COSC 594/650
Variational diffusion models

Dongsheng Ding
dongshed@utk.edu

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

Recap: learning is stochastic optimization

Lectures 16–19: learning $=$ minimizing a risk $L(\theta) = \mathbb{E}_\xi\big[\ell(\xi, \theta)\big]$ by SGD

Lecture 26: Adam with clipping: the default optimizer for large networks

- SO FAR: PREDICT
  - ⋆ data (x, y): learn a map $x \mapsto y$ – ERM, Lecture 16
- NEXT THREE LECTURES: GENERATE
  - ⋆ data $x \sim p_{\text{data}}$ only: learn to draw new samples of x – diffusion models

Today: which risk does a diffusion model minimize?

QUESTION: How do we train a model that generates new data?

maximize a lower bound on the likelihood – variational view

Maximum likelihood with a latent variable

Model: latent $z \sim p(z) = \mathcal{N}(0, I)$, data $x \sim p_\theta(x \mid z)$

Likelihood: $p_\theta(x) = \int p_\theta(x \mid z)\, p(z)\, dz$

- MAXIMUM LIKELIHOOD

$$\underset{\theta}{\text{maximize}} \;\; \mathbb{E}_{x \sim p_{\text{data}}}\big[\log p_\theta(x)\big] \;=\; -\,\mathrm{KL}\big(p_{\text{data}} \,\|\, p_\theta\big) + \text{const}$$

- WHY IT IS HARD
  - $\star$ $\log p_\theta(x)$: the log of an integral over z
  - $\star$ $\nabla_\theta \log p_\theta(x) = \mathbb{E}_{z \sim p_\theta(z \mid x)}\big[\nabla_\theta \log p_\theta(x, z)\big]$: needs the posterior

Maximum likelihood: the right objective, but intractable

Derivation: the evidence lower bound (ELBO)

Encoder: any $q(z \mid x)$, an approximation of the posterior $p_\theta(z \mid x)$

- MULTIPLY AND DIVIDE BY q, THEN JENSEN ($\log$ IS CONCAVE)

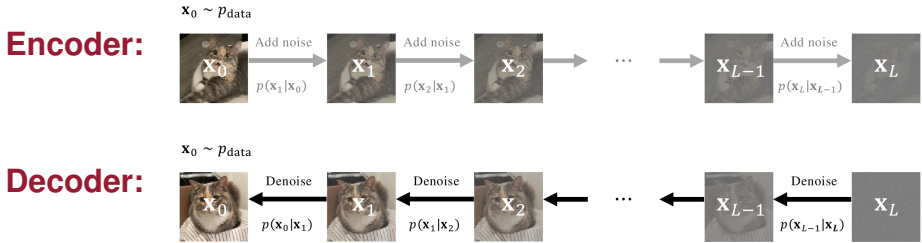
$$\log p_\theta(x) \;=\; \log \mathbb{E}_{q(z|x)}\left[\frac{p_\theta(x,z)}{q(z \mid x)}\right] \;\geq\; \underbrace{\mathbb{E}_{q(z|x)}\left[\log \frac{p_\theta(x,z)}{q(z \mid x)}\right]}_{\mathrm{ELBO}(x;\theta,q)}$$

- THE GAP: USE $p_\theta(x,z) = p_\theta(z \mid x)\, p_\theta(x)$

$$\log p_\theta(x) - \mathrm{ELBO} \;=\; \mathbb{E}_{q(z|x)}\left[\log \frac{q(z \mid x)\, p_\theta(x)}{p_\theta(x,z)}\right] \;=\; \mathrm{KL}\big(q(z \mid x) \,\|\, p_\theta(z \mid x)\big) \;\geq\; 0$$

Maximize the ELBO: a lower bound, tight when $q(z \mid x) = p_\theta(z \mid x)$

DDPM: a VAE whose encoder adds noise



Figures 2.4–2.5, Lai, Song, Kim, Mitsufuji, Ermon, *The Principles of Diffusion Models*, 2026

- LATENTS $z = (x_1, \ldots, x_L)$: NOISIER AND NOISIER COPIES OF x_0
  - ⋆ encoder: fixed, a little Gaussian noise per step
  - ⋆ decoder: learned

DDPM: fix the encoder, learn the decoder

DDPM: the ELBO maximization problem

Encoder (fixed): $p(x_{1:L} \,|\, x_0) = \prod_{i=1}^{L} p(x_i \,|\, x_{i-1}), \quad p(x_i \,|\, x_{i-1}) = \mathcal{N}(\alpha_i x_{i-1},\, \beta_i^2 I)$

Decoder (learned): $p_\theta(x_{0:L}) = \mathcal{N}(x_L; 0, I) \prod_{i=1}^{L} p_\theta(x_{i-1} \,|\, x_i)$

- PROBLEM: SLIDE 4 WITH $q(z \,|\, x_0) = p(x_{1:L} \,|\, x_0)$

$$\underset{\theta}{\text{maximize}} \;\; \mathbb{E}_{x_0 \sim p_{\text{data}}}\big[\text{ELBO}(x_0; \theta)\big], \quad \text{ELBO}(x_0; \theta) \;=\; \mathbb{E}_{p(x_{1:L}|x_0)}\left[\log \frac{p_\theta(x_{0:L})}{p(x_{1:L} \,|\, x_0)}\right]$$

  - $\alpha_i^2 + \beta_i^2 = 1, \;\; i = 1, \ldots, L$
  - variable: θ, one network for all steps

DDPM: maximize the ELBO over the decoder only

Derivation: the ELBO compares two chains

Forward joint: $p(x_{0:L}) = p_{\text{data}}(x_0)\, p(x_{1:L} \mid x_0)$, p_i: density of x_i

- AVERAGE THE ELBO OVER THE DATA

$$\begin{aligned}\mathbb{E}_{x_0}\big[-\text{ELBO}(x_0;\theta)\big] &= \mathbb{E}\left[\log \frac{p(x_{0:L})}{p_\theta(x_{0:L})}\right] - \mathbb{E}\big[\log p_{\text{data}}(x_0)\big] \\ &= \text{KL}\big(p(x_{0:L}) \,\|\, p_\theta(x_{0:L})\big) + \text{const}\end{aligned}$$

- WRITE THE FORWARD CHAIN BACKWARD: $p(x_{0:L}) = p_L(x_L) \prod_i p(x_{i-1} \mid x_i)$

$$\text{KL}\big(p(x_{0:L}) \,\|\, p_\theta(x_{0:L})\big) = \text{KL}\big(p_L \,\|\, \mathcal{N}(0, I)\big) + \sum_{i=1}^{L} \mathbb{E}_{x_i} \text{KL}\big(p(x_{i-1} \mid x_i) \,\|\, p_\theta(x_{i-1} \mid x_i)\big)$$

Maximize the ELBO = match every reverse step

A simple KL problem

$$\underset{\theta}{\text{minimize}} \quad \sum_{i=1}^{L} \mathbb{E}_{x_i \sim p_i} \Big[\mathrm{KL}\big(p(x_{i-1} \,|\, x_i) \,\|\, p_\theta(x_{i-1} \,|\, x_i)\big) \Big]$$

- MATCH EVERY REVERSE STEP
  - ⋆ minimum 0: $p_\theta(x_{i-1} \,|\, x_i) = p(x_{i-1} \,|\, x_i)$ for every i (rich enough model)
- NEXT, STEPS 1–5: FROM MATCHING DISTRIBUTIONS TO LEAST SQUARES
  - ⋆ each term, $i \geq 2$: $w_i \, \mathbb{E}\|\varepsilon - \varepsilon_\theta(x_i, i)\|^2 + \text{const}$

The KL problem: the ELBO, written step by step

Step 1: jump to any noise level

One step: $x_i = \alpha_i\, x_{i-1} + \beta_i\, \varepsilon_i, \quad \alpha_i^2 + \beta_i^2 = 1, \quad \varepsilon_1, \varepsilon_2, \ldots$ independent $\mathcal{N}(0, I)$

- TWO STEPS: THE NOISE VARIANCE IS $\alpha_2^2\beta_1^2 + \beta_2^2 = 1 - \alpha_1^2\alpha_2^2$

$$x_2 \;=\; \alpha_2\big(\alpha_1 x_0 + \beta_1\varepsilon_1\big) + \beta_2\varepsilon_2 \;=\; \alpha_2\alpha_1\, x_0 + \big(\alpha_2\beta_1\varepsilon_1 + \beta_2\varepsilon_2\big)$$

- BY INDUCTION

$$x_i \;=\; \bar\alpha_i\, x_0 + \sigma_i\, \varepsilon, \qquad \bar\alpha_i \;=\; \alpha_1\alpha_2\cdots\alpha_i, \qquad \sigma_i \;=\; \sqrt{1-\bar\alpha_i^2}, \qquad \varepsilon \sim \mathcal{N}(0, I)$$

⋆ $L = 1000$: $\bar\alpha_L \approx 0.006$, so $x_L \approx \mathcal{N}(0, I)$

Any noise level in one shot: $p_i(x_i \,|\, x_0) = \mathcal{N}\big(\bar\alpha_i x_0,\, \sigma_i^2 I\big)$

Step 2: the target is unknown

KL problem (slide 8): $\displaystyle\operatorname*{minimize}_{\theta} \sum_i \mathbb{E}_{x_i \sim p_i} \mathrm{KL}\big(p(x_{i-1} \,|\, x_i) \,\|\, p_\theta(x_{i-1} \,|\, x_i)\big)$

- BAYES' RULE

$$p(x_{i-1} \,|\, x_i) \;=\; p(x_i \,|\, x_{i-1})\, \frac{p_{i-1}(x_{i-1})}{p_i(x_i)}, \qquad p_i(x_i) \;=\; \int p_i(x_i \,|\, x_0)\, p_{\mathrm{data}}(x_0)\, dx_0$$

- ⋆ the marginals p_{i-1}, p_i average over all of p_{data}
- ⋆ for a fixed x_i: a mixture over all x_0 that could have produced x_i

The reverse step mixes over all data: intractable

Step 3: condition on the clean sample x_0

Known target: $p(x_{i-1} \mid x_i, x_0) = \dfrac{p(x_i \mid x_{i-1})\, p_{i-1}(x_{i-1} \mid x_0)}{p_i(x_i \mid x_0)}$: Gaussian

- SAME OBJECTIVE UP TO A CONSTANT (THEOREM 2.2.1), $i \geq 2$

$$\mathbb{E}_{x_i} \mathrm{KL}\big(p(x_{i-1}|x_i) \,\|\, p_\theta\big) = \mathbb{E}_{x_0}\, \mathbb{E}_{x_i|x_0}\, \mathrm{KL}\big(p(x_{i-1}|x_i, x_0) \,\|\, p_\theta\big) - C$$

- WHY: SPLIT THE LOG, THEN AVERAGE OVER $x_0 \sim p(x_0 \mid x_i)$

$$\log \frac{p(x_{i-1}|x_i, x_0)}{p_\theta(x_{i-1}|x_i)} = \underbrace{\log \frac{p(x_{i-1}|x_i, x_0)}{p(x_{i-1}|x_i)}}_{\text{average: } C, \text{ no } \theta} + \underbrace{\log \frac{p(x_{i-1}|x_i)}{p_\theta(x_{i-1}|x_i)}}_{\text{average: left-hand KL}}$$

Same minimizer, but a known target

Step 3: the conditional target in closed form

Step 1: $p(x_i \,|\, x_{i-1}) = \mathcal{N}(\alpha_i x_{i-1},\, \beta_i^2 I), \quad p_{i-1}(x_{i-1} \,|\, x_0) = \mathcal{N}(\bar{\alpha}_{i-1} x_0,\, \sigma_{i-1}^2 I)$

- COMPLETE THE SQUARE IN x_{i-1}, USE $\sigma_i^2 = \beta_i^2 + \alpha_i^2 \sigma_{i-1}^2$

$$-2 \log p(x_{i-1} \,|\, x_i, x_0) \;=\; \frac{\|x_i - \alpha_i x_{i-1}\|^2}{\beta_i^2} + \frac{\|x_{i-1} - \bar{\alpha}_{i-1} x_0\|^2}{\sigma_{i-1}^2} + \text{const}$$

$$p(x_{i-1} \,|\, x_i, x_0) \;=\; \mathcal{N}\big(\mu(x_i, x_0),\, \tilde{\beta}_i^2 I\big), \qquad \frac{1}{\tilde{\beta}_i^2} \;=\; \frac{\alpha_i^2}{\beta_i^2} + \frac{1}{\sigma_{i-1}^2} \;=\; \frac{\sigma_i^2}{\beta_i^2\, \sigma_{i-1}^2}$$

$$\mu(x_i, x_0) \;=\; \tilde{\beta}_i^2 \Big(\frac{\alpha_i}{\beta_i^2}\, x_i + \frac{\bar{\alpha}_{i-1}}{\sigma_{i-1}^2}\, x_0 \Big) \;=\; \frac{\alpha_i \sigma_{i-1}^2}{\sigma_i^2}\, x_i + \frac{\bar{\alpha}_{i-1} \beta_i^2}{\sigma_i^2}\, x_0$$

The target: a Gaussian with known mean and variance

Step 4: a Gaussian model turns KL into squared error

Model: $p_\theta(x_{i-1} \mid x_i) = \mathcal{N}\big(\mu_\theta(x_i, i),\, \tilde{\beta}_i^2 I\big)$: same variance, learn the mean

- TWO GAUSSIANS WITH THE SAME VARIANCE

$$\mathrm{KL}\Big(\mathcal{N}(\mu, \tilde{\beta}_i^2 I) \,\Big\|\, \mathcal{N}(\mu_\theta, \tilde{\beta}_i^2 I)\Big) \;=\; \frac{1}{2\tilde{\beta}_i^2}\, \|\mu - \mu_\theta\|^2$$

- SUM OVER THE LEVELS $i \geq 2$: ONE NETWORK $\mu_\theta(x, i)$

$$L_{\mathrm{DDPM}}(\theta) \;=\; \sum_{i=2}^{L} \mathbb{E}_{x_0}\, \mathbb{E}_{x_i \mid x_0} \left[\frac{1}{2\tilde{\beta}_i^2}\, \|\mu_\theta(x_i, i) - \mu(x_i, x_0)\|^2 \right]$$

Matching distributions = regressing the mean

Step 5: predict the noise – the DDPM loss

- WRITE x_0 THROUGH THE NOISE (STEP 1): $x_0 = (x_i - \sigma_i \varepsilon)/\bar{\alpha}_i$

$$\mu(x_i, x_0) \;=\; \frac{1}{\alpha_i}\Big(x_i - \frac{\beta_i^2}{\sigma_i}\,\varepsilon\Big), \qquad \text{choose} \quad \mu_\theta(x_i, i) \;=\; \frac{1}{\alpha_i}\Big(x_i - \frac{\beta_i^2}{\sigma_i}\,\varepsilon_\theta(x_i, i)\Big)$$

- THE MEAN ERROR IS A NOISE ERROR

$$\frac{1}{2\tilde{\beta}_i^2}\|\mu_\theta - \mu\|^2 \;=\; w_i\,\|\varepsilon - \varepsilon_\theta(x_i, i)\|^2, \qquad w_i \;=\; \frac{\beta_i^2}{2\alpha_i^2\sigma_{i-1}^2}$$

DDPM loss (drop w_i; Ho et al. 2020):

$$L_{\text{simple}}(\theta) \;=\; \mathbb{E}_{i,x_0,\varepsilon}\|\varepsilon - \varepsilon_\theta\big(\bar{\alpha}_i x_0 + \sigma_i\varepsilon,\; i\big)\|^2$$

The network predicts the noise that was added

Training: SGD on the DDPM loss

repeat: $x_0 \sim p_{\text{data}}, \quad i \sim \text{Uniform}\{1, \dots, L\}, \quad \varepsilon \sim \mathcal{N}(0, I)$

$x_i = \bar{\alpha}_i \, x_0 + \sigma_i \, \varepsilon$ (Step 1)

gradient step on $\|\varepsilon - \varepsilon_\theta(x_i, i)\|^2$ (SGD or Adam)

- A RISK $\mathbb{E}_\xi\big[\ell(\xi, \theta)\big]$ (LECTURE 16), $\xi = (x_0, i, \varepsilon)$: UNBIASED GRADIENTS
- WHAT IT LEARNS: ANY f

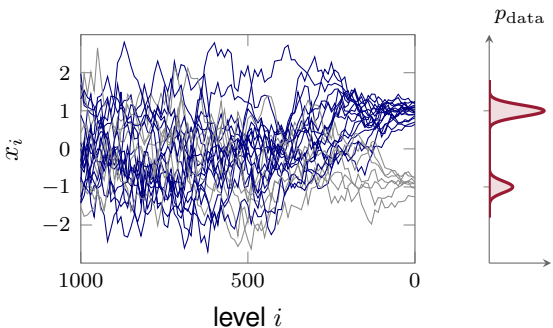
$$\mathbb{E}\|\varepsilon - f(x_i)\|^2 \;=\; \mathbb{E}\|\varepsilon - \mathbb{E}[\varepsilon \,|\, x_i]\|^2 + \mathbb{E}\|\mathbb{E}[\varepsilon \,|\, x_i] - f(x_i)\|^2$$

  - ⋆ minimizer $\varepsilon^\star(x_i, i) = \mathbb{E}[\varepsilon \,|\, x_i]$; the loss at $\varepsilon^\star$ is > 0

Training: SGD on a stochastic least-squares risk

Sampling: run the learned chain backward

Step $i \to i-1$: $x_{i-1} = \mu_\theta(x_i, i) + \tilde{\beta}_i\, \varepsilon'$ (Step 5), $\varepsilon' \sim \mathcal{N}(0, I)$, $x_L \sim \mathcal{N}(0, I)$



- EXACT $\varepsilon^\star$, $L = 1000$
  - 70.0% near $+1$ (data 70%)
  - cost: 1000 network calls

Sampling: L small denoising steps, one network call each

From the ELBO to the DDPM loss

Slide	Objective	Optimization view
4, 6	$\text{maximize}_\theta \ \mathbb{E}\,\text{ELBO}(x_0;\theta)$	a computable lower bound
7–8	$\sum_i \mathbb{E}\,\text{KL}\big(p(x_{i-1}\|x_i) \,\|\, p_\theta\big)$	same problem, unknown target
11	$\sum_i \mathbb{E}\,\text{KL}\big(p(x_{i-1}\|x_i, x_0) \,\|\, p_\theta\big)$	same minimizer, known target
13	$\sum_i \frac{1}{2\tilde{\beta}_i^2}\,\mathbb{E}\|\mu_\theta - \mu\|^2$	weighted least squares
14	$\mathbb{E}\|\varepsilon - \varepsilon_\theta(x_i, i)\|^2$	least squares, SGD/Adam

DDPM loss: least squares on the added noise

Questions?

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