Diffusion models

(Hierarchical) latent variable model

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p(x_{t-1} | x_t)$$

$$p_\theta(x_{t-1} | x_t) \sim \mathcal{N}(\mu_\theta(x_t, t) ; \Sigma_\theta(x_t ; t)).$$

Fix approximate posterior as

$$q(x_{1:T} | x_0) = \prod q(x_t | x_{t-1})$$

$$q(x_t | x_{t-1}) \sim \mathcal{N}(\sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$$

$\beta_1, \dots, \beta_T$ fixed.

$$\mathbb{E} - \log p_\theta(x_0) \leq \mathbb{E}_q \left[- \log \frac{p_\theta(x_{0:T})}{q(x_{1:T} | x_0)} \right]$$

$$= \mathbb{E}_q \left[- \log p(x_T) - \sum_{t \geq 1} \frac{p_\theta(x_{t-1} | x_t)}{q(x_t | x_{t-1})} \right]$$

$$= \mathbb{E}_q\Big[\overbrace{D_{KL}(q(x_T|x_0)\,\|\,p(x_T))}^{L_T}$$

$$+ \sum_{t>1} \overbrace{D_{KL}(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t))}^{L_t}$$

$$\underbrace{-\log p_\theta(x_0|x_1)}_{L_0}\Big].$$

Derivation: expand definition of conditional probability.

Note we have closed form

$$q(x_t|x_0) \sim \mathcal{N}(\sqrt{\bar\alpha_t}\,x_0,\ (1-\bar\alpha_t)I)$$

where $\alpha_t = (1-\beta_t)$, $\bar\alpha_t = \prod_{s=1}^{t} \alpha_s$.

Similar argument to obtain

$$q(x_{t-1}|x_t,x_0) \sim \mathcal{N}(\tilde\mu_t(x_t,x_0),\ \tilde\beta_t I)$$

$$\tilde\mu_t(x_t,x_0) = \frac{\sqrt{\bar\alpha_{t-1}}}{1-\bar\alpha_t}\beta_t x_0 + \frac{\sqrt{\bar\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t} x_t$$

$$\tilde\beta_t = \frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\beta_t.$$

Parametrization of model.

Variance: either fixed (Ho et al.),

or parametrize as interpolation

$$\Sigma_\theta(x_t, t) = \exp\left(v \log \beta_t + (1-v)\log \tilde{\beta}_t\right)$$

optimal for $x_0 \sim \mathcal{N}$

optimal for x_0 dirac.

For moderate t, $\beta_t \approx \tilde{\beta}_t$

(low large contribution at beginning to likelihood term).

$$L_{t-1} = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\right\|^2\right] + C$$

(KL of two Gaussians).

Reparametrize (x_t, x_0) as:

$$x_t, \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon\right)$$

See that μ_θ predicts $\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon\right)$

hence choose parametrization

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t, t)\right).$$

Sampling process becomes

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t, t)\right) + \sigma_t z.$$

Note: connection to Langevin dynamics.

Final loss:

$$L_{t-1} = \mathbb{E}_{x_0,\varepsilon}\left(\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t\right)\right\|^2\right)$$

weighted L_2 loss.

$$L_{simple} = \mathbb{E}_{x_0,\varepsilon}\sum_t \left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t\right)\right\|^2$$

$$= \mathbb{E}_{x_0,\varepsilon,t\sim U}\left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t\right)\right\|^2$$

Note: training on L_{simple} is better (and simpler).

→ But no variance signal.

(Nichol and Dhariwal): use L_{VLB} for variance, use IS to reduce MC variance in t.

Parametrization of ϵ_θ.

ϵ_θ is chosen to be a U-net.

t parametrized w sinusoidal features

(full parameter sharing across time).

→ Denoising diffusion models, Ho et al. 2020

→ Improved denoising diffusion models, Nichol and Dhariwal 2021

A detour to score-based generative models.

"Generative modeling by estimating gradients of the data distribution", Song and Ermon 2019

"Improved techniques for training score-based generative models", Song and Ermon 2019.

$$x_t = x_{t-1} + \alpha \nabla_x \log p(x_{t-1}) + \sqrt{2\alpha}\, z_t$$

$$dX = \nabla \log p(X) + \sqrt{2}\, dW$$

then stationary distribution is p.

Generative model; controlled by gradient process

$$\nabla \log q(x) = s_\theta(x).$$

Score matching.

Suppose we wish to learn some score-based generative model from data, natural to consider

$$J(\theta) := \min_\theta \mathbb{E}_p \| s_\theta(x) - \nabla_x \log p(x) \|^2$$

But: only access to samples from p, how to compute $\nabla_x \log p(x)$?

Claim: $J(\theta) = \mathbb{E}\left(\operatorname{tr}\left[\nabla_x s_\theta(x)\right] + \frac{1}{2} \| s_\theta(x) \|^2 \right)$

Proof: only need to consider cross-term.

$$\mathbb{E}\ s_\theta(x)^T \nabla_x \log p(x).$$

1-D argument:

$$\mathbb{E}_p (\log p)' \cdot f = \int p (\log p)' f$$

$$= \int p \frac{p'}{p} \cdot f$$

$$= \int p' \cdot f$$

$$= - \int f' \cdot p \quad \text{(IBP)}$$

$$= -\mathbb{E} f'$$

Note: the gradient of $tr(\nabla_x s_\theta(x))$ is expensive to compute.

Two alternative losses:

1) Denoising score matching

$$\mathbb{E}_{q_\sigma(\tilde{x}|x) p(x)} \| s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x) \|^2$$

where $q_\sigma(\tilde{x}|x)$ is some noise process

(note: learns score of $\log q_\sigma$, not $\log p$).

2) sliced score matching.

$$\mathbb{E}_{v} \, \mathbb{E}_{x \sim p}\left[v^T \nabla_x s_\theta(x) \, v + \frac{1}{2} \| s_\theta(x) \|^2 \right]$$

Issue with naïve score matching.

→ if high-dimensional data sparsely supported, then bad estimation in regions of low probability.

→ LD can be tricky if modes mix slowly.

Noise-conditional score networks.

Consider family of perturbed data distributions

$$p_\sigma(x) = \int p(t) \, \mathcal{N}(x \mid t, \sigma^2)$$

Will learn family of models $s_\theta(x, \sigma)$

via score matching.

Consider denoising score matching objective

$$\ell(\theta, \sigma) = \mathbb{E}_{x\sim p}\,\mathbb{E}_{\tilde{x}\sim q_\sigma(\tilde{x}|x)} \left\| s_\theta(\tilde{x}, \sigma) - \frac{\tilde{x} - x}{\sigma} \right\|^2$$

For all noise levels, then have

$$\sum_{i=1}^{L} \ell(\theta, \sigma_i) \cdot \lambda(\sigma_i)$$

weights.

typical choice $\lambda(\sigma) = \sigma^2$.

Note: almost same loss as diffusion models!

Sampling via annealed LD

for $i = 1:L$

for $t = 1:T$

$$\tilde{x}_{i,t} = \tilde{x}_{i,t-1} + \frac{\alpha_i}{2} s_\theta(\tilde{x}_{i,t-1}, \sigma_i) + \sqrt{\alpha_i}\, z_{i,t}$$

$$\tilde{x}_{i+1,0} = \tilde{x}_{i,T}$$

Note: simple modification to do inpainting by projecting onto observed at each step.

A unified view through SDEs

"score-based generative modeling through stochastic differential equations".

Song et al. 2021.

Consider forward diffusion SDE

$$dX = f(x,t)\,dt + g(t)\,dW$$

then have backwards SDE

$$dX = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right]dt + g(t)\,dW$$

To reverse SDE, we must thus perform score matching

$$\theta^* = \underset{\theta}{\arg\min}\ \mathbb{E}_{t\sim U[0,T]}\ \lambda(t)\ \mathbb{E}_{x_0}\mathbb{E}_{x_t|x_0}\ \| s_\theta(x_t,t) - \nabla_{x_t}\log p_{0t}(x_t|x_0)\|^2$$

What SDE to choose?

Score matching.

$$x_i = x_0 + \sigma_i z_i$$

$$= x_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\ z_{i-1}'$$

$$dx = \sqrt{(\sigma^2)'}\ dW$$

Diffusion

$$x_i = \sqrt{1-\beta_i}\ x_{i-1} + \sqrt{\beta_i}\ z_{i-1}$$

$$dx = -\frac{1}{2}\beta(t)\, x\, dt + \sqrt{\beta(t)}\ dW$$

Sampling the reverse SDE.

Choose standard discretization

↳ similar to ancestral sampling used in diffusion models

Additionally, can adjust marginal distribution at each step.
↳ e.g. LD as in score matching models.

Probability Flow

SDE → ODE with same marginals.

$$dx = \left[f(x,t) - \frac{1}{2} g^2(t) \nabla_x \log p_t(x)\right] dt$$

Connection to continuous normalizing flow.
↳ enables explicit evaluation of likelihood.

$$dz = f(z,t)\, dt$$

$$\frac{\partial \log p(z(t))}{\partial t} - \operatorname{tr} \frac{df}{dz}$$

Controllability

$$dx = \left[f(x,t) - g(t)^2 \left[\nabla_x \log p_t(x) + \nabla_x \log p_t(y|x)\right]\right] dt + g(t)\, d\bar{w}.$$