There is a quantity that captures the Signal-to-Noise ratio indeed, and it was named the same. It is related to the noise coefficients that we are free to choose in the $q(X_t|X_0)$ samples. To recall, $X_t = \alpha_t X_0 + \sigma_t \epsilon$.

Previously we had $\alpha_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t = \sqrt{1-\bar{\alpha}_t}$, where the $\bar{\alpha}_t$ is a sequence that approaches 0 as $t \to \infty$ or approximately 0 when $t \to T$.

What we require about $\alpha_t$ and $\sigma_t$ is that $\alpha_t \to 0$, $\sigma_t \to 1$ as $t \to T$ because we have to approach a standard normal $\mathcal{N}(0, I)$.

So, this Signal-to-Noise ratio, denoted $SNR(t) = \frac{\alpha_t^2}{\sigma_t^2}$. It is a measure of noise.

Although, if $SNR(T) = \frac{\alpha_T^2}{\sigma_T^2} = 0$ because $\alpha_t \to 0$, so it actually tells us, in a sense, ... the initial image is still have much from the initial image is still present at step $t$.

Also, another quantity of interest "the intensity" is $\lambda = \log$ SNR. Given that $t$ is a random variable (uniform as in DDPM), $\lambda$ above is a change of coordinates in the PDF's.

$$p(\lambda) = p(t)\left|\frac{dt}{d\lambda}\right|$$

Assuming $\lambda$ is decreasing with $t$ (which is natural to assume under the limits above):

$$p(\lambda) = -\frac{dt}{d\lambda}.$$

As an example, the cosine schedule is $\alpha_t = \cos\left(\frac{\pi t}{2}\right)$, $\sigma_t = \sin\left(\frac{\pi t}{2}\right)$ ($\forall t \in [0,1]$, otherwise $\frac{\pi t}{2T}$, for $T = 1000$ or other). So, $\lambda = -2 \ln \tan \frac{\pi}{2}$, $t = \frac{2}{\pi} \arctan e^{-\lambda/2}$. Thus, $p(\lambda) = -\frac{dt}{d\lambda}$

$= \frac{1}{2\pi} \operatorname{sech} \frac{\lambda}{2}$.

Also, $t = 1 - \int_{-\infty}^{\infty} p(\lambda) d\lambda = P(\lambda)$

$\Rightarrow \lambda = P^{-1}(t)$.

We can also obtain the noise schedule $\lambda$ by the inverse $P^{-1}$.

Under a general SNR we still have to look at how the ELBO changes.

Let $\mathcal{L}(x_0) = \sum_{i=1}^{T} E_{q(x_t|x_0)} D_{KL}[q(x_{i-1}|x_i, x_0)$

$p(x_{i-1}|x_i, x_0)]$.

There is a similarity between this and an integral if the steps are small and $T \to \infty$, transforming the sum into an integral.

For it, we will adopt a different notation by discretizing the $[0,1]$ interval into steps

$$s(i) = \frac{i-1}{T} := s \quad \text{and} \quad t(i) = \frac{i}{T} := t$$

$$p(x_s | x_t) = q(x_s | x_t, \hat{x}_\theta(x_t; t))$$

$$= \mathcal{N}(x_s; \mu_q(x_s, x_0; s, t), \sigma_q^2(s,t) I)$$

and $p(x_s | x_t) = \mathcal{N}(x_s; \mu_\theta(x_t; s, t), \sigma_q^2(s,t) I)$.

Recall this is the setup that we had in DDPM.

$$\sigma_{t|s}^2 = \sigma_t^2 - \frac{\alpha_t^2}{\alpha_s^2}\sigma_s^2$$

$$\mu_q = \frac{\frac{\alpha_t}{\alpha_s}\sigma_s^2}{\sigma_t^2} x_t + \frac{\alpha_s}{\sigma_t^2}\left(\sigma_t^2 - \frac{\alpha_t^2}{\alpha_s^2}\sigma_s^2\right) x_0$$

$$\mu_\theta = \frac{\frac{\alpha_t}{\alpha_s}\sigma_s^2}{\sigma_t^2} x_t + \frac{\alpha_s}{\sigma_t^2}\left(\sigma_t^2 - \frac{\alpha_t^2}{\alpha_s^2}\sigma_s^2\right) \hat{x}_\theta$$

$$\sigma_q^2 = \left(\sigma_t^2 - \frac{\alpha_t^2}{\alpha_s^2}\sigma_s^2\right)\frac{\sigma_s^2}{\sigma_t^2}$$

After computing the KL-dive, we end up with:

$$D_{KL} = \frac{1}{2}\left(SNR(s) - SNR(t)\right)\|x - \hat{x}_\theta(x_t;t)\|_2^2$$

With $x_t = \alpha_t x_0 + \sigma_t \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$

$$\mathcal{L}(x_0) = \frac{T}{2}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I),\, i\sim\{1,\dots,T\}}\left[(SNR(s) - SNR(t))\right.$$
$$\left.\|x - \hat{x}_\theta(x_t;t)\|_2^2\right]$$

When $T \to \infty$ we see have the sum converges to an integral

$$\mathcal{L}_\infty(x_0) = -\frac{1}{2}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\int_0^1 SNR'(t)\|x - \hat{x}_\theta\|_2^2 dt$$

The new variational lower bound indicates an equivalence under scaling which might we use as a new loss, where during training a range of noisy samples are prioritized.

$$SNR'(t) = (e^{\lambda})' = \frac{d\lambda}{dt} SNR \text{ or } \frac{1}{\frac{dt}{d\lambda}} SNR$$

$$= \frac{w(\lambda)}{\quad}$$ , if we compactify the above expression

$p(\lambda)$ s.t. everything except $p(\lambda)$ is some general weighting $w(\lambda)$!

Our new weighted loss becomes

$$\mathcal{L}_w(x_0) = -\frac{1}{2}\mathbb{E}_{t\sim \mathcal{U}(0,1),\, \varepsilon\sim\mathcal{N}(0,I)}\left[\frac{w(\lambda_t)}{p(\lambda_t)}\|\hat{\varepsilon}_\theta - \varepsilon\|_2^2\right]$$

The minus comes from the fact we want to minimize, and now time is sampled uniformly in $(0,1)$, although for practical reasons can be converted to $[0,T]$ in implementation. Additionally, if we replace $\frac{d\lambda}{dt}$ in $\mathcal{L}_\infty(x_0)$ we see how the integral no longer depends on $p(\lambda)$ and only $w(\lambda)$. So, the integral is independent of the noise schedule (remember, $w(\lambda)$ we want to make it arbitrary), although from our deduction looks like ... Also the bounds change

$w(\lambda) \sim SNR$). Also, the

$\int_{\lambda_{min}}^{\lambda_{max}}$ so as long as two noise schedulers $p(\lambda)$ define same $\lambda_{max}$, $\lambda_{min}$, the integral is the same (and $w(\lambda)$ is unchanged).

The last lines can be ignored, since we aren't computing the time integral, but only sampling as a Monte Carlo approximation. Due to sampling, the gradients and variance of the expectation depends on $p(\lambda)$ or noise scheduler. Hence, there is a reason to search for different noise schedulers to probably optimise sampling.

The final loss function becomes:

$$\mathcal{L}_w(\theta) = \frac{1}{2} \mathbb{E}_{x \sim \mathcal{D}, \epsilon \sim \mathcal{N}(0, I), \lambda \sim p(\lambda)} \left[ \frac{w(\lambda)}{p(\lambda)} \| \epsilon_\theta(x_\lambda; \lambda) - \epsilon \|_2^2 \right].$$

The noise schedule that was tested to achieve the best sampling quality is Laplace:

$$p(\lambda) = \frac{e^{-\frac{|\lambda-\mu|}{b}}}{2b}$$

$$\lambda(t) = \mu - b\,\mathrm{sgn}(0{,}5 - t)\ln(1 - 2|t - 0{,}5|)$$