Score Matching Models

Instructor: John Thickstun

Discussion Board: Available on Ed

Office Hours This Week: Friday 4:30pm

Zoom Link: Available on Canvas

Instructor Contact: thickstn@cs.washington.edu

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

Score Matching

- Want to learn $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ such that $s_\theta(x) \approx s(x) = \nabla_x \log p(x)$.

- What is a good way to quantify $s_\theta \approx s$? How about MSE?

$$\mathop{\mathbb{E}}_{x \sim p}\left[\frac{1}{2}\|s_\theta(x) - \nabla_x \log p(x)\|_2^2\right].$$

- Sample from a trained model $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ via Langevin dynamics:

$$\begin{aligned} x_{t+1} &= x_t - \eta \nabla_x \log p_\theta(x_t) + \sqrt{2\eta}\varepsilon_t \\ &= x_t - \eta s_\theta(x_t) + \sqrt{2\eta}\varepsilon_t \end{aligned}$$

Fisher Divergence

- What is a good way to quantify $s_\theta \approx s$? How about MSE?

$$\mathop{\mathbb{E}}_{x \sim p} \left[\frac{1}{2} \| s_\theta(x) - \nabla_x \log p(x) \|_2^2 \right].$$

- What is this quantity? It is the Fisher divergence $D_{\text{Fisher}}(p \parallel p_\theta)$.

- For any two probability distributions p, q we define

$$D_{\text{Fisher}}(p \parallel q) \equiv \mathop{\mathbb{E}}_{x \sim p} \left[\frac{1}{2} \left\| \nabla_x \log \frac{p(x)}{q(x)} \right\|^2 \right].$$

A Connection to KL Divergence

- Let $x \sim p$, $y \sim q$, and $\varepsilon_x, \varepsilon_y \sim \mathcal{N}(0, I)$ (independent samples).
- Define noisy samples: $\tilde{x}_t = x + \sqrt{t}\varepsilon_x, \quad \tilde{y}_t = y + \sqrt{t}\varepsilon_y.$
- Let $p_t(\tilde{x}_t), q_t(\tilde{y}_t)$ denote the densities of $\tilde{x}_t, \tilde{y}_t$ respectively.
- Adding noise to samples corresponds to convolution of densities.
- Gaussian convolution of densities causes smoothing.
- Proposition [Lyu, 2012]: $\frac{d}{dt} D(p_t \parallel q_t) = -D_{\text{Fisher}}(p_t \parallel q_t).$

Score Matching Revisited

- Want to learn $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ such that $s_\theta(x) \approx s(x) = \nabla_x \log p(x)$.

- Minimize the Fisher divergence, using implicit score matching:

$$\begin{aligned}\arg\min_\theta D_{\text{Fisher}}(p \parallel p_\theta) &= \arg\min_\theta \mathop{\mathbb{E}}_{x\sim p}\left[\frac{1}{2}\|s_\theta(x) - \nabla_x \log p(x)\|_2^2\right] \\ &= \arg\min_\theta \mathop{\mathbb{E}}_{x\sim p}\left[\operatorname{tr}\left(\nabla_x s_\theta(x)\right) + \frac{1}{2}\|s_\theta(x)\|_2^2\right].\end{aligned}$$

- This will work, in theory, but calculating the trace costs $O(d)$ calls to $s_\theta(x)$.

Sliced Score Matching

- Problem: calculating $\mathrm{tr}(\nabla_x s_\theta(x))$ requires $O(d)$ calls to $s_\theta(x)$.
- Don't match the full score vector $s_\theta(x) \approx s(x)$, just match $v^T s_\theta(x) \approx v^T s(x)$.
- The vector $v \in \mathbb{R}^d$ is chosen randomly.
- The random projection $v^T s(x) \in \mathbb{R}$ is a "slice" of the score function.
- Sliced score matching: $L(\theta, v) \equiv \mathop{\mathbb{E}}_{x \sim p}\left[\frac{1}{2}\left(v^T s_\theta(x) - v^T \nabla_x \log p(x)\right)^2\right].$

A Tractable Objective

- Sliced score matching: $L(\theta, v) \equiv \mathop{\mathbb{E}}_{x \sim p}\left[\frac{1}{2}\left(v^T s_\theta(x) - v^T \nabla_x \log p(x)\right)^2\right].$

- We'll minimize this objective over random projections $v \sim r$.

$$\begin{aligned}
\arg\min_\theta \mathop{\mathbb{E}}_{v \sim r} L(\theta, v) &= \arg\min_\theta \mathop{\mathbb{E}}_{v \sim r} v^T \mathop{\mathbb{E}}_{x \sim p}\left[\frac{1}{2}\|s_\theta(x) - \nabla_x \log p(x)\|^2\right] v \\
&= \arg\min_\theta \mathop{\mathbb{E}}_{v \sim r} v^T \mathop{\mathbb{E}}_{x \sim p}\left[\operatorname{tr}\left(\nabla_x s_\theta(x)\right) + \frac{1}{2}\|s_\theta(x)\|_2^2\right] v \\
&= \arg\min_\theta \mathop{\mathbb{E}}_{\substack{v \sim r \\ x \sim p}}\left[v^T \nabla_x s_\theta(x) v + \frac{1}{2}\left(v^T s_\theta(x)\right)^2\right].
\end{aligned}$$

This Does What We Want!

- Suppose $p(x) = p_{\theta^*}(x)$ for some parameter setting θ^*.
- Theorists call this the “realizable” setting.
- Assume that $\mathop{\mathbb{E}}_{v \sim r}[vv^T] \succ 0$ (r is a positive definite distribution).
- This assumption holds for, e.g. $r(v) = \mathcal{N}(v; 0, I)$.
- Proposition [Song, Garg, Shi, Ermon, 2019]:

$$\mathop{\mathbb{E}}_{v \sim r} L(\theta, v) = 0 \text{ iff } \theta = \theta^*.$$

The Sliced Objective is Correct

- Proposition [Song, Garg, Shi, Ermon, 2019]: $\mathop{\mathbb{E}}_{v \sim r} L(\theta, v) = 0$ iff $\theta = \theta^*$.

- Proof: Suppose $\mathop{\mathbb{E}}_{v \sim r} L(\theta, v) = 0$ (the converse is clearly true).

Because $L(\theta, v) \geq 0$, we see that for any value of x,

$$
\begin{aligned}
0 &= \mathop{\mathbb{E}}_{v \sim r} \left[\frac{1}{2} \left(v^T s_\theta(x) - v^T \nabla_x \log p(x) \right)^2 \right] \\
&= \mathop{\mathbb{E}}_{v \sim r} \left[\frac{1}{2} v^T \left(s_\theta(x) - \nabla_x \log p(x) \right) \left(s_\theta(x) - \nabla_x \log p(x) \right)^T v \right] \\
&= \frac{1}{2} \left(s_\theta(x) - \nabla_x \log p(x) \right)^T \mathop{\mathbb{E}}_{v \sim r} \left[v v^T \right] \left(s_\theta(x) - \nabla_x \log p(x) \right).
\end{aligned}
$$

And because $\mathop{\mathbb{E}}_{v \sim r} \left[v v^T \right] \succ 0$, if follows that $s_\theta(x) - \nabla_x \log p(x) = 0$.

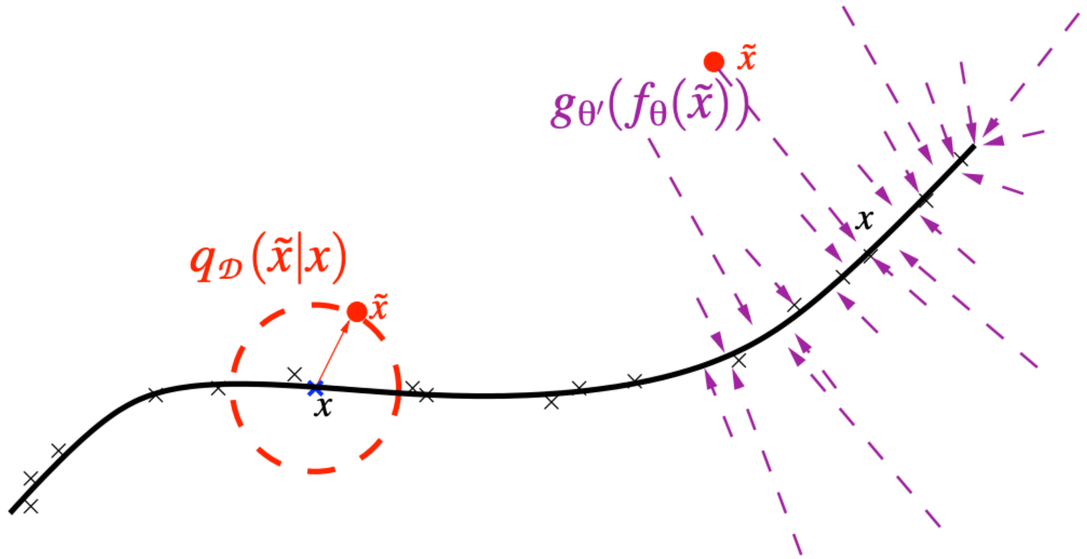
5-Minute Break

Denoising Autoencoders

- Learn to denoise noise-corrupted data: recover x given $\tilde{x} = x + \varepsilon$.

- Suppose data lives on a low-dimensional manifold.

- Learn to project noise-corrupted data back onto the manifold



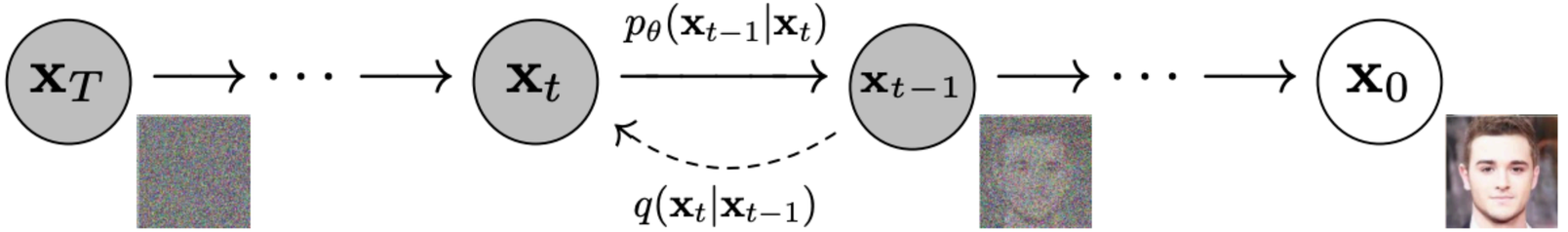
Vincent et. al., JMLR 2010

- Optimize a reconstruction objective:

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \mathop{\mathbb{E}}_{(x,\tilde{x})\sim p} \|x - g_\theta(f_\phi(\tilde{x}))\|^2.$$

Motivation

- Can we turn denoising autoencoders into a generative model?
- Idea: construct a Markov chain of progressively less noisy samples:



Ho, Jain, and Abbeel, Neurips 2020

- What if each transition $p_\theta(x_{t-1}|x_t)$ were given by a denoising autoencoder?

Score Matching in Denoising AE

- Denoising autoencoders are convenient to work with.
- Suppose that I would be satisfied modeling a smoothed version of $p(x)$, e.g.

$$q_\sigma(\tilde{x}) = \int_{\mathcal{X}} p_\sigma(\tilde{x}|x)p(x)\,dx.$$

- Where for example $p_\sigma(\tilde{x}|x) = \mathcal{N}(\tilde{x}; x, \sigma^2 I)$.
- We can implement score matching to model $q_\sigma(\tilde{x})$ pretty directly.
- Proposition [Vincent, 2011]:

$$\arg\min_\theta D_{\text{Fisher}}(q_\sigma \parallel p_\theta) = \arg\min_\theta \mathop{\mathbb{E}}_{\substack{x\sim p \\ \tilde{x}\sim p_\sigma(\cdot|x)}} \left[\frac{1}{2}\|s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x)\|_2^2\right].$$

Denoising Score Matching

- Proposition [Vincent, 2011]:

$$\arg\min_\theta D_{\text{Fisher}}(q_\sigma \parallel p_\theta) = \arg\min_\theta \mathop{\mathbb{E}}_{\substack{x\sim p \\ \tilde{x}\sim p_\sigma(\cdot|x)}} \left[\frac{1}{2}\|s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x)\|_2^2\right].$$

- Proof: Expand the quadratic:

$$\arg\min_\theta \mathop{\mathbb{E}}_{\tilde{x}\sim q_\sigma} \left[\frac{1}{2}\|s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})\|_2^2\right] = \arg\min_\theta \mathop{\mathbb{E}}_{\tilde{x}\sim q_\sigma} \left[\frac{1}{2}\|s_\theta(\tilde{x})\|^2 - s_\theta(\tilde{x})^T \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})\right].$$

Two applications of the log-derivative trick:

$$\mathop{\mathbb{E}}_{\tilde{x}\sim q_\sigma} \left[s_\theta(\tilde{x})^T \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})\right] = \int_{\mathcal{X}} s_\theta(\tilde{x})^T \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}) q_\sigma(\tilde{x})\, d\tilde{x} = \int_{\mathcal{X}} s_\theta(\tilde{x})^T \nabla_{\tilde{x}} q_\sigma(\tilde{x})\, d\tilde{x}$$

$$= \int_{\mathcal{X}} s_\theta(\tilde{x})^T \nabla_{\tilde{x}} \int_{\mathcal{X}} p(x) p_\sigma(\tilde{x}|x)\, dx\, d\tilde{x} = \int_{\mathcal{X}} s_\theta(\tilde{x})^T \int_{\mathcal{X}} p(x) p_\sigma(\tilde{x}|x) \nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x)\, dx\, d\tilde{x}$$

$$= \iint_{\mathcal{X}\times\mathcal{X}} p(x) p_\sigma(\tilde{x}|x) s_\theta(\tilde{x})^T \nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x)\, d(x, \tilde{x}) = \mathop{\mathbb{E}}_{\substack{x\sim p \\ \tilde{x}\sim p_\sigma(\tilde{x}|x)}} \left[s_\theta(\tilde{x})^T \nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x)\right].$$