Generalized Energy-Based Models

Arthur Gretton

Gatsby Computational Neuroscience Unit,
Deepmind

Columbia Statistics, 2023
Training generative models

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

Role of divergence $D(P, Q)$?
Visual notation: GAN setting
Visual notation: GAN setting
Outline

Divergences $D(P, Q)$
- $\phi$-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

Integral prob. metrics

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

\( \Phi \)-divergences

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

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

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

Integral prob. metrics

$\phi$-divergences

- Hellinger
- KL
- Pearson chi$^2$
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]

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

**Notation:** 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**

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

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

- For a convex l.s.c. φ we have

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

\textbf{Notation:} 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:

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

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

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

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

\[
D_{\phi}(P, Q) = \int q(z) \phi \left( \frac{p(z)}{q(z)} \right) \, dz \\
= \int q(z) \sup_{f_z} \left( \frac{p(z)}{q(z)} f_z - \phi^*(f_z) \right) \\
\geq \sup_{f \in \mathcal{H}} \mathbb{E}_P f(X) - \mathbb{E}_Q \phi^*(f(Y))
\]

(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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This is a

K

A

L

E

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

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

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

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

The KALE divergence

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

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

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

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

Glaser, Arbel, G. “KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support,” (NeurIPS 2021, Section 2)
Empirical properties of KALE

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

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

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

Glaser, Arbel, G. “KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support,” (NeurIPS 2021, Section 2)
Empirical properties of KALE

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

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

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

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

Glaser, Arbel, G. “KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support,” (NeurIPS 2021, Section 2)
Empirical properties of KALE

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

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

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

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

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

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

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

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

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

$$- \int \underbrace{(\exp(-f) + f - 1)}_{\geq 0} \, dQ$$

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

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

Figure 1: DCGAN generator used for LSUN scene modeling. A 100 dimensional uniform distribution $Z$ is projected to a small spatial extent convolutional representation with many feature maps. A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions) then convert this high level representation into a $64 \times 64$ pixel image. Notably, no fully connected or pooling layers are used.

4.1 LSUN

As visual quality of samples from generative image models has improved, concerns of over-fitting and memorization of training samples have risen. To demonstrate how our model scales with more data and higher resolution generation, we train a model on the LSUN bedrooms dataset containing a little over 3 million training examples. Recent analysis has shown that there is a direct link between how fast models learn and their generalization performance (Hardt et al., 2015). We show samples from one epoch of training (Fig.2), mimicking online learning, in addition to samples after convergence (Fig.3), as an opportunity to demonstrate that our model is not producing high quality samples via simply overfitting/memorizing training examples. No data augmentation was applied to the images.

4.1.1 DEDUPLICATION

To further decrease the likelihood of the generator memorizing input examples (Fig.2) we perform a simple image de-duplication process. We fit a 3072-128-3072 de-noising dropout regularized RELU autoencoder on 32x32 downsampled center-crops of training examples. The resulting code layer activations are then binarized via thresholding the ReLU activation which has been shown to be an effective information preserving technique (Srivastava et al., 2014) and provides a convenient form of semantic-hashing, allowing for linear time de-duplication. Visual inspection of hash collisions showed high precision with an estimated false positive rate of less than 1 in 100. Additionally, the technique detected and removed approximately 275,000 near duplicates, suggesting a high recall.

4.2 FACES

We scraped images containing human faces from random web image queries of peoples names. The people names were acquired from dbpedia, with a criterion that they were born in the modern era. This dataset has 3M images from 10K people. We run an OpenCV face detector on these images, keeping the detections that are sufficiently high resolution, which gives us approximately 350,000 face boxes. We use these face boxes for training. No data augmentation was applied to the images.
Generalized Energy-Based Models - the idea

Target distribution $P$

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

GAN (generator)

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

correct support but wrong mass

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

Log energy function and $Q_\theta$

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

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

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

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

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

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

- 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 $\mathcal{X}$, 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 $Q_\theta$ **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:

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

One last trick...(convexity of exponential)

$$- \log \int \exp(-E) \, dQ \geq -c - e^{-c} \int \exp(-E) \, dQ + 1$$
tight whenever

$$c = \log \int \exp(-E) \, dQ.$$

Generalized Log-Likelihood has the lower bound:

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

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

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

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

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:

\[ X = \mathcal{B}_\theta(Z), \quad Z \sim \eta \]

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:

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

where \( \mathcal{E}^* \) 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 $X = B_\theta(Z)$, $Z \sim \eta$,

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

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

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

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

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

- **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
Research support

Work supported by:

The Gatsby Charitable Foundation

Deepmind
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 $B_\theta$ is real NVP with closed-form density.