dQ(x) - f(x') \, dP(x') \leq LW_1(P, Q)$$

Liu, Bousquet, Chaudhuri. “Approximation and Convergence Properties of Generative Adversarial Learning” (NeurIPS 2017); Arbel, Liang, G. (ICLR 2021, Proposition 1)
How to train your GAN
Generalized Energy-Based Model
Visual notation: GAN setting
Visual notation: GAN setting
Reminder: the generator

Radford, Metz, Chintala, ICLR 2016
Energy function to improve generator: demo

Target distribution $P$

Example thanks to M. Arbel
Energy function to improve generator: demo

GAN (generator) $Q_\theta$, correct support but wrong mass

Example thanks to M. Arbel
Energy function to improve generator: demo

Log energy function and $Q_\theta$

Key:
- **Orange**: increase mass
- **Blue**: reduce mass

Example thanks to M. Arbel
Energy function to improve generator: demo

Target distribution $P$ and GAN (generator) $Q_\theta$, wrong support and wrong mass

Example thanks to M. Arbel
Energy function to improve generator: demo

Log energy function, $P$, and $Q_\theta$

Key:

- **Orange**: increase weight
- **Blue**: reduce weight

Example thanks to M. Arbel
Generalized energy-based models

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

- Sample from generator with parameters $\theta$

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

- Reweight the samples according to importance weights:

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

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

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

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

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

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

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

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

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

**Theorem:** \( \kappa \) is lipschitz and differentiable for almost all \( \theta \in \Theta \) with:

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

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 \( \Psi \) compact,
  - jointly continous w.r.t. \((\psi, x)\), \(L\)-lipschitz and \(L\)-smooth w.r.t. \(x\).
- \((\theta, z) \mapsto B_\theta(z)\) jointly continuous wrt \((\theta, z)\), \(z \mapsto B_\theta(z)\) uniformly Lipschitz w.r.t. \(z\), lipschitz and smooth wrt \(\theta\) (see paper: constants depend on \(z\))
Sampling from the model

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

$$f_{B,E}(x) := \frac{\exp(-E(x))}{Z_{Q,E}}$$
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 $g$,
\[ \int g(x) d Q_{B,E}(x) = \int g(B(z)) f_{B,E}(B(z)) \eta(z) dz \]
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 \( X = B_\theta(Z) \), \( Z \sim \eta \),

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

For a test function \( g \),

\[
\int g(x) d Q_{B,E}(x) = \int g(B(z)) f_{B,E}(B(z)) \eta(z) dz
\]

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: \( \frac{\text{FID}(Q_{B_{\theta},E})}{\text{FID}(B_{\theta})} \)

For a given generator \( B_{\theta} \) 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

**Generalized Energy Based Models**

Michael Arbel, Liang Zhou, Arthur Gretton
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
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 $B_{AI}$ is real NVP with closed-form density.