Deep Learning

Vazgen Mikayelyan

December 22, 2020

FAST DISCOVERING THE FUTURE

WGANs

Let $\mathcal{X}$ be a compact metric set (e.g $[0,1]^d$) and let Σ denote the set of all Borel subsets of $\mathcal{X}$.

WGANs

Let $\mathcal{X}$ be a compact metric set (e.g $[0,1]^d$) and let Σ denote the set of all Borel subsets of $\mathcal{X}$.

- Total Variation (TV) distance

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)|.$$

WGANs

Let $\mathcal{X}$ be a compact metric set (e.g $[0,1]^d$) and let Σ denote the set of all Borel subsets of $\mathcal{X}$.

- Total Variation (TV) distance

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)|.$$

- Kullback-Leibler (KL) divergence

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int_{\mathbb{R}} \log\left(\frac{p_r(x)}{p_g(x)}\right) p_r(x)\, d\mu(x).$$

WGANs

Let $\mathcal{X}$ be a compact metric set (e.g $[0,1]^d$) and let Σ denote the set of all Borel subsets of $\mathcal{X}$.

- Total Variation (TV) distance

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)|.$$

- Kullback-Leibler (KL) divergence

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int_{\mathbb{R}} \log\left(\frac{p_r(x)}{p_g(x)}\right) p_r(x)\, d\mu(x).$$

- Jensen-Shannon (JS) distance

$$JS(\mathbb{P}_r \| \mathbb{P}_g) = \frac{1}{2}\left(KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m)\right)$$

where $\mathbb{P}_m = \dfrac{\mathbb{P}_r + \mathbb{P}_g}{2}$.

- The Earth-Mover (EM) distance or Wasserstein-1

$$W\left(\mathbb{P}_r, \mathbb{P}_g\right) = \inf_{\gamma \in \Pi\left(\mathbb{P}_r, \mathbb{P}_g\right)} \mathbb{E}_{(x,y)\sim\gamma}\left(\|x - y\|\right)$$

where $\Pi\left(\mathbb{P}_r, \mathbb{P}_g\right)$ denotes the set of all joint distributions $\gamma\left(x, y\right)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$.

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \infty, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \infty, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $JS(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \log 2, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \infty, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $JS(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \log 2, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$

WGANs

Note that

WGANs

Note that

- All distances other than EM are not continuous.

Note that

- All distances other than EM are not continuous.
- When $\theta_t \to 0$, the sequence $(\mathbb{P}_{\theta_t})_{t \in \mathbb{N}}$ converges to $\mathbb{P}_0$ under the EM distance, but does not converge at all under either the JS, KL, reverse KL or TV divergences.

WGANs

Note that

- All distances other than EM are not continuous.
- When $\theta_t \to 0$, the sequence $(\mathbb{P}_{\theta_t})_{t \in \mathbb{N}}$ converges to $\mathbb{P}_0$ under the EM distance, but does not converge at all under either the JS, KL, reverse KL or TV divergences.
- Only EM distance has informative gradient.

Lipschitz functions

Definition 1

Let X and Y are normed vector spaces. A function $f : X \to Y$ is called

- *K-Lipschitz if there exists a real constant $K > 0$ such that, for all x_1 and x_2 in X*

$$\|f(x_1) - f(x_2)\| \leq K\|x_1 - x_2\|.$$

- *local Lipschitz if for every $x \in X$ there exists a neighbourhood U of x such that f is Lipschitz on U.*

Lipschitz functions

Definition 1

Let X and Y are normed vector spaces. A function $f : X \to Y$ is called

- *K-Lipschitz if there exists a real constant $K > 0$ such that, for all x_1 and x_2 in X*

$$\|f(x_1) - f(x_2)\| \leq K\|x_1 - x_2\|.$$

- *local Lipschitz if for every $x \in X$ there exists a neighbourhood U of x such that f is Lipschitz on U.*

Theorem 1

If function $f : \mathbb{R}^n \to \mathbb{R}$ has bounded gradient, then f is a Lipschitz function.

EM properties

Theorem 1. *Let $\mathbb{P}_r$ be a fixed distribution over $\mathcal{X}$. Let Z be a random variable (e.g Gaussian) over another space $\mathcal{Z}$. Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a function, that will be denoted $g_\theta(z)$ with z the first coordinate and θ the second. Let $\mathbb{P}_\theta$ denote the distribution of $g_\theta(Z)$. Then,*

1. *If g is continuous in θ, so is $W(\mathbb{P}_r, \mathbb{P}_\theta)$.*
2. *If g is locally Lipschitz and satisfies regularity assumption 1, then $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere, and differentiable almost everywhere.*
3. *Statements 1-2 are false for the Jensen-Shannon divergence $JS(\mathbb{P}_r, \mathbb{P}_\theta)$ and all the KLs.*

EM properties

Theorem 1. *Let $\mathbb{P}_r$ be a fixed distribution over $\mathcal{X}$. Let Z be a random variable (e.g Gaussian) over another space $\mathcal{Z}$. Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a function, that will be denoted $g_\theta(z)$ with z the first coordinate and θ the second. Let $\mathbb{P}_\theta$ denote the distribution of $g_\theta(Z)$. Then,*

1. *If g is continuous in θ, so is $W(\mathbb{P}_r, \mathbb{P}_\theta)$.*
2. *If g is locally Lipschitz and satisfies regularity assumption 1, then $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere, and differentiable almost everywhere.*
3. *Statements 1-2 are false for the Jensen-Shannon divergence $JS(\mathbb{P}_r, \mathbb{P}_\theta)$ and all the KLs.*

Corollary 1. *Let g_θ be any feedforward neural network[4] parameterized by θ, and $p(z)$ a prior over z such that $\mathbb{E}_{z \sim p(z)}[\|z\|] < \infty$ (e.g. Gaussian, uniform, etc.). Then assumption 1 is satisfied and therefore $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere and differentiable almost everywhere.*

EM properties

Theorem 2. *Let $\mathbb{P}$ be a distribution on a compact space $\mathcal{X}$ and $(\mathbb{P}_n)_{n\in\mathbb{N}}$ be a sequence of distributions on $\mathcal{X}$. Then, considering all limits as $n \to \infty$,*

1. *The following statements are equivalent*
   - *$\delta(\mathbb{P}_n, \mathbb{P}) \to 0$ with δ the total variation distance.*
   - *$JS(\mathbb{P}_n, \mathbb{P}) \to 0$ with JS the Jensen-Shannon divergence.*
2. *The following statements are equivalent*
   - *$W(\mathbb{P}_n, \mathbb{P}) \to 0$.*
   - *$\mathbb{P}_n \xrightarrow{\mathcal{D}} \mathbb{P}$ where $\xrightarrow{\mathcal{D}}$ represents convergence in distribution for random variables.*
3. *$KL(\mathbb{P}_n \| \mathbb{P}) \to 0$ or $KL(\mathbb{P} \| \mathbb{P}_n) \to 0$ imply the statements in (1).*
4. *The statements in (1) imply the statements in (2).*

EM properties

Summary

Wasserstein (or EM) loss for neural networks is continuous and differentiable almost everywhere. Moreover, Convergence in KL implies convergence in TV and JS which implies convergence in EM.

Training

Kantorovich-Rubinstein duality

$$W\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{x \sim \mathbb{P}_r}\left[f\left(x\right)\right] - \mathbb{E}_{x \sim \mathbb{P}_\theta}\left[f\left(x\right)\right]\right)$$

where the supremum is over all the 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$.

Training

Kantorovich-Rubinstein duality

$$W\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{x \sim \mathbb{P}_r}\left[f\left(x\right)\right] - \mathbb{E}_{x \sim \mathbb{P}_\theta}\left[f\left(x\right)\right]\right)$$

where the supremum is over all the 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$.

Note that if we replace $\|f\|_L \leq 1$ for $\|f\|_L \leq K$ (consider K-Lipschitz for some constant K) then we end up with $K \cdot W\left(\mathbb{P}_r, \mathbb{P}_g\right)$.

Training

Kantorovich-Rubinstein duality

$$W\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{x \sim \mathbb{P}_r}\left[f\left(x\right)\right] - \mathbb{E}_{x \sim \mathbb{P}_\theta}\left[f\left(x\right)\right]\right)$$

where the supremum is over all the 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$.

Note that if we replace $\|f\|_L \leq 1$ for $\|f\|_L \leq K$ (consider K-Lipschitz for some constant K) then we end up with $K \cdot W\left(\mathbb{P}_r, \mathbb{P}_g\right)$.
Therefore, if we have a parameterized family of functions $\{f_w\}_{w \in \mathcal{W}}$ that are all K-Lipschitz for some K, we could consider solving the problem

$$\max_{w \in \mathcal{W}} \left(\mathbb{E}_{x \sim \mathbb{P}_r}\left[f_w\left(x\right)\right] - \mathbb{E}_{z \sim p(z)}\left[f_w\left(g_\theta\left(z\right)\right)\right]\right)$$

for estimating $W\left(\mathbb{P}_r, \mathbb{P}_\theta\right)$.

Training

Theorem 3. *Let $\mathbb{P}_r$ be any distribution. Let $\mathbb{P}_\theta$ be the distribution of $g_\theta(Z)$ with Z a random variable with density p and g_θ a function satisfying assumption 1. Then, there is a solution $f : \mathcal{X} \to \mathbb{R}$ to the problem*

$$\max_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

and we have

$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) = -\mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))]$$

when both terms are well-defined.

Training

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.

Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

while θ has not converged **do**
  for $t = 0, ..., n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end for
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while

Training

Note that

Training

Note that

- By clipping the weight parameters in the critic we force critic to have bounded gradients,

Training

Note that

- By clipping the weight parameters in the critic we force critic to have bounded gradients,
- we can train the critic till optimality,

Training

Note that

- By clipping the weight parameters in the critic we force critic to have bounded gradients,
- we can train the critic till optimality,
- if the clipping parameter is large, then it can take a long time for any weights to reach their limit, thereby making it harder to train the critic till optimality. If the clipping is small, this can easily lead to vanishing gradients when the number of layers is big.

WGAN Results

- A meaningful loss metric that correlates with the generator's convergence and sample quality,

WGAN Results

- A meaningful loss metric that correlates with the generator's convergence and sample quality,
- improved stability of the optimization process,

WGAN Results

- A meaningful loss metric that correlates with the generator's convergence and sample quality,
- improved stability of the optimization process,
- results of WGAN are better than results of DCGAN (with the same generator and discriminator),

WGAN Results

- A meaningful loss metric that correlates with the generator's convergence and sample quality,
- improved stability of the optimization process,
- results of WGAN are better than results of DCGAN (with the same generator and discriminator),
- does not work well with momentum-based optimizers e.g. Adam.

WGAN Results

- A meaningful loss metric that correlates with the generator's convergence and sample quality,
- improved stability of the optimization process,
- results of WGAN are better than results of DCGAN (with the same generator and discriminator),
- does not work well with momentum-based optimizers e.g. Adam.
- slower to converge than KL loss.