COSC 594/650
Score-based optimization

Dongsheng Ding
dongshed@utk.edu

Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville

Recap: DDPM learns a conditional mean

Lecture 27: $\underset{\theta}{\text{minimize}}\ \mathbb{E}\|\varepsilon - \varepsilon_\theta(x_i, i)\|^2, \quad x_i = \bar{\alpha}_i\, x_0 + \sigma_i\, \varepsilon$

Minimizer: $\varepsilon^\star(x_i, i) = \mathbb{E}[\varepsilon \,|\, x_i]$

- TRAINING
  - ⋆ SGD on a stochastic least-squares risk
- OPEN QUESTION
  - ⋆ what is $\mathbb{E}[\varepsilon \,|\, x_i]$, as a property of p_{data}?

Today: the same least squares, from the score of the data

QUESTION: How do we learn the score of the data, and sample with it?

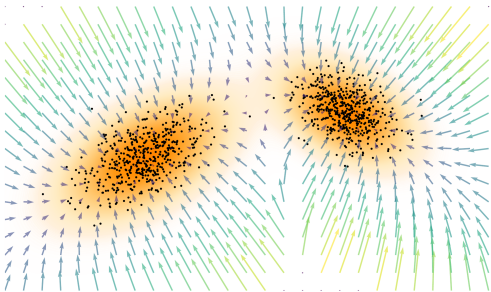
learn: least squares on the score – score matching

make it tractable: add noise – denoising score matching

sample: gradient ascent on $\log p$, plus noise – Langevin dynamics

The score: a vector field toward high density

Score: $s(x) = \nabla_x \log p(x), \quad s : \mathbb{R}^D \to \mathbb{R}^D$



- NO NORMALIZING CONSTANT

$$p = \frac{\tilde{p}}{Z} \implies \nabla_x \log p = \nabla_x \log \tilde{p}$$

  - $Z = \int \tilde{p}(x)\, dx$: never computed

- POINTS UPHILL
  - toward regions of higher density

Figure 3.2, Lai, Song, Kim, Mitsufuji, Ermon, *The Principles of Diffusion Models*, 2026

The score: the gradient of $\log p$, free of Z

Score matching: the optimization problem

Data: samples $x \sim p_{\text{data}}$, score $s(x) = \nabla_x \log p_{\text{data}}(x)$

Model (learned): a network $s_\theta(x) \in \mathbb{R}^D$

- PROBLEM: MATCH THE SCORE IN MEAN SQUARE

$$\underset{\theta}{\text{minimize}} \quad L_{\text{SM}}(\theta) \;=\; \frac{1}{2}\,\mathbb{E}_{x\sim p_{\text{data}}}\|s_\theta(x) - s(x)\|^2$$

  - ⋆ minimum 0: $s_\theta = s$ (rich enough model)
  - ⋆ the target $s(x)$ is unknown: we only have samples

- NEXT, STEPS 1–2: REMOVE THE UNKNOWN TARGET

Score matching: least squares with an unknown target

Step 1: expand the square

- EXPAND, WITH THE UNKNOWN $s = \nabla_x \log p_{\text{data}}$

$$\frac{1}{2}\|s_\theta - s\|^2 \;=\; \frac{1}{2}\|s_\theta\|^2 \;-\; s_\theta^\top s \;+\; \frac{1}{2}\|s\|^2$$

  - $\star$ $\frac{1}{2}\,\mathbb{E}\|s\|^2$: no θ, a constant
  - $\star$ $\mathbb{E}\big[s_\theta^\top s\big]$: the only term with both θ and the unknown s

- THE CROSS TERM: USE $s\,p_{\text{data}} = \nabla p_{\text{data}}$

$$\mathbb{E}_{x\sim p_{\text{data}}}\big[s_\theta^\top s\big] \;=\; \int s_\theta^\top \frac{\nabla p_{\text{data}}}{p_{\text{data}}}\, p_{\text{data}}\, dx \;=\; \int s_\theta^\top \nabla p_{\text{data}}\, dx$$

Only the cross term needs the unknown score

Step 2: integrate by parts (Hyvärinen 2005)

Assume: $s_\theta(x)\, p_{\text{data}}(x) \to 0$ as $\|x\| \to \infty$

- INTEGRATE BY PARTS IN EACH COORDINATE x_j

$$\int s_\theta^{(j)}(x)\; \partial_j\, p_{\text{data}}(x)\, dx \;=\; -\int \partial_j\, s_\theta^{(j)}(x)\; p_{\text{data}}(x)\, dx$$

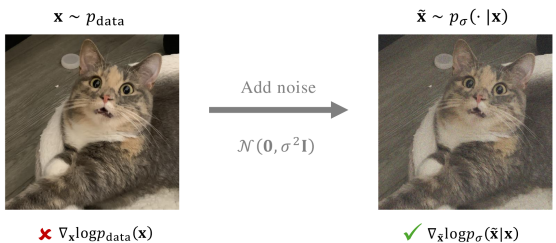
  - ⋆ sum over j: $\mathbb{E}\big[s_\theta^\top s\big] = -\,\mathbb{E}\big[\operatorname{tr} \nabla_x s_\theta(x)\big]$

$$L_{\text{SM}}(\theta) \;=\; \mathbb{E}_{x\sim p_{\text{data}}}\Big[\operatorname{tr} \nabla_x s_\theta(x) + \frac{1}{2}\|s_\theta(x)\|^2\Big] \;+\; \text{const}$$

No unknown target, but $\operatorname{tr} \nabla_x s_\theta$ costs D derivatives

Step 3: add noise, then condition on the clean sample

Noisy data: $\tilde{x} = x + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \quad p_\sigma(\tilde{x} \mid x) = \mathcal{N}(\tilde{x};\, x,\, \sigma^2 I)$



■ CONDITIONAL SCORE: KNOWN

$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x} \mid x) \;=\; \frac{x - \tilde{x}}{\sigma^2} \;=\; -\frac{\varepsilon}{\sigma}$$

⋆ $\nabla \log p_\sigma$: still unknown

Figure 3.5, Lai, Song, Kim, Mitsufuji, Ermon, *The Principles of Diffusion Models*, 2026

Adding noise makes the target known

The denoising score matching problem

Marginal: $p_\sigma(\tilde{x}) = \int p_\sigma(\tilde{x} \mid x)\, p_{\text{data}}(x)\, dx$, model $s_\theta(\tilde{x}, \sigma)$

- IDEAL: SCORE MATCHING ON p_σ (SLIDE 4)

$$L_{\text{SM}}(\theta; \sigma) \;=\; \frac{1}{2}\, \mathbb{E}_{\tilde{x} \sim p_\sigma} \| s_\theta(\tilde{x}, \sigma) - \nabla \log p_\sigma(\tilde{x}) \|^2$$

- TRACTABLE: THE TARGET OF STEP 3

$$\underset{\theta}{\text{minimize}} \quad L_{\text{DSM}}(\theta; \sigma) \;=\; \frac{1}{2}\, \mathbb{E}_{x,\, \varepsilon} \| s_\theta(x + \sigma\varepsilon, \sigma) + \frac{\varepsilon}{\sigma} \|^2$$

Next, Steps 4–5: $L_{\text{DSM}} = L_{\text{SM}} + C$

Step 4: the conditional score averages to the score

Conditional score: $Y = \nabla_{\tilde{x}} \log p_\sigma(\tilde{x} \,|\, x) = -\,\varepsilon/\sigma, \quad$ fixed σ

- DIFFERENTIATE UNDER THE INTEGRAL

$$\nabla \log p_\sigma(\tilde{x}) \;=\; \frac{\nabla p_\sigma(\tilde{x})}{p_\sigma(\tilde{x})} \;=\; \frac{1}{p_\sigma(\tilde{x})} \int \nabla_{\tilde{x}}\, p_\sigma(\tilde{x} \,|\, x)\; p_{\text{data}}(x)\, dx$$

- USE $\nabla p = p\, \nabla \log p$, THEN BAYES' RULE

$$\nabla \log p_\sigma(\tilde{x}) \;=\; \int \nabla_{\tilde{x}} \log p_\sigma(\tilde{x} \,|\, x)\; \underbrace{\frac{p_\sigma(\tilde{x} \,|\, x)\, p_{\text{data}}(x)}{p_\sigma(\tilde{x})}}_{p(x \,|\, \tilde{x})}\, dx \;=\; \mathbb{E}\big[\, Y \,|\, \tilde{x} \,\big]$$

The score = the average conditional score

Step 5: same minimizer, same gradient

- LEAST SQUARES (LECTURE 27), ANY f

$$\mathbb{E}\|f(\tilde{x}) - Y\|^2 \;=\; \mathbb{E}\|f(\tilde{x}) - \mathbb{E}[Y \mid \tilde{x}]\|^2 \;+\; \mathbb{E}\|Y - \mathbb{E}[Y \mid \tilde{x}]\|^2$$

- TAKE $f = s_\theta(\cdot, \sigma)$, USE STEP 4: $\mathbb{E}[Y \mid \tilde{x}] = \nabla \log p_\sigma(\tilde{x})$

$$L_{\mathrm{DSM}}(\theta;\sigma) \;=\; L_{\mathrm{SM}}(\theta;\sigma) \;+\; C, \qquad C \;=\; \frac{1}{2}\,\mathbb{E}\|Y - \mathbb{E}[Y \mid \tilde{x}]\|^2 \;\geq\; 0$$

  - ⋆ same gradient in θ: SGD on L_{DSM} is SGD on L_{SM}
  - ⋆ same minimizer: $s^\star(\tilde{x}, \sigma) = \nabla \log p_\sigma(\tilde{x})$

DSM = score matching + a constant floor C

Step 6: predicting the noise is predicting the score

Step 4: $\nabla \log p_\sigma(\tilde{x}) = \mathbb{E}[Y \mid \tilde{x}] = -\,\mathbb{E}[\varepsilon \mid \tilde{x}]/\sigma$

- NOISE PREDICTOR $\varepsilon_\theta = -\sigma s_\theta$: HALF THE DDPM LOSS (LECTURE 27)

$$\sigma^2 \cdot \frac{1}{2}\|s_\theta(\tilde{x},\sigma) + \frac{\varepsilon}{\sigma}\|^2 \;=\; \frac{1}{2}\|\varepsilon - \varepsilon_\theta(\tilde{x},\sigma)\|^2, \qquad \varepsilon^\star \;=\; -\,\sigma\, s^\star \;=\; \mathbb{E}[\varepsilon \mid \tilde{x}]$$

- DENOISER: TWEEDIE'S FORMULA, FROM $x = \tilde{x} - \sigma\varepsilon$

$$\mathbb{E}[x \mid \tilde{x}] \;=\; \tilde{x} - \sigma\,\mathbb{E}[\varepsilon \mid \tilde{x}] \;=\; \tilde{x} + \sigma^2\,\nabla \log p_\sigma(\tilde{x})$$

  - ⋆ one ascent step on $\log p_\sigma$ with step size σ^2

Noise predictor, score, denoiser: one object

Training: many noise levels (NCSN)

repeat: $x \sim p_{\text{data}}, \quad i \sim \text{Uniform}\{1, \dots, L\}, \quad \varepsilon \sim \mathcal{N}(0, I)$

$\tilde{x} = x + \sigma_i \varepsilon$ (Step 3)

gradient step on $\lambda(\sigma_i)\, \frac{1}{2} \| s_\theta(\tilde{x}, \sigma_i) + \varepsilon / \sigma_i \|^2$ (SGD or Adam)

- LEVELS $0 < \sigma_1 < \cdots < \sigma_L$, ONE NETWORK $s_\theta(\tilde{x}, \sigma)$

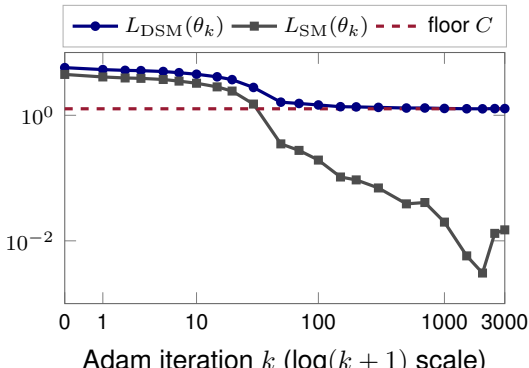
$$L_{\text{NCSN}}(\theta) \;=\; \sum_{i=1}^{L} \lambda(\sigma_i)\, L_{\text{DSM}}(\theta; \sigma_i), \qquad \lambda(\sigma_i) \;>\; 0$$

  - ⋆ minimizer: $s^\star(\cdot, \sigma_i) = \nabla \log p_{\sigma_i}$ for every i (Step 5)

Training: SGD on one least-squares risk over all levels

Experiment: the training loss has a floor

Setup: running example, one level $\sigma = 0.3$, network 1-64-64-1, Adam, batch 128

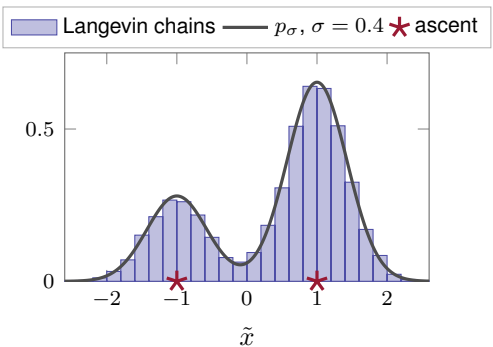


- WHAT WE SEE: L_{DSM}
  - ⋆ stalls at $C = 1.27$
  - ⋆ batch loss: 1.27 ± 0.39
- WHAT WE WANT: L_{SM}
  - ⋆ $L_{\mathrm{DSM}} - C$: $4.5 \to 0.01$

Judge training by the gap above the floor

Sampling: Langevin dynamics

Langevin: $\tilde{x}_{n+1} = \tilde{x}_n + \eta\, s_\theta(\tilde{x}_n, \sigma) + \sqrt{2\eta}\,\varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0, I)$



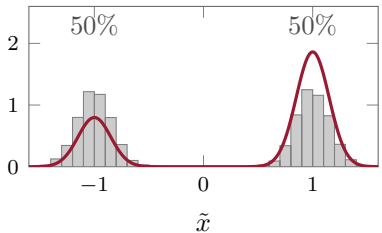
- WITHOUT NOISE
  - ascent stops at a mode
- WITH NOISE: SAMPLES FROM p_σ
  - 69.8% at $\tilde{x} > 0$ (true 69.6%)

Sampling = gradient ascent + the right amount of noise

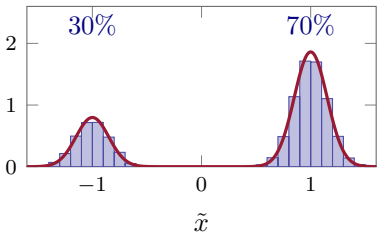
Sampling: annealed Langevin dynamics

Annealing: for $i = L, \dots, 1$: Langevin with $s_\theta(\cdot, \sigma_i)$, warm start from level $i+1$

Langevin with the score of p_{data}



annealed: $\sigma = 1.5 \to 0.01$, 20 levels



- SMOOTH FIRST, THEN SHARPEN
  - ⋆ avoids the wrong basin (Lecture 23)

Large noise first: smoothing removes the barrier

From score matching to the DDPM loss

Slide	Objective	Optimization view
4	$\frac{1}{2}\,\mathbb{E}\|s_\theta(x) - \nabla \log p_{\text{data}}(x)\|^2$	least squares, unknown target
6	$\mathbb{E}\big[\operatorname{tr} \nabla_x s_\theta + \frac{1}{2}\|s_\theta\|^2\big]$	no target, D derivatives
8–10	$\frac{1}{2}\,\mathbb{E}\|s_\theta(\tilde{x}, \sigma) + \varepsilon/\sigma\|^2$	known target, same minimizer
11	$\frac{1}{2}\,\mathbb{E}\|\varepsilon - \varepsilon_\theta(\tilde{x}, \sigma)\|^2$	half the DDPM loss, $\lambda = \sigma^2$
14–15	$\tilde{x} + \eta\, s_\theta + \sqrt{2\eta}\,\varepsilon$	noisy gradient ascent

Train by least squares, sample by noisy gradient ascent

Questions?

(**Email: dongshed@utk.edu**)