Generative Adversarial Nets

Instructor: John Thickstun

Discussion Board: Available on Ed

Zoom Link: Available on Canvas

Instructor Contact: thickstn@cs.washington.edu

Course Webpage: https://courses.cs.washington.edu/courses/cse599i/20au/

Latent Variable Models

- Given finite samples $x_1, \ldots, x_n \sim p$, and unlimited samples $z \sim r$.

Variational Autoencoder

- Generative latent variable model:
  1. $z \sim r$,
  2. $x \sim p_\theta(\cdot|z)$.
- Learn the marginal defined by

$$p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x|z) r(z)\, dz.$$

Generative Adversarial Net

- Generative latent variable model
  1. $z \sim r$,
  2. $x = g_\theta(z) \sim p_\theta(x)$.
- Learn the pushforward defined by:

$$p_\theta(x) = r(g_\theta^{-1}(x)) |\nabla_x g_\theta^{-1}(x)|.$$

Maximize the Likelihood?

- Generative latent variable model:

$$1. \quad z \sim r,$$
$$2. \quad x = g_\theta(z) \sim p_\theta(x).$$

- The maximum likelihood estimator:

$$\hat{\theta}_{\text{mle}} \equiv \arg\max_\theta \mathop{\mathbb{E}}_{x \sim p} \log p_\theta(x) \approx \arg\max_\theta \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i), \text{ where } x_i \sim p.$$

- The log-likelihood of a sample:

$$\log p_\theta(x) = \log r(g_\theta^{-1}(x)) + \text{logdet} \left(\nabla_x g_\theta^{-1}(x) \right).$$

What are our options?

- The maximum likelihood estimator:

$$\arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log r(g_\theta^{-1}(x_i)) + \text{logdet} \left(\nabla_x g_\theta^{-1}(x_i) \right) .$$

- Write down parameterized families with simple inverses and Jacobians?

- Work hard and compute the inverses and Jacobians?

- Generative adversarial nets: Give up on the MLE and try something else.

Revisiting the MLE

- The MLE minimizes KL divergence:

$$\begin{aligned}\arg\min_\theta D(p \parallel p_\theta) &= \arg\min_\theta H(p) + D(p \parallel p_\theta) \\ &= \arg\min_\theta \mathop{\mathbb{E}}_{x\sim p} -\log \frac{p_\theta(x)}{p(x)} p(x) \\ &= \arg\max_\theta \mathop{\mathbb{E}}_{x\sim p} \log p_\theta(x).\end{aligned}$$

- Where KL divergence is given by $D(p \parallel q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$

- What's so special about KL anyway?

What's so special about MLE?

- The MLE is consistent:

1. $D(p \,\|\, p_\theta) \geq 0$ and $D(p \,\|\, p_\theta) = 0$ iff $p = p_\theta$.
2. $\arg\max_\theta \mathop{\mathbb{E}}_{x \sim p} \log p_\theta(x) = \arg\min_\theta D(p \,\|\, p_\theta)$.
3. $\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x) = \mathbb{E}_{x \sim p} \log p_\theta(x)$.

- That's nice, but the same argument applies to the reverse KL $D(p_\theta \,\|\, p)$.

- Or to the symmetrized KL $D(p_\theta \,\|\, p) + D(p \,\|\, p_\theta)$.

Monte Carlo Estimation

- We can construct a Monte Carlo estimate of the MLE:

$$\arg\min_\theta D(p \parallel p_\theta) = \arg\max_\theta \mathop{\mathbb{E}}_{x\sim p} \log p_\theta(x)$$

$$\approx \arg\max_\theta \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i), \text{ where } x_i \sim p.$$

- How would we do this for reverse-KL?

$$\arg\min_\theta D(p_\theta \parallel p) = \arg\min_\theta \mathop{\mathbb{E}}_{x\sim p_\theta} \log \frac{p_\theta(x)}{p(x)}$$

$$\approx \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_\theta(x)}{p(x)}, \text{ where } x_i \sim p_\theta.$$

Information Divergences

- We can't compute $p_\theta(x)$ so the Monte Carlo estimator isn't that useful.
- This frees us to think about other information divergences. E.g.

$$D_f(p \parallel q) \equiv \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx.$$

- We can construct lower bounds on an f-divergence.
- Does the choice of information divergence make a difference? Unclear!

f-Divergences

Definition 1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex, lower-semicontinuous function, such that $f(1) = 0$. We define the* ***f-divergence*** *between two distributions with densities p and q on $\mathcal{X}$ by*

$$D_f(p \parallel q) \equiv \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx.$$

- For example, suppose $f(x) = x \log x$. Then $D_f(p \parallel q) = D(p \parallel q)$.
- If $f(x) = -\log x$ then $D_f(p \parallel q) = D(q \parallel p)$.
- Non-negative: $D_f(p \parallel q) = \mathop{\mathbb{E}}_{x \sim q}\left[f\left(\frac{p(x)}{q(x)}\right)\right] \geq f\left(\mathop{\mathbb{E}}_{x \sim q}\left[\frac{p(x)}{q(x)}\right]\right) = f(1) = 0.$

A Divergence Lower Bound

- For *any* function $T : \mathcal{X} \to \mathbb{R}$, $D_f(p \parallel q) \geq \mathop{\mathbb{E}}_{x \sim p} T(x) - \mathop{\mathbb{E}}_{x \sim q} f^*(T(x))$.
- The function $f^* : \mathbb{R} \to \mathbb{R}$ is the convex conjugate of f, defined by:

$$f^*(t) \equiv \sup_x \{tx - f(x)\}.$$

- Furthermore [Nguyen, Wainwright, and Jordan 2010]:

$$D_f(p \parallel q) = \sup_{T : \mathcal{X} \to \mathbb{R}} \left[\mathop{\mathbb{E}}_{x \sim p} T(x) - \mathop{\mathbb{E}}_{x \sim q} f^*(T(x)) \right].$$

Generative Adversarial Nets

- Solve a saddle-point problem:

$$\theta_f = \arg\min_\theta D_f(p \parallel p_\theta) = \arg\min_\theta \sup_\varphi \left[\mathop{\mathbb{E}}_{x\sim p} T_\varphi(x) - \mathop{\mathbb{E}}_{x\sim p_\theta} f^*(T_\varphi(x)) \right].$$

- Use an expressive parameterized family of functions $T_\varphi : \mathcal{X} \to \mathbb{R}$.

- Adversarial: optimize g_θ to minimize the objective, and T_ϕ to maximize it.

- The objective only requires samples from p_θ: $g_\theta(z) \sim p_\theta, \text{ where } z \sim r$.

5-Minute Break

Generative Adversarial Nets

- Solve a saddle-point problem:

$$\theta_f = \arg\min_\theta D_f(p \parallel p_\theta) = \arg\min_\theta \sup_\varphi \left[\mathop{\mathbb{E}}_{x\sim p} T_\varphi(x) - \mathop{\mathbb{E}}_{x\sim p_\theta} f^*(T_\varphi(x)) \right].$$

- Using the fact that:

$$D_f(p \parallel q) = \sup_{T:\mathcal{X}\to\mathbb{R}} \left[\mathop{\mathbb{E}}_{x\sim p} T(x) - \mathop{\mathbb{E}}_{x\sim q} f^*(T(x)) \right].$$

- Where does this fact come from?

Fenchel Duality

- The function $f^* : \mathbb{R} \to \mathbb{R}$ is the convex conjugate of f, defined by:

$$f^*(t) \equiv \sup_x \{tx - f(x)\}.$$

- For a convex, lower-semicontinuous function f: $f = f^{**}$ (Fenchel duality).

- A variational representation of f: $f(t) = f^{**}(t) \equiv \sup_t \{tx - f^*(t)\}$.

- Lift a variational representation of f to a variational representation of

$$D_f(p \parallel q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx.$$

Proof of the Lower Bound

Proposition [Nguyen, Wainwright, and Jordan 2010]:

$$D_f(p \parallel q) = \sup_{T:\mathcal{X}\to\mathbb{R}} \left[\mathop{\mathbb{E}}_{x\sim p} T(x) - \mathop{\mathbb{E}}_{x\sim q} f^*(T(x)) \right].$$

Proof:

$$\begin{aligned} D_f(p \parallel q) &= \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx = \int_{\mathcal{X}} q(x) \sup_t \left[t\frac{p(x)}{q(x)} - f^*(t) \right] dx \\ &= \int_{\mathcal{X}} \sup_t \left[tp(x) - f^*(t)q(x) \right] dx = \sup_{T:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} \left(T(x)p(x) - f^*(T(x))q(x) \right) dx \\ &= \sup_{T:\mathcal{X}\to\mathbb{R}} \left[\mathop{\mathbb{E}}_{x\sim p} T(x) - \mathop{\mathbb{E}}_{x\sim q} f^*(T(x)) \right]. \end{aligned}$$

The Goodfellow GAN

- Choose $f(x) = x \log x - (x+1)\log(x+1)$, resulting in

$$D_f(p \parallel q) = 2\mathrm{JSD}(p,q) - \log(4).$$

- The Jensen-Shannon Divergence is given by

$$\mathrm{JSD}(p,q) = \frac{1}{2} D_{\mathrm{KL}}\left(p \middle\| \frac{p+q}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(q \middle\| \frac{p+q}{2}\right).$$

- The convex conjugate of f is $f^*(t) = -\log(1-e^t)$.

- Parameterize $T_\varphi(x) = \log(d_\varphi(x))$. Then

$$\theta_f = \arg\min_\theta \sup_\varphi \left[\mathop{\mathbb{E}}_{x\sim p} \log d_\varphi(x) + \mathop{\mathbb{E}}_{z\sim r} \log(1 - d_\varphi(g_\theta(z))) \right].$$