Implicit generative models: dual vs. primal approaches

Ilya Tolstikhin
MPI for Intelligent Systems
ilya@tue.mpg.de

Machine Learning Summer School 2017
Tübingen, Germany
Contents

1. Unsupervised generative modelling and implicit models
2. Distances on probability measures
3. GAN and $f$-GAN: minimizing $f$-divergences (dual formulation)
4. WGAN: minimizing the optimal transport (dual formulation)
5. VAE: minimizing the KL-divergence (primal formulation)
6. POT: minimizing the optimal transport (primal formulation)
7. Dual vs. primal: precision vs. recall? Unifying VAE and GAN

Most importantly:

WE NEED AN ADEQUATE WAY TO EVALUATE THE MODELS
Contents

1. Unsupervised generative modelling and implicit models
2. Distances on probability measures
3. GAN and $f$-GAN: minimizing $f$-divergences (dual formulation)
4. WGAN: minimizing the optimal transport (dual formulation)
5. VAE: minimizing the KL-divergence (primal formulation)
6. POT: minimizing the optimal transport (primal formulation)
7. Dual vs. primal: precision vs. recall? Unifying VAE and GAN

Most importantly:

WE NEED AN ADEQUATE WAY TO EVALUATE THE MODELS
The task:

- There exists an unknown distribution $P_X$ over the data space $\mathcal{X}$ and we have an i.i.d. sample $X_1, \ldots, X_n$ from $P_X$.
- Find a model distribution $P_G$ over $\mathcal{X}$ similar to $P_X$.

We will work with latent variable models $P_G$ defined by 2 steps:

1. Sample a code $Z$ from the latent space $\mathcal{Z}$;
2. Map $Z$ to $G(Z) \in \mathcal{X}$ with a (random) transformation $G: \mathcal{Z} \rightarrow \mathcal{X}$.

$$p_G(x) := \int_{\mathcal{Z}} p_G(x|z)p_z(z)dx.$$ 

All techniques mentioned in this talk share two features:

- While $P_G$ has no analytical expression, it is easy to sample from;
- The objective allows for SGD training.
1. Unsupervised generative modelling and implicit models
2. Distances on probability measures
3. GAN and $f$-GAN: minimizing $f$-divergences (dual formulation)
4. WGAN: minimizing the optimal transport (dual formulation)
5. VAE: minimizing the KL-divergence (primal formulation)
6. POT: minimizing the optimal transport (primal formulation)
7. Dual vs. primal: precision vs. recall? Unifying VAE and GAN

**Most importantly:**

**WE NEED AN ADEQUATE WAY TO EVALUATE THE MODELS**
How to measure a similarity between $P_X$ and $P_G$?

- **f-divergences** Take any convex $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$.

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

- **Integral Probability Metrics**
  Take any class $\mathcal{F}$ of bounded real-valued functions on $\mathcal{X}$.

\[
\gamma_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]|
\]

- **Optimal transport** Take any cost $c(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. 

\[
W_c(P, Q) := \inf_{\Gamma \in \mathcal{P}(X \sim P, Y \sim Q)} \mathbb{E}_{(X,Y) \sim \Gamma}[c(X,Y)],
\]

where $\mathcal{P}(X \sim P, Y \sim Q)$ is a set of all joint distributions of $(X,Y)$ with marginals $P$ and $Q$ respectively.
Contents

1. Unsupervised generative modelling and implicit models
2. Distances on probability measures
3. GAN and $f$-GAN: minimizing $f$-divergences (dual formulation)
4. WGAN: minimizing the optimal transport (dual formulation)
5. VAE: minimizing the KL-divergence (primal formulation)
6. POT: minimizing the optimal transport (primal formulation)
7. Dual vs. primal: precision vs. recall? Unifying VAE and GAN

Most importantly:

WE NEED AN ADEQUATE WAY TO EVALUATE THE MODELS
The goal: minimize $D_f(P_X \| P_G)$ with respect to $P_G$

Variational (dual) representation of $f$-divergences:

$$D_f(P \| Q) = \sup_{T: X \rightarrow \text{dom}(f^*)} \mathbb{E}_{X \sim P}[T(X)] - \mathbb{E}_{Y \sim Q}[f^*(T(Y))]$$

where $f^*(x) := \sup_u x \cdot u - f(u)$ is a convex conjugate of $f$.

Solving $\inf_{P_G} D_f(P_X \| P_G)$ is equivalent to

$$\inf_{G} \sup_{T} \mathbb{E}_{X \sim P_X}[T(X)] - \mathbb{E}_{Z \sim P_Z}[f^*(T(G(Z)))]$$

(\*)

1. Estimate expectations with samples:

$$\approx \inf_{G} \sup_{T} \frac{1}{N} \sum_{i=1}^{N} T(X_i) - \frac{1}{M} \sum_{j=1}^{M} f^*(T(G(Z_j))).$$

2. Parametrize $T = T_\omega$ and $G = G_\theta$ using any flexible functions (eg. deep nets) and run SGD on (\*).
The goal: minimize $D_f(P_X \parallel P_G)$ with respect to $P_G$

**Variational (dual) representation** of $f$-divergences:

$$D_f(P \parallel Q) = \sup_{T : \mathcal{X} \to \text{dom}(f^*)} \mathbb{E}_{X \sim P}[T(X)] - \mathbb{E}_{Y \sim Q}[f^*(T(Y))]$$

where $f^*(x) := \sup_u x \cdot u - f(u)$ is a **convex conjugate** of $f$.

Solving $\inf_{P_G} D_f(P_X \parallel P_G)$ is equivalent to

$$\inf_{G} \sup_{T} \mathbb{E}_{X \sim P_X}[T(X)] - \mathbb{E}_{Z \sim P_Z}[f^*(T(G(Z)))] \quad (*)$$

1. Estimate expectations with samples:

$$\approx \inf_{G} \sup_{T} \frac{1}{N} \sum_{i=1}^{N} T(X_i) - \frac{1}{M} \sum_{j=1}^{M} f^*(T(G(Z_j))).$$

2. Parametrize $T = T_\omega$ and $G = G_\theta$ using any flexible functions (eg. deep nets) and run SGD on $(*)$. 
**Variational (dual) representation** of $f$-divergences:

\[
D_f(P_X \| P_G) = \sup_{T: \mathcal{X} \to \text{dom}(f^*)} \mathbb{E}_{X \sim P}[T(X)] - \mathbb{E}_{Z \sim P_Z}[f^*(T(G(Z)))]
\]

where $f^*(x) := \sup_u x \cdot u - f(u)$ is a convex conjugate of $f$.

1. Take $f(x) = -(x + 1) \log\frac{x+1}{2} + x \log x$ and $f^*(t) = -\log(2 - e^t)$. The domain of $f^*$ is $(-\infty, \log 2)$;
2. Take $T = g_f \circ T_\omega$, where $g_f(v) = \log 2 - \log(1 + e^{-v})$;
3. Parametrize $G = G_\theta$ and $T_\omega$ with deep nets

Up to additive $2 \log 2$ term $\inf_{P_G} D_f(P_X \| P_G)$ is equivalent to

\[
\inf_{G_\theta} \sup_{T_\omega} \mathbb{E}_{X \sim P_X} \log \frac{1}{1 + e^{-T_\omega(X)}} + \mathbb{E}_{Z \sim P_Z} \log \left(1 - \frac{1}{1 + e^{-T_\omega(G_\theta(Z))}}\right)
\]

Compare to the original GAN objective

\[
\inf_{G_\theta} \sup_{T_\omega} \mathbb{E}_{X \sim P_d}[\log T_\omega(X)] + \mathbb{E}_{Z \sim P_Z}[\log(1 - T_\omega(G_\theta(Z)))]
\]
Theory vs. practice: do we know what GANs do?

**Variational (dual) representation** of $f$-divergences:

$$D_f(P_X \| P_G) = \sup_{T: \mathcal{X} \rightarrow \text{dom}(f^*)} \mathbb{E}_{X \sim P}[T(X)] - \mathbb{E}_{Z \sim P_Z}[f^*(T(G(Z)))]$$

where $f^*(x) := \sup_u x \cdot u - f(u)$ is a convex conjugate of $f$.

$$\inf_{G_\theta} \sup_{T_\omega} \mathbb{E}_{X \sim P_d}[\log T_\omega(X)] + \mathbb{E}_{Z \sim P_Z}[\log(1 - T_\omega(G_\theta(Z)))]$$

GANs are not precisely solving $\inf_{P_G} \text{JS}(P_X \| P_G)$, because:

1. GANs replace expectations with sample averages. Uniform lows of large numbers may not apply, as our function classes are huge;

2. Instead of taking supremum over all possible witness functions $T$ GANs optimize over classes of DNNs;

3. In practice GANs never optimize $T_\omega$ “to the end” because of various computational/numerical reasons.
A possible criticism of $f$-divergences:

- When $P_X$ and $P_G$ are supported on disjoint manifolds $f$-divergences often max out.

- This leads to numerical instabilities: no useful gradients for $G$.

- Consider $P_{G'}$ and $P_{G''}$ supported on manifolds $M'$ and $M''$. Suppose $d(M', M_X) < d(M', M_X)$, where $M_X$ is the true manifold. $f$-divergences will often give the same numbers.

Possible solutions:

1. **The smoothing**: add a noise to both $P_X$ and $P_G$ before comparing.

2. Use other divergences, including IPMs and the optimal transport.
Minimizing MMD between $P_X$ and $P_G$

- Take any reproducing kernel $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Let $\mathcal{B}_k$ be a unit ball of the corresponding RKHS $\mathcal{H}_k$.

- Maximum Mean Discrepancy is the following IPM:

$$\gamma_k(P_X, P_G) := \sup_{T \in \mathcal{B}_k} |\mathbb{E}_{P_X}[T(X)] - \mathbb{E}_{P_G}[T(Y)]|$$

(MMD)

- This optimization problem has a closed form analytical solution.

One can play the adversarial game using (MMD) instead of $D_f(P_X \parallel P_G)$:

- No need to train the discriminator $T$;
- On the other hand, $\mathcal{B}_k$ is a rather restricted class;
- One can also train $k$ adversarially, resulting in a stronger objective:

$$\inf_{P_G} \max_k \gamma_k(P_X, P_G).$$
1. Unsupervised generative modelling and implicit models
2. Distances on probability measures
3. GAN and $f$-GAN: minimizing $f$-divergences (dual formulation)
4. WGAN: minimizing the optimal transport (dual formulation)
5. VAE: minimizing the KL-divergence (primal formulation)
6. POT: minimizing the optimal transport (primal formulation)
7. Dual vs. primal: precision vs. recall? Unifying VAE and GAN

Most importantly:

WE NEED AN ADEQUATE WAY TO EVALUATE THE MODELS
1-Wasserstein distance is defined by

\[ W_1(P, Q) := \inf_{\Gamma \in \mathcal{P}(X \sim P, Y \sim Q)} \mathbb{E}_{(X,Y) \sim \Gamma}[d(X,Y)], \]

where \( \mathcal{P}(X \sim P, Y \sim Q) \) is a set of all joint distributions of \((X, Y)\) with marginals \(P\) and \(Q\) respectively and \((\mathcal{X}, d)\) is a metric space.

Kantorovich-Rubinstein duality:

\[ W_1(P, Q) = \sup_{T \in \mathcal{F}_L} |\mathbb{E}_{P_X}[T(X)] - \mathbb{E}_{P_G}[T(Y)]|, \tag{KR} \]

where \( \mathcal{F}_L \) are all the bounded 1-Lipschitz functions on \((\mathcal{X}, d)\).

WGAN: In order to solve \( \inf_{P_G} W_1(P_X, P_G) \) let’s play the adversarial training card on (KR). Parametrize \( T = T_\omega \) using the weight clipping or perform the gradient penalization.

Unfortunately, (KR) holds only for the 1-Wasserstein distance.
Contents

1. Unsupervised generative modelling and implicit models
2. Distances on probability measures
3. GAN and $f$-GAN: minimizing $f$-divergences (dual formulation)
4. WGAN: minimizing the optimal transport (dual formulation)
5. VAE: minimizing the KL-divergence (primal formulation)
6. POT: minimizing the optimal transport (primal formulation)
7. Dual vs. primal: precision vs. recall? Unifying VAE and GAN

Most importantly:

WE NEED AN ADEQUATE WAY TO EVALUATE THE MODELS
VAE: Maximizing the marginal log-likelihood

\[
\inf_{P_G} \text{KL}(P_X \| P_G) \iff \inf_{P_G} - \mathbb{E}_{P_X}[\log p_G(X)].
\]

**Variational upper bound:** for any conditional distribution \(Q(Z|X)\)

\[-\mathbb{E}_{P_X}[\log p_G(X)] = \mathbb{E}_{P_X}\left[\text{KL}(Q(Z|X), P_Z) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right]
- \mathbb{E}_{P_X}\left[\text{KL}(Q(Z|X), P_G(Z|X))\right]
\leq \mathbb{E}_{P_X}\left[\text{KL}(Q(Z|X), P_Z) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right].\]

In particular, if \(Q\) is not restricted:

\[-\mathbb{E}_{P_X}[\log p_G(X)] = \inf_Q \mathbb{E}_{P_X}\left[\text{KL}(Q(Z|X), P_Z) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right].\]

**Variational Auto-Encoders** use the upper bound and

- Latent variable models with any \(P_G(X|Z)\), eg. \(\mathcal{N}(X; G(Z), \sigma^2 \cdot I)\)
- Set \(P_Z(Z) = \mathcal{N}(Z; 0, I)\) and \(Q(Z|X) = \mathcal{N}(Z; \mu(X), \Sigma(X))\)
- Parametrize \(G = G_\theta, \mu, \) and \(\Sigma\) with deep nets. Run SGD.
AVB: reducing the gap in the upper bound

**Variational upper bound:**

\[-E_{P_X}[\log p_G(X)] \leq \inf_{Q \in Q} E_{P_X}[KL(Q(Z|X), P_Z) - E_{Q(Z|X)}[\log p_G(X|Z)]]\]

**Adversarial Variational Bayes** reduces the variational gap by

- Allowing for flexible encoders $Q_e(Z|X)$, defined implicitly by random variables $e(X, \epsilon)$, where $\epsilon \sim P_\epsilon$;
- Replacing the KL divergence in the objective by the adversarial approximation (any of the ones discussed above);
- Parametrize $e$ with a deep net. Run SGD.

**Downsides** of VAE and AVB:

- Literature reports blurry samples. This is caused by the combination of KL objective and the Gaussian decoder.
- Importantly, $P_G(X|Z)$ is trained only for encoded training points, i.e. for $Z \sim Q(Z|X)$ and $X \sim P_X$. But we sample from $Z \sim P_Z$. 
Unregularized Auto-Encoders

Variational upper bound:

\[-\mathbb{E}_{P_X}[\log p_G(X)] \leq \inf_{Q \in \mathcal{Q}} \mathbb{E}_{P_X}[\text{KL}(Q(Z|X), P_Z) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]]\]

- The KL term in the upper bound may be viewed as a regularizer;
- Dropping it results in classical auto-encoders, where the encoder-decoder pair tries to reconstruct all training images;
- In this case training images $X$ often end up being mapped to different spots chaotically scattered in the $Z$ space;
- As a result, $Z$ captures no useful representations. Sampling is hard.
Contents

1. Unsupervised generative modelling and implicit models
2. Distances on probability measures
3. GAN and $f$-GAN: minimizing $f$-divergences (dual formulation)
4. WGAN: minimizing the optimal transport (dual formulation)
5. VAE: minimizing the KL-divergence (primal formulation)
6. POT: minimizing the optimal transport (primal formulation)
7. Dual vs. primal: precision vs. recall? Unifying VAE and GAN

Most importantly:

WE NEED AN ADEQUATE WAY TO EVALUATE THE MODELS
Minimizing the optimal transport

**Optimal transport** for a cost function \( c(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \) is

\[
W_c(P_X, P_G) := \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_G)} \mathbb{E}_{(X,Y) \sim \Gamma}[c(X,Y)],
\]

If \( P_G(Y|Z = z) = \delta_{G(z)} \) for all \( z \in \mathcal{Z} \), where \( G : \mathcal{Z} \to \mathcal{X} \), we have

\[
W_c(P_X, P_G) = \inf_{Q : Q_Z = P_Z} \mathbb{E}_{P_X} \mathbb{E}_{Q(Z|X)}[c(X, G(Z))],
\]

where \( Q_Z \) is the marginal distribution of \( Z \) when \( X \sim P_X, Z \sim Q(Z|X) \).
Minimizing the optimal transport

Optimal transport for a cost function \( c(x, y): \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \) is

\[
W_c(P_X, P_G) := \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_G)} \mathbb{E}_{(X,Y) \sim \Gamma}[c(X,Y)],
\]

If \( P_G(Y|Z = z) = \delta_{G(z)} \) for all \( z \in \mathcal{Z} \), where \( G: \mathcal{Z} \to \mathcal{X} \), we have

\[
W_c(P_X, P_G) = \inf_{Q: \mathbb{Q}_Z = \mathbb{P}_Z} \mathbb{E}_{P_X} \mathbb{E}_{\mathbb{Q}(Z|X)}[c(X, G(Z))],
\]

where \( \mathbb{Q}_Z \) is the marginal distribution of \( Z \) when \( X \sim P_X, Z \sim Q(Z|X) \).
Relaxing the constraint

\[
W_c(P_X, P_G) = \inf_{Q: Q_Z = P_Z} \mathbb{E}_{P_X} \mathbb{E}_{Q(Z|X)} [c(X, G(Z))],
\]

Penalized Optimal Transport replaces the constraint with a penalty:

\[
\text{POT}(P_X, P_G) := \inf_{Q} \mathbb{E}_{P_X} \mathbb{E}_{Q(Z|X)} [c(X, G(Z))] + \lambda \cdot D(Q_Z, P_Z)
\]

and uses the adversarial training in the \( Z \) space to approximate \( D \).

- For the 2-Wasserstein distance \( c(X, Y) = \|X - Y\|_2^2 \) POT recovers Adversarial Auto-Encoders;

- For the 1-Wasserstein distance \( c(X, Y) = \|X - Y\|^2 \) POT and WGAN are solving the same problem from the primal and dual forms respectively.

- Importantly, unlike VAE, POT does not force \( Q(Z|X = x) \) to intersect for different \( x \), which is known to lead to the blurriness.
Contents

1. Unsupervised generative modelling and implicit models
2. Distances on probability measures
3. GAN and $f$-GAN: minimizing $f$-divergences (dual formulation)
4. WGAN: minimizing the optimal transport (dual formulation)
5. VAE: minimizing the KL-divergence (primal formulation)
6. POT: minimizing the optimal transport (primal formulation)
7. Dual vs. primal: precision vs. recall? Unifying VAE and GAN

Most importantly:

WE NEED AN ADEQUATE WAY TO EVALUATE THE MODELS
GANs approach the problem from a dual perspective.

They are known to produce very sharply looking images.

$$\max_G \mathbb{E}_{Z \sim P_Z} [T^*(G(Z))]$$

But notoriously hard to train, unstable (although many would disagree), and sometimes lead to mode collapses.

GANs come without an encoder.
(Gulrajani et al., 2017) aka Improved WGAN, 32X32 CIFAR-10
(Radford et al., 2015) aka DCGAN, 64x64 LSUN
- VAEs approach the problem from its primal.
- They enjoy a **very stable training** and often lead to **diverse samples**.

$$\max_G \mathbb{E}_{X \sim P_X} \mathbb{E}_{Z \sim Q(Z|X)} [c(X, G(Z))]$$

- But the **samples look blurry**
- VAEs come with **encoders**.

Various papers are trying to combine a stability and recall of VAEs with the precision of GANs:

- Choose an adversarially trained cost function $c$;
- Combine AE costs with the GAN criteria;
- ...
(Mescheder et al., 2017) aka AVB, CelebA
VAE trained on CIFAR-10, $\mathcal{Z}$ of 20 dim.
(Bousquet et al., 2017) aka POT, CIFAR-10, same architecture
(Bousquet et al., 2017) aka POT, CIFAR-10, test reconstruction
Contents

1. Unsupervised generative modelling and implicit models
2. Distances on probability measures
3. GAN and $f$-GAN: minimizing $f$-divergences (dual formulation)
4. WGAN: minimizing the optimal transport (dual formulation)
5. VAE: minimizing the KL-divergence (primal formulation)
6. POT: minimizing the optimal transport (primal formulation)
7. Dual vs. primal: precision vs. recall? Unifying VAE and GAN

Most importantly:

WE NEED AN ADEQUATE WAY TO EVALUATE THE MODELS
Literature


