Diffusion

samrudhdhi.rangrej

February 2024

1 Forward Process

The forward process, also known as the diffusion process, transforms clean image (x_0) to Gaussian noise image (X_T). The forward process (or encoding process) is *generally a fixed* Markov chain - it is not learnt (but it can also be learned). The forward process is an approximate posterior as described below.

$$\begin{aligned}
q(x_{1:T}|x_0) &= \prod_{t=1}^{T} q(x_t|x_{t-1}) && (1)\\
q(x_t|x_{t-1}) &= \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) && (2)\\
&= \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}) \quad \text{(based on eq 5)} && (3)
\end{aligned}$$

The forward process variance β_t can be learned by reparameterization or held constant as hyperparameters (i.e. non-learnable). When β_t is small reverse process also becomes Gaussian. One nice property of forward process is that we can find x_t in a closed form using the formula for sum of two Gaussian random variables, which leads to:

$$\begin{aligned}
q(x_t|x_0) &= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) && (4)\\
\alpha_t &= 1-\beta_t && (5)\\
\bar{\alpha}_t &= \prod_{s=1}^{t} \alpha_s && (6)
\end{aligned}$$

As mentioned earlier, we can derive eq 4 based on the sum of two Gaussian random variable - which is as follows.

$$\begin{aligned}
Z &= X \pm Y && (7)\\
p(X) &= \mathcal{N}(\mu_x, \sigma_x^2) && (8)\\
p(Y) &= \mathcal{N}(\mu_y, \sigma_y^2) && (9)\\
p(Z) &= \mathcal{N}(\mu_x \pm \mu_y, \sigma_x^2 + \sigma_y^2) && (10)
\end{aligned}$$

Let's write x_t as a noisy version of x_{t-1}.

$$\begin{aligned}
x_1 &= \sqrt{\alpha_1}x_0 + \sqrt{1-\alpha_1}\epsilon_0 \quad (\because \text{eq 3 and reparameterization trick}) && (11)\\
x_2 &= \sqrt{\alpha_2}x_1 + \sqrt{1-\alpha_2}\epsilon_1 && (12)\\
&= \sqrt{\alpha_2}(\sqrt{\alpha_1}x_0 + \sqrt{1-\alpha_1}\epsilon_0) + \sqrt{1-\alpha_2}\epsilon_1 && (13)\\
&= \sqrt{\alpha_2\alpha_1}x_0 + \sqrt{\alpha_2-\alpha_2\alpha_1}\epsilon_0 + \sqrt{1-\alpha_2}\epsilon_1 && (14)\\
&= \sqrt{\alpha_2\alpha_1}x_0 + \sqrt{\sqrt{\alpha_2-\alpha_2\alpha_1}^2 + \sqrt{1-\alpha_2}^2}\epsilon_1^* \quad (\because \text{sum of Gaussian random variables}) && (15)\\
&= \sqrt{\alpha_2\alpha_1}x_0 + \sqrt{\alpha_2-\alpha_2\alpha_1+1-\alpha_2}\epsilon_1^* && (16)\\
&= \sqrt{\alpha_2\alpha_1}x_0 + \sqrt{1-\alpha_2\alpha_1}\epsilon_1^* && (17)
\end{aligned}$$

Repeat recursively to get,

$$\begin{aligned}
x_t &= \sqrt{\alpha_t \dots \alpha_1}x_0 + \sqrt{1-\alpha_t \dots \alpha_1}\epsilon && (18)\\
q(x_t|x_0) &= \mathcal{N}(x_t; \sqrt{\alpha_t \dots \alpha_1}x_0, (1-\alpha_t \dots \alpha_1)\mathbf{I}) && (19)\\
&= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) && (20)
\end{aligned}$$

2 Reverse Process

The reverse process transforms Gaussian noise image (X_T) to clean image (x_0). The reverse process is a Markov chain with *learned* Gaussian transitions. It is a joint distribution as described next.

$$
\begin{aligned}
p(x_{0:T}) &= p(x_T)\prod_{t=1}^{T} p(x_{t-1}|x_t) \qquad (21)\\
p(x_T) &= \mathcal{N}(x_T; 0, \mathbf{I}) \qquad (22)\\
p(x_{t-1}|x_t) &= \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \mathbf{\Sigma}_\theta(x_t, t)) \qquad (23)
\end{aligned}
$$

3 ELBO

$$
L = \underbrace{\mathbb{KL}\big[q(x_T|x_0)||p(x_T)\big]}_{L_T} + \sum_{t>1} \underbrace{\mathbb{KL}\big[q(x_{t-1}|x_t, x_0)||p(x_{t-1}|x_t)\big]}_{L_{t-1}} - \underbrace{\mathbb{E}_q\big[log\ p(x_0|x_1)\big]}_{L_0} \qquad (24)
$$

ELBO can be derived as follows.

$$
\begin{aligned}
-log\big(p(x_0)\big) &= -log\Big(\int p(x_{0:T})\ dx_{1:T}\Big)\Big) \qquad (25)\\
&= -log\Big(\int \frac{p(x_{0:T})}{q(x_{1:T}|x_0)} q(x_{1:T}|x_0)\ dx_{1:T}\Big)\Big) \qquad (26)\\
&= -log\Big[\mathbb{E}_q \frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\Big] \qquad (27)\\
&\geq -\mathbb{E}_q\Big[log\Big\{\frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\Big\}\Big] \qquad (28)\\
&= -\mathbb{E}_q\Big[log\Big\{p(x_T)\prod_{t>1}\frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1}, x_0)}\ \frac{p(x_0|x_1)}{q(x_1|x_0)}\Big\}\Big] \qquad (29)\\
&= -\mathbb{E}_q\Big[log\Big\{p(x_T)\prod_{t>1}\frac{p(x_{t-1}|x_t)q(x_{t-1}|x_0)}{q(x_{t-1}|x_t, x_0)q(x_t|x_0)}\ \frac{p(x_0|x_1)}{q(x_1|x_0)}\Big\}\Big] \quad (\because \text{Bayes' rule}) \qquad (30)\\
&= -\mathbb{E}_q\Big[log\Big\{p(x_T)\prod_{t>1}\frac{p(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\prod_{t>1}\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}\ \frac{p(x_0|x_1)}{q(x_1|x_0)}\Big\}\Big] \qquad (31)\\
&= -\mathbb{E}_q\Big[log\Big\{p(x_T)\prod_{t>1}\frac{p(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\ \frac{q(x_1|x_0)}{q(x_T|x_0)}\ \frac{p(x_0|x_1)}{q(x_1|x_0)}\Big\}\Big] \qquad (32)\\
&= -\mathbb{E}_q\Big[log\Big\{\frac{p(x_T)}{q(x_T|x_0)}\prod_{t>1}\frac{p(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\ p(x_0|x_1)\Big\}\Big] \qquad (33)\\
&= \mathbb{KL}\big[q(x_T|x_0)||p(x_T)\big] + \sum_{t>1}\mathbb{KL}\big[q(x_{t-1}|x_t, x_0)||p(x_{t-1}|x_t)\big] - \mathbb{E}_q\big[log\ p(x_0|x_1)\big] \qquad (34)
\end{aligned}
$$

3.1 Forward process and L_T

When β_t are fixed (non-learnable), $q(x_T|x_0)$ has no learnable parameters i.e. L_T is constant during training and can be ignored.

3.2 Reverse process and L_{t-1}

L_{t-1} is a KL-divergence between $q(x_{t-1}|x_t, x_0)$ and $p(x_{t-1}|x_t)$. $q(x_{t-1}|x_t, x_0)$ are tractable when conditioned on x_0 and is defined as follows (derivation in Sec A).

$$
\begin{aligned}
q(x_{t-1}|x_t, x_0) &= \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbf{I}) \qquad (35)\\
\tilde{\mu}_t(x_t, x_0) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 \qquad (36)\\
\tilde{\beta} &= \frac{(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\beta_t = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)} \qquad (37)
\end{aligned}
$$

As shown in eq 23, $p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \mathbf{\Sigma}_\theta(x_t, t))$ is learnable. Though, since variance of forward process ($\tilde{\beta}_t\mathbf{I}$) is fixed and independent of x_0, we can set variance of reverse process $\mathbf{\Sigma}_\theta(x_t, t)$ to be a constant too [1], i.e. $\mathbf{\Sigma}_\theta(x_t, t) = \tilde{\beta}_t\mathbf{I}$. However, we must learn mean of reverse process $\mu_\theta(x_t, t)$ since mean of forward process depends on x_0.

Following the formula of KL-divergence between two Gaussian distributions, we get,

$$\begin{aligned}
L_{t-1} &= \mathbb{KL}\big[q(x_{t-1}|x_t, x_0)||p(x_{t-1}|x_t)\big] \qquad (38)\\
&= \frac{1}{2}\big[\log\frac{|\Sigma_\theta(x_t, t)|}{\tilde{\beta}_t} - k + \big(\mu_\theta(x_t, x_0) - \tilde{\mu}(x_t, x_0)\big)^T \Sigma_\theta(x_t, t)^{-1}\big(\mu_\theta(x_t, x_0) - \tilde{\mu}(x_t, x_0)\big) + tr\{\tilde{\beta}_t\Sigma_\theta(x_t, t)^{-1}\}\big] \qquad (39)\\
&= \frac{1}{2}\big[\big(\mu_\theta(x_t, x_0) - \tilde{\mu}(x_t, x_0)\big)^T \Sigma_\theta(x_t, t)^{-1}\big(\mu_\theta(x_t, x_0) - \tilde{\mu}(x_t, x_0)\big)\big] + C' \qquad (\because \Sigma_\theta \text{ is held fixed}) \qquad (40)\\
&= \frac{1}{2\tilde{\beta}_t}||\mu_\theta(x_t, x_0) - \tilde{\mu}(x_t, x_0)||^2 + C \qquad (\text{when } \Sigma_\theta = \tilde{\beta}_t\mathbf{I}) \qquad (41)
\end{aligned}$$

Let's simplify $\tilde{\mu}(x_t, x_0)$ using reparameterization trick,

$$\begin{aligned}
x_t &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \qquad (42)\\
x_0 &= \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon) \qquad (43)\\
\tilde{\mu}_t(x_t, x_0) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 \qquad (44)\\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon) \qquad (45)\\
&= \frac{1}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}(\alpha_t - \alpha_t\bar{\alpha}_{t-1} + 1 - \alpha_t)x_t - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon \qquad (46)\\
&= \frac{1}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}(1-\bar{\alpha}_t)x_t - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon \qquad (47)\\
&= \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon) \qquad (48)
\end{aligned}$$

Hence, based on eq 41, $\mu_\theta(x_t, x_0)$ should predict $\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon)$. Since x_t is input to the mode, we can chose to parameterize $\mu_\theta(x_t, x_0)$ as follows.

$$\mu_\theta(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, x_0)) \qquad (49)$$

Thus, eq 41 becomes,

$$\begin{aligned}
L_{t-1} &\propto ||\epsilon - \epsilon_\theta(x_t, x_0)||^2 \qquad (50)\\
&= ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon)||^2 \qquad \text{(final simplified loss function)} \qquad (51)
\end{aligned}$$

In eq 50, we ignore the coefficient which depends on β_t, α_t, $\bar{\alpha}_t$. Due to ignoring the weight factor, finally, L_t would have higher weight than $L_{t'}$ where $t > t'$. This turns out to be a desirable property for model training. The model can focus more on denoising at large t which generally has more noise than small t.

3.3 Data scaling, reverse process decoder and L_0

Assuming x_0 has discrete values (e.g. integer image pixels in {0, 1, . . . , 255}) normalized to [-1,1], we may define $p(x_0|x_1)$ as a mapping function that discretizes x_1 into x_0. When in the forward process we create x_1 by first mapping x_0 to real values and then adding a small noise $< \frac{1}{255}$ (another reason why β_t starts with a small value!), during reverse process we can map scaled x_1 directly to nearest integer. This is a parameter-less operation. Hence, L_0 has no learnable parameters and is constant.

[1] In the ddpm paper, another choice is $\mathbf{\Sigma}_\theta(x_t, t) = \beta_t\mathbf{I}$. They mention that $\beta_t\mathbf{I}$ is the optimal when $x_0 \sim \mathcal{N}(0, \mathbf{I})$ and $\tilde{\beta}_t\mathbf{I}$ is optimal when x_0 is set deterministically to one point. Consequently, the above choices correspond to upper and lower bounds on reverse entropy for *data with coordinatewise unit variance.*

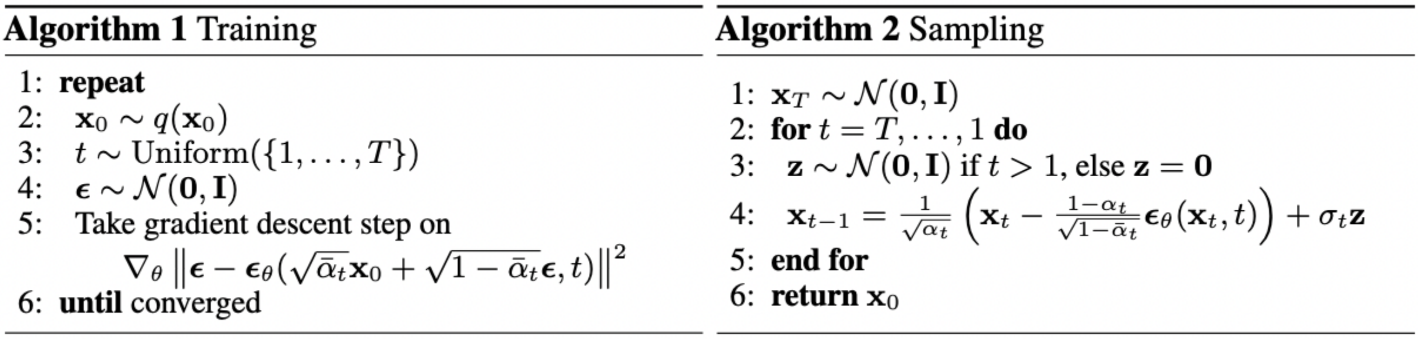
4 Sampling

Let's apply reparameterization trick to eq 23,

$$x_{t-1} = \mu_\theta(x_t, t) + \mathbf{\Sigma}_\theta(x_t, t)^{-\frac{1}{2}} z; \qquad \text{where} \;\; z \sim \mathcal{N}(0, I) \tag{52}$$

$$= \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, x_0)) + \sigma_t z; \qquad \text{where} \;\; \sigma_t^2 = \beta_t \;\; \text{or} \;\; \tilde{\beta}_t \tag{53}$$

Above we use parameterization of $\mu_\theta(x_t, t)$ defined in eq 49.



Algorithm 1 Training

```
repeat
    $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
    $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Take gradient descent step on
        $\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
until converged
```

Algorithm 2 Sampling

```
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ do
    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$
```

4.1 DDIM

Denoising diffusion implicit model (DDIM) argues that **simplified** ELBO does not depend on joint distribution $q(x_{1:T}|x_0)$ but only on marginal $q(x_t|x_0)$. Hence, one could come up with a forward process which is not Markovian yet have a same marginal as DDPM. Under this argument, they chose a specific formulation which gives same marginal as eq 3. The marginal and the reconstruction formula under this formulation is as follows.

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}}, \sigma_t^2 \mathbf{I}) \tag{54}$$

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, x_0)}{\sqrt{\bar{\alpha}_t}} \right)}_{\text{predicted } x_0} + \underbrace{\sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t, x_0)}_{\text{direction pointing to } x_t} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}} \tag{55}$$

Different values of σ results in different generative process while using the same diffusion model ϵ_θ.

$$\sigma_t = \begin{cases} \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}} \sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}} & \to \text{stochastic DDPM} \\ 0; \;\; \forall t & \to \text{deterministic} \end{cases}$$

One benefit of this formulation is that now we can draw samples at larger intervals.

$$x_{t-\Delta t} = \sqrt{\bar{\alpha}_{t-\Delta t}} \underbrace{\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, x_0)}{\sqrt{\bar{\alpha}_t}} \right)}_{\text{predicted } x_0} + \underbrace{\sqrt{1 - \bar{\alpha}_{t-\Delta t} - \sigma_t^2} \cdot \epsilon_\theta(x_t, x_0)}_{\text{direction pointing to } x_t} + \underbrace{\sigma_t \epsilon_t}_{randomnoise} \tag{56}$$

$$\sigma_t = \sqrt{\frac{1-\bar{\alpha}_{t-\Delta t}}{1-\bar{\alpha}_t}} \sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-\Delta t}}} \qquad \text{(for stochastic DDPM)} \tag{57}$$

5 Diffusion Guidance

$$p(x_t|y) = \frac{p(y|x_t)p(x_t)}{p(y)} \tag{58}$$

$$\log p(x_t|y) = \log p(y|x_t) + \log p(x_t) - \log p(y) \tag{59}$$

$$\nabla_{x_t} \log p(x_t|y) = \nabla_{x_t} \log p(y|x_t) + \nabla_{x_t} \log p(x_t) \tag{60}$$

A temperature scaled version is as follows.

$$\begin{align} \nabla_{x_t} \log p_\gamma(x_t|y) &= \gamma \nabla_{x_t} \log p(y|x_t) + \nabla_{x_t} \log p(x_t) \tag{61} \\ p_\gamma(x_t|y) &\propto p(y|x_t)^\gamma p(x_t) \tag{62} \end{align}$$

Higher value of γ makes $p(y|x_t)$ more peaky, increasing influence of a conditional guidance.

5.1 Classifier Guidance

Assuming backward is estimating forward process,

$$\begin{align} p(x_t|x_0) &= \mathcal{N}(x_t; \mu_\theta(x_0), \mathbf{\Sigma}_\theta(x_0)) \tag{63} \\ &= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{64} \\ & \tag{65} \end{align}$$

Let's find $\nabla_{x_t} \log p(x_t)$ to plug in eq 60.

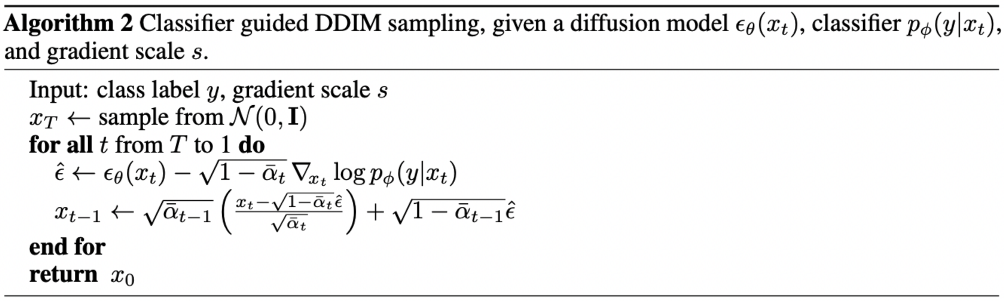
$$\begin{align} p(x_t) &= \frac{p(x_t|x_0)p(x_0)}{p(x_0|x_t)} \tag{66} \\ &= \frac{p(x_t|x_0)p(x_0)}{p(x_0)} \tag{67} \\ &= p(x_t|x_0) \tag{68} \\ \nabla_{x_t} \log p(x_t) &= \nabla_{x_t} \log p(x_t|x_0) \tag{69} \\ &= \nabla_{x_t} \Big[- \frac{(x_t - \sqrt{\bar{\alpha}_t} x_0)^2}{2(1-\bar{\alpha}_t)} + const \Big] \qquad (\because \text{ formula of a log of Gaussian}) \tag{70} \\ &= - \frac{(x_t - \sqrt{\bar{\alpha}_t} x_0)}{(1-\bar{\alpha}_t)} \tag{71} \\ &= - \frac{(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{(1-\bar{\alpha}_t)} \epsilon_\theta(x_t, t) - \sqrt{\bar{\alpha}_t} x_0)}{(1-\bar{\alpha}_t)} \qquad (\because \text{ Reparameterization trick}) \tag{72} \\ &= - \frac{\epsilon_\theta(x_t, t)}{\sqrt{1-\bar{\alpha}_t}} \tag{73} \end{align}$$

Thus, based on eq 60 and 73,

$$\nabla_{x_t} \log p(x_t|y) = \nabla_{x_t} \log p(y|x_t) - \frac{\epsilon_\theta(x_t, t)}{\sqrt{1-\bar{\alpha}_t}} \tag{74}$$

Hence, we can define a new $\hat{\epsilon}$.

$$\hat{\epsilon}(x_t, t) = \epsilon_\theta(x_t, t) - \sqrt{1-\bar{\alpha}_t} \nabla_{x_t} \log p(y|x_t) \qquad \text{(based on eq 73)} \tag{75}$$



Algorithm 2 Classifier guided DDIM sampling, given a diffusion model $\epsilon_\theta(x_t)$, classifier $p_\phi(y|x_t)$, and gradient scale s.

```
Input: class label $y$, gradient scale $s$
$x_T \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
for all $t$ from $T$ to 1 do
    $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t) - \sqrt{1-\bar{\alpha}_t} \nabla_{x_t} \log p_\phi(y|x_t)$
    $x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left( \frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}} \hat{\epsilon}$
end for
return $x_0$
```

In classifier guidance, we learn a classifier $p(y|x_t)$ that predicts y given x_t $\forall t$ (i.e. all noise levels), which is difficult. Hence, there are practical limitation of classifier guidance. This is where classifier free guidance wins. However, classifier guidance is still useful because it does not require conditional diffusion model. It is capable of generating conditional samples from unconditional diffusion model!

5.2 Classifier-free Guidance

Based on 60,

$$\nabla_{x_t} \log p(y|x_t) = \nabla_{x_t} \log p(x_t|y) - \nabla_{x_t} \log p(x_t) \tag{76}$$

Let's replace above formula in eq 61.

$$\begin{align} \nabla_{x_t} \log p_\gamma(x_t|y) &= \gamma(\nabla_{x_t} \log p(x_t|y) - \nabla_{x_t} \log p(x_t)) + \nabla_{x_t} \log p(x_t) \tag{77} \\ &= \gamma \nabla_{x_t} \log p(x_t|y) + (1-\gamma)\nabla_{x_t} \log p(x_t) \tag{78} \end{align}$$

Above is a barycentric combination of the conditional and the unconditional score functions. For $\gamma = 0$, we recover the unconditional model, and for $\gamma = 1$ we get the standard conditional model. We can increase the quality of the generated sample by increasing influence of conditional term i.e. $\gamma > 1$.

Classifier-free guidance does not require training a separate classifier. Instead, we train a conditional diffusion model, with conditioning dropout i.e. the conditioning information is dropped for some iterations. The resulting model works both as conditional and unconditional model depending on whether the conditional input is provided or not. Final sampling is done as follows.

Hence, we can define a new $\hat{\epsilon}$.

$$\hat{\epsilon}(x_t, t) = \gamma \epsilon_\theta(x_t, t, c) - (1-\gamma)\epsilon_\theta(x_t, t, \phi) \tag{79}$$

Sampling algorithm is same as classifier guided DDIM, except $\hat{\epsilon}$ is computed using above formula.

The main reason why classifier-free guidance works better than classifier guidance is that in former we construct the "classifier" from a generative model. While standard classifiers can take shortcuts and ignore most of the input while still obtaining competitive classification results, generative models do not have such luxury. This makes the resulting gradient much more robust. As a bonus, we only have to train a single (generative) model, and conditioning dropout is trivial to implement.

A Derivation of $q(x_{t-1}|x_t, x_0)$

$$\begin{align} q(x_{t-1}|x_t, x_0) &= \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{80} \\ &= \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{81} \\ &= \frac{\mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})\mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})} \tag{82} \\ &\propto \exp\Big\{-\frac{1}{2}\Big[\frac{(x_t - \sqrt{\alpha_t}x_{t-1})^2}{(1-\alpha_t)} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0)^2}{(1-\bar{\alpha}_{t-1})} - \frac{(x_t - \sqrt{\bar{\alpha}_t}x_0)^2}{(1-\bar{\alpha}_t)}\Big]\Big\} \tag{83} \\ &= \exp\Big\{-\frac{1}{2}\Big[\frac{(x_t^2 - 2\sqrt{\alpha_t}x_t x_{t-1} + \alpha_t x_{t-1}^2)}{(1-\alpha_t)} + \frac{(x_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}x_{t-1}x_0 + \bar{\alpha}_{t-1}x_0^2)}{(1-\bar{\alpha}_{t-1})} \tag{84} \\ &\qquad - \frac{(x_t^2 - 2\sqrt{\bar{\alpha}_t}x_t x_0 + \bar{\alpha}_t x_0^2)}{(1-\bar{\alpha}_t)}\Big]\Big\} \tag{85} \\ &= \exp\Big\{-\frac{1}{2}\Big[\frac{x_t^2}{(1-\alpha_t)} - \frac{2\sqrt{\alpha_t}x_t x_{t-1}}{(1-\alpha_t)} + \frac{\alpha_t x_{t-1}^2}{(1-\alpha_t)} + \frac{x_{t-1}^2}{(1-\bar{\alpha}_{t-1})} - \frac{2\sqrt{\bar{\alpha}_{t-1}}x_{t-1}x_0}{(1-\bar{\alpha}_{t-1})} + \frac{\bar{\alpha}_{t-1}x_0^2}{(1-\bar{\alpha}_{t-1})} \tag{86} \\ &\qquad - \frac{x_t^2}{(1-\bar{\alpha}_t)} + \frac{2\sqrt{\bar{\alpha}_t}x_t x_0}{(1-\bar{\alpha}_t)} - \frac{\bar{\alpha}_t x_0^2}{(1-\bar{\alpha}_t)}\Big]\Big\} \tag{87} \end{align}$$

Let's focus on blue terms.

$$\begin{align}
\frac{\alpha_t x_{t-1}^2}{(1-\alpha_t)} + \frac{x_{t-1}^2}{(1-\bar{\alpha}_{t-1})} &= \frac{(\alpha_t(1-\bar{\alpha}_{t-1}) + (1-\alpha_t))}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})} x_{t-1}^2 \tag{88}\\
&= \frac{(\alpha_t - \alpha_t\bar{\alpha}_{t-1} + 1 - \alpha_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})} x_{t-1}^2 \tag{89}\\
&= \frac{(\alpha_t - \bar{\alpha}_t + 1 - \alpha_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})} x_{t-1}^2 \qquad (\because \alpha_t\bar{\alpha}_{t-1} = \bar{\alpha}_t) \tag{90}\\
&= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})} x_{t-1}^2 \tag{91}\\
&= \frac{1}{\beta_t} x_{t-1}^2 \tag{92}
\end{align}$$

Let's focus on red terms.

$$\begin{align}
-\frac{2\sqrt{\alpha_t} x_t x_{t-1}}{(1-\alpha_t)} - \frac{2\sqrt{\bar{\alpha}_{t-1}} x_{t-1} x_0}{(1-\bar{\alpha}_{t-1})} &= -2\left(\frac{\sqrt{\alpha_t} x_t}{(1-\alpha_t)} + \frac{\sqrt{\bar{\alpha}_{t-1}} x_0}{(1-\bar{\alpha}_{t-1})}\right) x_{t-1} \tag{93}\\
&= -2\left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) x_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})} + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t) x_0}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right) x_{t-1} \tag{94}\\
&= -2\frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) x_t}{(1-\bar{\alpha}_t)} + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t) x_0}{(1-\bar{\alpha}_t)}\right) x_{t-1} \tag{95}\\
&= -2\frac{1}{\beta_t}(\tilde{\mu}(x_t, x_0) x_{t-1}) \tag{96}
\end{align}$$

Let's focus on green terms.

$$\begin{align}
\frac{x_t^2}{(1-\alpha_t)} - \frac{x_t^2}{(1-\bar{\alpha}_t)} &= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{(1-\bar{\alpha}_t)(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)^2} - \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)^2}\Big) x_t^2 \tag{97}\\
&= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{1 - \bar{\alpha}_{t-1} - \bar{\alpha}_t + \bar{\alpha}_t\bar{\alpha}_{t-1} - 1 + \bar{\alpha}_{t-1} + \alpha_t - \alpha_t\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2}\Big) x_t^2 \tag{98}\\
&= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{-\bar{\alpha}_t + \bar{\alpha}_t\bar{\alpha}_{t-1} + \alpha_t - \alpha_t\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2}\Big) x_t^2 \tag{99}\\
&= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{-\alpha_t\bar{\alpha}_{t-1} + \alpha_t\bar{\alpha}_{t-1}\bar{\alpha}_{t-1} + \alpha_t - \alpha_t\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2}\Big) x_t^2 \qquad (\because \alpha_t\bar{\alpha}_{t-1} = \bar{\alpha}_t) \tag{100}\\
&= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{\alpha_t(-\bar{\alpha}_{t-1} + \bar{\alpha}_{t-1}\bar{\alpha}_{t-1} + 1 - \bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)^2}\Big) x_t^2 \tag{101}\\
&= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{\alpha_t(1 - 2\bar{\alpha}_{t-1} + \bar{\alpha}_{t-1}^2)}{(1-\bar{\alpha}_t)^2}\Big) x_t^2 \tag{102}\\
&= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{\alpha_t(1-\bar{\alpha}_{t-1})^2}{(1-\bar{\alpha}_t)^2}\Big) x_t^2 \tag{103}\\
& \tag{104}
\end{align}$$

Let's focus on yellow terms.

$$\frac{\bar{\alpha}_{t-1}x_0^2}{(1-\bar{\alpha}_{t-1})} - \frac{\bar{\alpha}_t x_0^2}{(1-\bar{\alpha}_t)} = \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{(1-\alpha_t)(1-\bar{\alpha}_t)\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2} - \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})\bar{\alpha}_t}{(1-\bar{\alpha}_t)^2}\Big)x_0^2 \tag{105}$$

$$= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{(1-\alpha_t)(1-\bar{\alpha}_t)\bar{\alpha}_{t-1} - (1-\alpha_t)(1-\bar{\alpha}_{t-1})\bar{\alpha}_{t-1}\alpha_t}{(1-\bar{\alpha}_t)^2}\Big)x_0^2 \quad (\because \alpha_t\bar{\alpha}_{t-1} = \bar{\alpha}_t) \tag{106}$$

$$= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{\bar{\alpha}_{t-1}(1-\alpha_t)((1-\bar{\alpha}_t)-(1-\bar{\alpha}_{t-1})\alpha_t)}{(1-\bar{\alpha}_t)^2}\Big)x_0^2 \tag{107}$$

$$= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{\bar{\alpha}_{t-1}(1-\alpha_t)(1-\bar{\alpha}_t-\alpha_t-\bar{\alpha}_{t-1}\alpha_t)}{(1-\bar{\alpha}_t)^2}\Big)x_0^2 \tag{108}$$

$$= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{\bar{\alpha}_{t-1}(1-\alpha_t)(1-\bar{\alpha}_t-\alpha_t+\bar{\alpha}_t)}{(1-\bar{\alpha}_t)^2}\Big)x_0^2 \quad (\because \alpha_t\bar{\alpha}_{t-1} = \bar{\alpha}_t) \tag{109}$$

$$= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2}\Big)x_0^2 \tag{110}$$

$$\tag{111}$$

Let's focus on black terms.

$$\frac{2\sqrt{\bar{\alpha}_t}x_tx_0}{(1-\bar{\alpha}_t)} = \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{2\sqrt{\bar{\alpha}_{t-1}\alpha_t}x_tx_0}{(1-\bar{\alpha}_t)}\,\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\Big) \quad (\because \alpha_t\bar{\alpha}_{t-1} = \bar{\alpha}_t) \tag{112}$$

$$= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big(\frac{2\sqrt{\bar{\alpha}_{t-1}\alpha_t}(1-\alpha_t)(1-\bar{\alpha}_{t-1})x_tx_0}{(1-\bar{\alpha}_t)^2}\Big) \tag{113}$$

$$\tag{114}$$

Let's put green, black and yellow terms together.

$$\frac{x_t^2}{(1-\alpha_t)} - \frac{x_t^2}{(1-\bar{\alpha}_t)} + \frac{2\sqrt{\bar{\alpha}_t}x_tx_0}{(1-\bar{\alpha}_t)} + \frac{\bar{\alpha}_{t-1}x_0^2}{(1-\bar{\alpha}_{t-1})} - \frac{\bar{\alpha}_t x_0^2}{(1-\bar{\alpha}_t)} \tag{115}$$

$$= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big[\Big(\frac{\alpha_t(1-\bar{\alpha}_{t-1})^2}{(1-\bar{\alpha}_t)^2}\Big)x_t^2 + \Big(\frac{2\sqrt{\bar{\alpha}_{t-1}\alpha_t}(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)^2}\Big)x_tx_0 + \Big(\frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2}\Big)x_0^2\Big] \tag{116}$$

$$= \frac{(1-\bar{\alpha}_t)}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big[\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0\Big]^2 \tag{117}$$

$$= \frac{1}{\beta_t}\tilde{\mu}(x_t, x_0)^2 \tag{118}$$

Let's put all together. Replacing the sum of eq 92, 96 and 118 in eq 87,

$$q(x_{t-1}|x_t, x_0) \propto \exp\Big\{-\frac{1}{2}\Big[\frac{1}{\beta_t}x_{t-1}^2 - 2\frac{1}{\beta_t}(\tilde{\mu}(x_t, x_0)x_{t-1}) + \frac{1}{\beta_t}\tilde{\mu}(x_t, x_0)^2\Big]\Big\} \tag{119}$$

$$= \exp\Big\{-\frac{1}{2\beta_t}\Big[x_{t-1}^2 - 2x_{t-1}\tilde{\mu}(x_t, x_0) + \tilde{\mu}(x_t, x_0)^2\Big]\Big\} \tag{120}$$

$$= \exp\Big\{-\frac{1}{2\beta_t}\Big(x_{t-1} - \tilde{\mu}(x_t, x_0)\Big)^2\Big\} \tag{121}$$

$$= \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \beta_t\mathbf{I}) \tag{122}$$