Towards Principled Methods for Training Generative Adversarial Networks

Martin Arjovsky & Léon Bottou
Unsupervised learning

- We have samples $\{x^{(i)}\}_{i=1}^{m}$ from an unknown distribution $\mathbb{P}_r$
Unsupervised learning

- We have samples $\{x^{(i)}\}_{i=1}^m$ from an unknown distribution $\mathbb{P}_r$

- We want to approximate it by $\mathbb{P}_\theta$ a parametric distribution that’s close to $\mathbb{P}_r$ in some sense.
Unsupervised learning

- We have samples \( \{x^{(i)}\}_{i=1}^m \) from an unknown distribution \( \mathbb{P}_r \)

- We want to approximate it by \( \mathbb{P}_\theta \) a parametric distribution that’s close to \( \mathbb{P}_r \) in some sense.

- Close how?
Maximum Likelihood

- Maximum likelihood:

$$\max_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log P_{\theta}(x^{(i)})$$
Maximum Likelihood

- Maximum likelihood:

\[
\max_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log P_{\theta}(x^{(i)})
\]

- Assumptions: continuous with full support.
Maximum Likelihood

- Maximum likelihood:

\[
\max_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log P_\theta(x^{(i)})
\]

- Assumptions: continuous with full support.

- Problems: restricted capacity distributes mass.
  Modeling low dimensional distributions is impossible.
Kullback-Leibler Divergence

- Closeness measured by KL divergence (equivalent to ML):

\[
\min_{\theta \in \mathbb{R}^d} KL(P_r || P_\theta) = \int_X P_r(x) \log \frac{P_r(x)}{P_\theta(x)} \, dx
\]
Kullback-Leibler Divergence

- Closeness measured by KL divergence (equivalent to ML):

\[
\min_{\theta \in \mathbb{R}^d} KL(\mathbb{P}_r \| \mathbb{P}_\theta) = \int x \ P_r(x) \log \frac{P_r(x)}{P_\theta(x)} \ dx
\]

- When \( P_r(x) > 0, P_\theta(x) \to 0 \) integrand goes to infinity: high cost for mode dropping.
Kullback-Leibler Divergence

- Closeness measured by KL divergence (equivalent to ML):

\[
\min_{\theta \in \mathbb{R}^d} KL(P_r || P_\theta) = \int_{\mathcal{X}} P_r(x) \log \frac{P_r(x)}{P_\theta(x)} \, dx
\]

- When \( P_r(x) > 0, P_\theta(x) \to 0 \) integrand goes to infinity: high cost for mode dropping.

- When \( P_\theta(x) > 0, P_r(x) \to 0 \) integrand goes to 0: low cost for fake looking samples.
Generative Adversarial Networks (Goodfellow et al.)

- Let $\mathbb{P}_\theta$ be the dist of $g_\theta(Z)$ for some simple (e.g. Gaussian) r.v $Z$, passed through a complex function.
Generative Adversarial Networks (Goodfellow et al.)

- Let $\mathbb{P}_\theta$ be the dist of $g_\theta(Z)$ for some simple (e.g. Gaussian) r.v $Z$, passed through a complex function.

- Discriminator maximizes and generator minimizes

$$L(D, \theta) = \mathbb{E}_{x \sim \mathbb{P}_r} [\log D(x)] + \mathbb{E}_{z \sim \mathbb{P}_Z} [\log (1 - D(g_\theta(z)))]$$
Generative Adversarial Networks (Goodfellow et al.)

- Let $\mathbb{P}_\theta$ be the dist of $g_\theta(Z)$ for some simple (e.g. Gaussian) r.v Z, passed through a complex function.

- Discriminator maximizes and generator minimizes

$$L(D, \theta) = \mathbb{E}_{x \sim \mathbb{P}_r} [\log D(x)] + \mathbb{E}_{z \sim \mathbb{P}_Z} [\log (1 - D(g_\theta(z)))]$$

$$JSD(\mathbb{P}_r || \mathbb{P}_\theta) = \max_D \frac{1}{2} L(D, \theta) + \log 2$$
JSD seems maxed out..
Generative Adversarial Networks

- Under optimal discriminator, minimizes

$$\min_{\theta \in \mathbb{R}^d} JSD(P_r || P_\theta) = KL(P_r || P_m) + KL(P_\theta || P_m)$$

- Problems: vanishing gradients very quickly when D’s accuracy is high.
Discriminator is pretty good...
Vanishing gradients, original cost
Alternate update

- Alternate update that has less vanishing gradients

\[ \Delta \theta \propto \mathbb{E}_{z \sim p_Z} [\nabla_\theta \log(D_\phi(g_\theta(z)))] \]
Alternate update

- Alternate update that has less vanishing gradients

\[ \Delta \theta \propto \mathbb{E}_{z \sim p_z}[\nabla \theta \log(D_\phi(g_\theta(z)))] \]

- Under optimality optimizes

\[ KL(\mathbb{P}_\theta \| \mathbb{P}_r) - 2JS(D(\mathbb{P}_r \| \mathbb{P}_\theta)) \]
Alternate update

- Alternate update that has less vanishing gradients

\[ \Delta \theta \propto \mathbb{E}_{z \sim p_z} [\nabla_\theta \log(D_\phi(g_\theta(z)))] \]

- Under optimality optimizes

\[ KL(P_\theta \| P_r) - 2JSD(P_r \| P_\theta) \]

- Problems: JSD with the wrong sign, reverse KL has high mode dropping. Still unstable when D is good.
High variance updates
Problems of GANs (and divergences)

- When $\mathcal{P}_r$ and $\mathcal{P}_\theta$ lie on low dimensional manifolds, there’s always a perfect discriminator, that provides no usable gradients.
Manifold picture

- Real
- Generated
Problems of GANs (and divergences)

- When $\mathbb{P}_r$ and $\mathbb{P}_\theta$ lie on low dimensional manifolds, there's always a perfect discriminator, that provides no usable gradients.

**Theorem 2.2.** Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions that have support contained in two closed manifolds $\mathcal{M}$ and $\mathcal{P}$ that don't perfectly align and don't have full dimension. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are continuous in their respective manifolds, meaning that if there is a set $A$ with measure 0 in $\mathcal{M}$, then $\mathbb{P}_r(A) = 0$ (and analogously for $\mathbb{P}_g$). Then, there exists an optimal discriminator $D^* : \mathcal{X} \to [0, 1]$ that has accuracy 1 and for almost any $x$ in $\mathcal{M}$ or $\mathcal{P}$, $D^*$ is smooth in a neighbourhood of $x$ and $\nabla_x D^*(x) = 0$. 
Problems of GANs (and divergences)

- When $\mathbb{P}_r$ and $\mathbb{P}_\theta$ lie on low dimensional manifolds, there’s always a perfect discriminator, that provides no usable gradients.
Problems of GANs (and divergences)

- When $\mathbb{P}_r$ and $\mathbb{P}_\theta$ lie on low dimensional manifolds, there’s always a perfect discriminator, that provides no usable gradients.

- Under the same assumptions

$$\text{JSD}(\mathbb{P}_r \| \mathbb{P}_\theta) = \log 2$$
$$KL(\mathbb{P}_r \| \mathbb{P}_\theta) = +\infty$$
$$KL(\mathbb{P}_\theta \| \mathbb{P}_r) = +\infty$$
A first step to a solution

- Distributions are essentially disjoint
A first step to a solution

- Distributions are essentially disjoint
- Add noise during training to make them overlap!
A first step to a solution

- Distributions are essentially disjoint
- Add noise during training to make them overlap!
- Matching noisy distributions amounts to matching the underlying ones.
Manifold picture

- Real
- Generated
Manifold picture with noise

- Real
- Generated
A first step to a solution

**Theorem 3.2.** Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions with support on $\mathcal{M}$ and $\mathcal{P}$ respectively, with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then, the gradient passed to the generator has the form

$$E_{z \sim p(z)} \left[ \nabla_\theta \log(1 - D^*(g_\theta(z))) \right]$$

1. 

$$= E_{z \sim p(z)} \left[ a(z) \int_{\mathcal{M}} P_\epsilon(g_\theta(z) - y) \nabla_\theta \| g_\theta(z) - y \|^2 \, d\mathbb{P}_r(y) ight]$$

2. 

$$- b(z) \int_{\mathcal{P}} P_\epsilon(g_\theta(z) - y) \nabla_\theta \| g_\theta(z) - y \|^2 \, d\mathbb{P}_g(y)$$

We move our samples $g_\theta(z)$ towards point in the data manifold, weighted by their probability and distance to our samples.
Theoretical guarantee

**Theorem 3.3.** Let $P_r$ and $P_g$ be any two distributions, and $\epsilon$ be a random vector with mean 0 and variance $V$. If $P_{r+\epsilon}$ and $P_{g+\epsilon}$ have support contained on a ball of diameter $C$, then

$$W(P_r, P_g) \leq 2V^{\frac{1}{2}} + 2C \sqrt{JSD(P_{r+\epsilon} || P_{g+\epsilon})}$$

(6)
Theoretical guarantee

**Theorem 3.3.** Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be any two distributions, and $\epsilon$ be a random vector with mean 0 and variance $V$. If $\mathbb{P}_{r+\epsilon}$ and $\mathbb{P}_{g+\epsilon}$ have support contained on a ball of diameter $C$, then

$$W(\mathbb{P}_r, \mathbb{P}_g) \leq 2V^{\frac{1}{2}} + 2C \sqrt{\text{JSD}(\mathbb{P}_{r+\epsilon} \parallel \mathbb{P}_{g+\epsilon})}$$

(6)

- Wasserstein is well defined in the manifold setting.
Theoretical guarantee

**Theorem 3.3.** Let $P_r$ and $P_g$ be any two distributions, and $\epsilon$ be a random vector with mean 0 and variance $V$. If $P_{r+\epsilon}$ and $P_{g+\epsilon}$ have support contained on a ball of diameter $C$, then

$$W(P_r, P_g) \leq 2V^{\frac{1}{2}} + 2C \sqrt{JSD(P_{r+\epsilon} \parallel P_{g+\epsilon})}$$

- Wasserstein is well defined in the manifold setting.
- The noise method optimizes an upper bound of it.
Theoretical guarantee

**Theorem 3.3.** Let $P_r$ and $P_g$ be any two distributions, and $\epsilon$ be a random vector with mean 0 and variance $V$. If $P_{r+\epsilon}$ and $P_{g+\epsilon}$ have support contained on a ball of diameter $C$, then

$$W(P_r, P_g) \leq 2V^{\frac{1}{2}} + 2C \sqrt{JSD(P_{r+\epsilon} \| P_{g+\epsilon})}$$

- Wasserstein is well defined in the manifold setting.
- The noise method optimizes an upper bound of it.
- We can reduce the first summand by annealing the noise, the second one by optimizing with noise.
Loads of work done since then!

- Now we have more understanding of the relationship between Wasserstein, JSD and the rest: Weak vs strong.
Loads of work done since then!

- Now we have more understanding of the relationship between Wasserstein, JSD and the rest: Weak vs strong.
- Optimizing an approximation of Wasserstein directly is doable. (Arjovsky, Chintala & Bottou, 2017)
Loads of work done since then!

- Now we have more understanding of the relationship between Wasserstein, JSD and the rest: Weak vs strong.
- Optimizing an approximation of Wasserstein directly is doable. (Arjovsky, Chintala & Bottou, 2017)
- Different ways to do this. (Gulrajani et al. 2017)
Loads of work done since then!

- Now we have more understanding of the relationship between Wasserstein, JSD and the rest: Weak vs strong.
- Optimizing an approximation of Wasserstein directly is doable. (Arjovsky, Chintala & Bottou, 2017)
- Different ways to do this. (Gulrajani et al. 2017)
- Time to scale up!
That's all Folks!