Diffusion Models

Agenda

DDPM

- DDPM
  - Score Matching
  - VAE
  - SDE
- Conditional DDPM
  - Classifier Guidance
  - Classifier-Free Guidance

Sampling: Langevin Dynamics

p(x) — unnormalized probability density.

$$x_{t+1} = x_t + \eta \left(\frac{1}{2} \frac{\partial}{\partial x} \log p(x) + \frac{\epsilon_t}{\sqrt{\eta}} \right), \quad \epsilon_t \sim N(0, I).$$

Problems:

- We don't have $p(x)$, only samples $x^{(i)}$, $i = 1, \dots, N$.
- Burn-in can be too long: $x_0 \to x_1 \to x_2 \to \dots$

Connection to SGD:
`https://francisbach.com/gradient-flows/`

Diffusion

Consider the following process

$$x_0 \to x_1 \to x_2 \to \ldots \to x_T$$

$$p(x) = p_0(x), \ \ldots, \ p_T(x) = N(0, I)$$

Example (variance preserving). For $0 < \beta_1 \leq \beta_2 \leq \ldots \leq \beta_T \ll 1$

$$x_{t+1} = \sqrt{1 - \beta_{t+1}} x_t + \sqrt{\beta_{t+1}} \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N(0, I).$$

Lemma

$q(x_t | x_0) = N\left(\sqrt{\overline{\alpha_t}} x_0, (1 - \overline{\alpha_t}) I\right)$, *where* $\alpha_s = 1 - \beta_s$, $\overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s$

Diffusion

Lemma

$$q(x_t|x_0) = N\left(\sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I\right), \ \textit{where}\ \alpha_s = 1-\beta_s,\ \overline{\alpha_t} = \prod_{s=1}^{t}\alpha_s$$

$$x_{t+1} = \sqrt{1-\beta_{t+1}}x_t + \sqrt{\beta_{t+1}}\varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N(0, I).$$

is equivalent to $q(x_{t+1}|x_t) = N\left(\sqrt{1-\beta_{t-1}}x_t, \beta_{t+1}I\right)$.

$$s = 1:\ \ x_1 = \sqrt{1-\beta_1}x_0\sqrt{\beta_1}\varepsilon_1 = \sqrt{\overline{\alpha_1}}x_0 + \sqrt{1-\overline{\alpha_t}}\varepsilon_1.$$

$$\begin{aligned} s = t+1:\ \ x_{t+1} &= \sqrt{1-\beta_{t+1}}x_t + \sqrt{\beta_{t+1}}\varepsilon_{t+1} \\ &= \sqrt{\alpha_{t+1}}\left(\sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\varepsilon_t\right) + \sqrt{1-\alpha_{t+1}}\varepsilon_{t+1} \\ &= \sqrt{\overline{\alpha_{t+1}}}x_0 + \sqrt{\alpha_{t+1}-\overline{\alpha_{t+1}}}\varepsilon_t + \sqrt{1-\alpha_{t+1}}\varepsilon_{t+1} \end{aligned}$$

$$E[x_{t+1}|x_0] = \sqrt{\overline{\alpha_{t+1}}}x_0$$
$$Var[x_{t+1}|x_0] = \alpha_{t+1} - \overline{\alpha_{t+1}} + 1 - \alpha_{t+1} = 1 - \overline{\alpha_{t+1}}.$$

Denoising

$x_T^{(M)} \sim N(0, I)$;
for $t = T-1 \ldots 0$ **do**
  $x_t^{(1)} = x_{t+1}^{(M)}$;
  for *m=1...M-1* **do**
    $x_t^{(m+1)} = x_t^{(m)} + \eta \left(\frac{1}{2} \frac{\partial}{\partial x} \log p_t(x_t^{(m)}) + \frac{\epsilon_t}{\sqrt{\eta}} \right), \; \epsilon_t \sim N(0, I)$.
  end
end
return $x_0^{(M)}$

Score Function

$$s_\theta(x, t) \approx \frac{\partial}{\partial x} \log p_t(x)$$

$$\theta = \operatorname*{argmin}_\theta \sum_{t=1}^{T-1} \int p_t(x) \left\| s_\theta(x, t) - \frac{\partial}{\partial x} \log p_t(x) \right\|^2 dx.$$

Score Function

$$\int p_t(x)\left\|s_\theta(x,t)-\frac{\partial}{\partial x}\log p_t(x)\right\|^2 dx = \int p_t(x)\left[s_\theta^T s_\theta - 2s_\theta^T \frac{\partial}{\partial x}\log p_t(x)\right] dx + con$$

$$= \int p_t(x)s_\theta^T s_\theta\, dx - 2\int p_t(x)s_\theta^T \frac{\partial p_t(x)}{\partial x}\frac{1}{p_t(x)} dx + const$$

$$p_t(x)=\int q_t(x|x_0)p_0(x_0)dx_0$$

$$= \int\int q_t(x|x_0)p_0(x_0)s_\theta^T s_\theta\, dx dx_0 - 2\int s_\theta^T \frac{\partial}{\partial x}\int q_t(x|x_0)p_0(x_0)dx_0 dx + const$$

$$= \iint q_t(x|x_0)p_0(x_0)s_\theta^T s_\theta\, dx dx_0 - 2\iint s_\theta^T \frac{\partial}{\partial x} q_t(x|x_0)p_0(x_0)dx_0 dx + const$$

$$= \iint q_t(x|x_0)p_0(x_0)s_\theta^T s_\theta\, dx dx_0 - 2\iint s_\theta^T q_t(x|x_0)\frac{\partial}{\partial x}\log q_t(x|x_0)p_0(x_0)dx_0 dx + con$$

$$= \iint q_t(x|x_0)p_0(x_0)\left[s_\theta^T s_\theta - 2s_\theta^T\frac{\partial}{\partial x}\log q_t(x|x_0) \pm \left(\tfrac{\partial}{\partial x}\log q_t(x|x_0)\right)^T \tfrac{\partial}{\partial x}\log q_t(x|x_0)\right] dx_0 dx$$

$$= \iint q_t(x|x_0)p_0(x_0)\left\|s_\theta - \frac{\partial}{\partial x}\log q_t(x|x_0)\right\|^2 dx_0 dx + const$$

Latent Variables

$$
\begin{aligned}
\log p_0(x_0) &= \int q(x_1, \ldots, x_T | x_0) \log p_0(x_0) dx_1 \ldots dx_T \\
&= \int q(x_1, \ldots, x_T | x_0) \log \frac{p_0(x_0) q(x_1, \ldots, x_T | x_0)}{q(x_1, \ldots, x_T | x_0)} dx_1 \ldots dx_T \\
&= \int q(x_1, \ldots, x_T | x_0) \log \frac{p(x_0, x_1, \ldots, x_T) q(x_1, \ldots, x_T | x_0)}{p(x_1, \ldots, x_T | x_0) q(x_1, \ldots, x_T | x_0)} dx_1 \ldots dx_T \\
&= \int q(x_1, \ldots, x_T | x_0) \log \frac{p(x_0, x_1, \ldots, x_T)}{q(x_1, \ldots, x_T | x_0)} dx_1 \ldots dx_T \\
&\qquad + \int q(x_1, \ldots, x_T | x_0) \log \frac{q(x_1, \ldots, x_T | x_0)}{p(x_1, \ldots, x_T | x_0)} dx_1 \ldots dx_T \\
&= \mathcal{L} + KL(q(x_1, \ldots, x_T | x_0) || p(x_1, \ldots, x_T | x_0)) \geq \mathcal{L} \ (ELBO)
\end{aligned}
$$

ELBO estimation

$$
\begin{aligned}
\log p_\theta(x_0) &\geq \int q(x_{1:T}|x_0) \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} dx_{1:T} \\
&= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right] = \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})} \right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right] = \mathcal{L}_\theta .
\end{aligned}
$$

Minimize (negative) variational lower bound

$$
-\log p_\theta(x_0) \leq -\mathcal{L}_\theta \to min. \tag{1}
$$

Or with averaging over the batch

$$
-\frac{1}{N} \sum_{x_0} \log p_\theta(x_0) = -\mathbb{E}_{q(x_0)} \log p_\theta(x_0) \leq -\mathbb{E}_{q(x_0)} \mathcal{L}_\theta \to min. \tag{2}
$$

ELBO estimation

$$\mathcal{L}_\theta = \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right].$$

Using the Markovian property, one can rewrite the denominator as follows

$$q(x_t|x_{t-1}) = q(x_t|x_{t-1},\ x_0) = \frac{q(x_t,\ x_{t-1}|x_0)}{q(x_{t-1}|x_0)} = \frac{q(x_{t-1}|x_t,\ x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}.$$

Thus,

$$\begin{aligned}
\mathcal{L}_\theta &= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}{q(x_{t-1}|x_t,\ x_0)q(x_t|x_0)} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p(x_T)}{q(x_T|x_0)} + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,\ x_0)} + \log p_\theta(x_0|x_1) \right].
\end{aligned}$$

ELBO estimation

$$\mathcal{L}_\theta = \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_T)}{q(x_T|x_0)} + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,\ x_0)} + \log p_\theta(x_0|x_1)\right].$$

Calculate the second term:

$$\begin{aligned}
\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,\ x_0)}\right] &= \int q(x_{1:T}|x_0) \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,\ x_0)} dx_{1:T} \\
&= \int q(x_{t-1},\ x_t|x_0) \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,\ x_0)} dx_{t-1} dx_t \\
&= \int q(x_{t-1}|\ x_t,\ x_0) q(x_t|x_0) \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,\ x_0)} dx_{t-1} dx_t \\
&= \int q(x_t|x_0) dx_t \int q(x_{t-1}|\ x_t,\ x_0) \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,\ x_0)} dx_{t-1} \\
&= -\int q(x_t|x_0) D_{KL}\left(q(x_{t-1}|x_t,\ x_0) || p_\theta(x_{t-1}|x_t)\right) dx_t \\
&= -\mathbb{E}_{q(x_t|x_0)}\left[D_{KL}\left(q(x_{t-1}|x_t,\ x_0) || p_\theta(x_{t-1}|x_t)\right)\right].
\end{aligned}$$

ELBO estimation

To calculate $q(x_{t-1}|x_t,\ x_0)$ one can use the Bayes rule

$$q(x_{t-1}|x_t,\ x_0) = \frac{q(x_t|x_{t-1},\ x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} = N\left(x_{t-1}|\tilde{\mu}_t(x_t,\ x_0),\ \tilde{\beta}_t I\right), \tag{3}$$

with

$$\tilde{\mu}_t(x_t,\ x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \tag{4}$$

and

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \tag{5}$$

If we assume, that in reverse process $p_\theta(x_{t-1}|x_t) = q(x_{t-1}|x_t, x_\theta(x_t, t)$, the covariance is the same as in the forward process (3), i.e., $\Sigma_\theta(x_t,\ t) = \tilde{\beta}_t I$, then

$$D_{KL}\left(q(x_{t-1}|x_t,\ x_0)||p_\theta(x_{t-1}|x_t)\right) = \frac{1}{2\tilde{\beta}_t}\left\|\tilde{\mu}_t(x_t,\ x_0) - \mu_\theta(x_t, t)\right\|^2 + const. \tag{6}$$

ELBO estimation

$$\frac{1}{2\tilde{\beta}_t}\left\|\tilde{\mu}_t(x_t,\ x_0) - \mu_\theta(x_t, t)\right\|^2 = \frac{\sqrt{\overline{\alpha_{t-1}}}\beta_t}{2\tilde{\beta}_t(1-\overline{\alpha_t})} = \left\|x_0 - x_\theta(x_t, t)\right\|^2$$

$$= \frac{\beta_t^2}{2\tilde{\beta}_t\alpha_t(1-\bar{\alpha}_t)}\left\|\varepsilon_t - \varepsilon_\theta(x_t, t)\right\|^2.$$

$$\mathcal{L}_{\theta,\text{simplified}} = -\sum_{t=1}^{T}\left\|\varepsilon_t - \varepsilon_\theta(x_t, t)\right\|^2$$

DDPM

$$x_t = \sqrt{\overline{\alpha_1}} x_0 + \sqrt{1 - \overline{\alpha_t}} \varepsilon_t \tag{7}$$

Sampling algorithm:
$\hat{x}_T \sim N(0,\ I)$
for $t = T, \dots, 1$:
  $\epsilon_\theta(x_t, t)$
  $x_0^{(\theta)} = (7)$ with ϵ_θ
  $\hat{x}_{t-1} \sim q(x_{t-1} | x_t, x_0 = x_0^\theta)$
$\hat{x}_0$

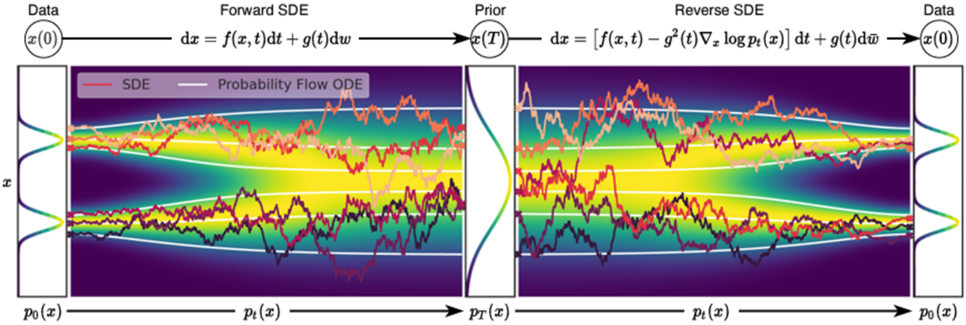
Training algorithm:
Take a batch of x_0
$t \sim U[1, \dots, T]$
  $\epsilon \sim N(0,\ I)$
  $\hat{x}_t = (7)$ with ϵ
  $\epsilon_\theta(x_t, t)$
$Loss(\epsilon_\theta, \epsilon) = \|\epsilon_\theta - \epsilon\|^2$

Connection with Score

$$x_t = \sqrt{\overline{\alpha_1}} x_0 + \sqrt{1 - \overline{\alpha_t}} \varepsilon_t, \qquad \varepsilon_t = \frac{x_t - \sqrt{\overline{\alpha_1}} x_0}{\sqrt{1 - \overline{\alpha_t}}}$$

$$\begin{aligned} \frac{\partial}{\partial x_t} \log q(x_t | x_0) &= \frac{\partial}{\partial x_t} \left(- \frac{(x_t - \sqrt{\overline{\alpha_1}} x_0)^2}{2(1 - \overline{\alpha_t})} \right) \\ &= - \frac{x_t - \sqrt{\overline{\alpha_1}} x_0}{1 - \overline{\alpha_t}} = - \frac{\varepsilon_t}{\sqrt{1 - \overline{\alpha_t}}} = s_\theta(x_t, t). \end{aligned}$$

Connection with SDE



Y. Song et all, Score-Based Generative Modeling through Stochastic Differential Equations https://arxiv.org/abs/2011.13456

Advantages of different approaches

Diffusion Model

Score Matching

1) *Connection* s_θ *and* ε_θ

2) *Classifier* – ~~*free*~~ *guidance*

VAE

1) *Different noise*

2) *Train decoder*

SDE

1) *DDIM*

2) $p_0(x)$

Conditional DM

Condition DM

Classifier Guidance

$$P(A|B) = \frac{P(AB)}{P(B)}$$

$$s_\theta(x) \approx \frac{\partial}{\partial x}\log p_t(x)$$

$$\frac{\partial}{\partial x}\log p_t(x|y) = \frac{\partial}{\partial x}\log\frac{p_t(y|x)p_t(x)}{p(y)}$$

$$= \lambda\underbrace{\frac{\partial}{\partial x}[\log p_t(y|x)]}_{\text{Classifier}} + \underbrace{\frac{\partial}{\partial x}\log p_t(x)}_{s_\theta} + const \qquad \lambda > 1$$

Conditional DM

Classifier – Free Guidance

$$(x,y)$$

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \| x_0 - x_\theta\left(x_t^{(i)}, t\right) \|^2 \rightarrow \min$$

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \| x_0 - x_\theta\left(x_t^{(i)}, t, y^{(i)}\right) \|^2 + \sum_{i=1}^{N} \sum_{t=1}^{T} \| x_0 - x_\theta\left(x_t^{(i)}, t, \varnothing\right) \|^2$$