A DERIVATIONS

A.1 DERIVATION OF THE FUNDAMENTAL POINTWISE DENOISING RELATION

We now derive the following pointwise denoising relation.

$$\frac{d}{d\gamma} D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)] = {}^1\!/_2 \, \mathrm{mmse}(\boldsymbol{x}, \gamma)$$

Recall the definition of pointwise MMSE, except we will use a shorthand notation for the analytic MMSE denoiser.

$$\begin{aligned} \mathrm{mmse}(\boldsymbol{x}, \gamma) &\equiv \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})}[||\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma]||_2^2] \\ \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma] &\equiv \int d\bar{\boldsymbol{x}} \; p(\bar{\boldsymbol{x}}|\boldsymbol{z}_\gamma) \; \bar{\boldsymbol{x}} = \int d\bar{\boldsymbol{x}} \; p(\boldsymbol{z}_\gamma|\bar{\boldsymbol{x}}) p(\bar{\boldsymbol{x}}) / p(\boldsymbol{z}_\gamma) \; \bar{\boldsymbol{x}} \end{aligned}$$

We begin by expanding the left-hand side.

$$\begin{aligned} \frac{d}{d\gamma} D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)] &= \frac{d}{d\gamma} \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})}[\log p(\boldsymbol{z}_\gamma|\boldsymbol{x})] - \frac{d}{d\gamma} \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})}[\log p(\boldsymbol{z}_\gamma)] \\ &= -\frac{d}{d\gamma} \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})}[\log p(\boldsymbol{z}_\gamma)] = -\int \; d\boldsymbol{z}_\gamma \frac{d}{d\gamma}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \log p(\boldsymbol{z}_\gamma)] \\ &= -\int \; d\boldsymbol{z}_\gamma (\frac{d}{d\gamma} p(\boldsymbol{z}_\gamma|\boldsymbol{x})[\log p(\boldsymbol{z}_\gamma)] + p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \frac{d}{d\gamma} \log p(\boldsymbol{z}_\gamma)) \end{aligned}$$

The first term in the first line is a constant that does not depend on γ. Then we expand with the product rule. The following is an easily verified identity for our Gaussian noise channel, $p(\boldsymbol{z}_\gamma|\boldsymbol{x})$.

$$\frac{d}{d\gamma} p(\boldsymbol{z}_\gamma|\boldsymbol{x}) = -\boldsymbol{x}/(2\sqrt{\gamma}) \cdot \nabla_{\boldsymbol{z}_\gamma} p(\boldsymbol{z}_\gamma|\boldsymbol{x})$$

We also need the following relation.

$$\begin{aligned} \frac{d}{d\gamma} \log p(\boldsymbol{z}_\gamma) &= 1/p(\boldsymbol{z}_\gamma) \frac{d}{d\gamma} p(\boldsymbol{z}_\gamma) = 1/p(\boldsymbol{z}_\gamma) \int d\bar{\boldsymbol{x}} \frac{d}{d\gamma} p(\boldsymbol{z}_\gamma|\bar{\boldsymbol{x}}) p(\bar{\boldsymbol{x}}) \\ &= -1/(2\sqrt{\gamma}) \; 1/p(\boldsymbol{z}_\gamma) \int d\bar{\boldsymbol{x}} \; \bar{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{z}_\gamma} p(\boldsymbol{z}_\gamma|\bar{\boldsymbol{x}}) p(\bar{\boldsymbol{x}}) \end{aligned}$$

Using these expressions in the derivation above, we get the following.

$$\begin{aligned} \frac{d}{d\gamma} D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)] = \int \; d\boldsymbol{z}_\gamma \Big(& \boldsymbol{x}/(2\sqrt{\gamma}) \cdot \nabla_{\boldsymbol{z}_\gamma} p(\boldsymbol{z}_\gamma|\boldsymbol{x})[\log p(\boldsymbol{z}_\gamma)] \\ & + 1/(2\sqrt{\gamma}) \; p(\boldsymbol{z}_\gamma|\boldsymbol{x})/p(\boldsymbol{z}_\gamma) \int d\bar{\boldsymbol{x}} \; \bar{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{z}_\gamma} p(\boldsymbol{z}_\gamma|\bar{\boldsymbol{x}}) p(\bar{\boldsymbol{x}}) \Big) \end{aligned}$$

Next we use integration by parts on $\boldsymbol{z}_\gamma$.

$$\begin{aligned} \frac{d}{d\gamma} D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)] &= -\int \; d\boldsymbol{z}_\gamma \Big(1/2 \; p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \; \boldsymbol{x}/\sqrt{\gamma} \cdot \nabla_{\boldsymbol{z}_\gamma} \log p(\boldsymbol{z}_\gamma) \\ &\qquad + 1/(2\sqrt{\gamma}) \; 1/p(\boldsymbol{x}) \int d\bar{\boldsymbol{x}} \; p(\boldsymbol{z}_\gamma|\bar{\boldsymbol{x}}) p(\bar{\boldsymbol{x}}) \; \bar{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{z}_\gamma} p(\boldsymbol{x}|\boldsymbol{z}_\gamma) \Big) \\ &= -\int \; d\boldsymbol{z}_\gamma \Big(1/2 \; p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \; \boldsymbol{x}/\sqrt{\gamma} \cdot \nabla_{\boldsymbol{z}_\gamma} \log p(\boldsymbol{z}_\gamma) \\ &\qquad + 1/(2\sqrt{\gamma}) \; p(\boldsymbol{z}_\gamma)/p(\boldsymbol{x}) \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma] \cdot \nabla_{\boldsymbol{z}_\gamma} p(\boldsymbol{x}|\boldsymbol{z}_\gamma) \Big) \end{aligned}$$

The gradients can be written,

$$\begin{aligned} \nabla_{\boldsymbol{z}_\gamma} p(\boldsymbol{x}|\boldsymbol{z}_\gamma) &= p(\boldsymbol{x}|\boldsymbol{z}_\gamma) \sqrt{\gamma} (\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma]) \\ \nabla_{\boldsymbol{z}_\gamma} \log p(\boldsymbol{z}_\gamma) &= \sqrt{\gamma} \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma] - \boldsymbol{z}_\gamma. \end{aligned}$$

$$\begin{aligned}
\frac{d}{d\gamma} D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)] = \int d\boldsymbol{z}_\gamma \Big(& 1/2\, p(\boldsymbol{z}_\gamma|\boldsymbol{x})\, \boldsymbol{x}/\sqrt{\gamma} \cdot (\boldsymbol{z}_\gamma - \sqrt{\gamma}\mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma]) \\
& - 1/(2\sqrt{\gamma})\, p(\boldsymbol{z}_\gamma|\boldsymbol{x})\mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma] \cdot \sqrt{\gamma}(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma]) \Big) \\
= 1/2 \int d\boldsymbol{z}_\gamma p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \Big(& - \boldsymbol{x} \cdot \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma] + \boldsymbol{x} \cdot \boldsymbol{z}_\gamma/\sqrt{\gamma} \\
& - \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma] \cdot (\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma]) \Big) \\
= 1/2 \int d\boldsymbol{z}_\gamma p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \Big(& - \boldsymbol{x} \cdot \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma] + \boldsymbol{x} \cdot \boldsymbol{x} \\
& - \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma] \cdot (\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma]) \Big) \\
= 1/2 \int d\boldsymbol{z}_\gamma\, p(\boldsymbol{z}_\gamma|\boldsymbol{x})\, & \|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma]\|_2^2 \\
= {}^1\!/\!_2\, \mathrm{mmse}(\boldsymbol{x}, \gamma) \quad \square &
\end{aligned}$$

A.2 High SNR KL divergence limit

In this appendix, we derive the following result.

$$\lim_{\gamma \to \infty} D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p_G(\boldsymbol{z}_\gamma)] - D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)] = \log \frac{p(\boldsymbol{x})}{p_G(\boldsymbol{x})}$$

In this expression, we have two "base distributions", $p(\boldsymbol{x}), p_G(\boldsymbol{x})$, and we consider the marginal distributions after injecting Gaussian noise, $p_G(\boldsymbol{z}_\gamma) = \int p(\boldsymbol{z}_\gamma|\boldsymbol{x}) p_G(\boldsymbol{x}) d\boldsymbol{x}$, $p(\boldsymbol{z}_\gamma) = \int p(\boldsymbol{z}_\gamma|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$. Start by expanding and canceling out terms.

$$\begin{aligned}
f(\boldsymbol{x}, \gamma) &\equiv D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p_G(\boldsymbol{z}_\gamma)] - D_{KL}[p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \parallel p(\boldsymbol{z}_\gamma)] && \text{(Cancel } \log p(\boldsymbol{z}_\gamma|\boldsymbol{x}) \text{ terms)} \\
&= \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})}[\log\left((p(\boldsymbol{z}_\gamma)\gamma^{d/2})/(p_G(\boldsymbol{z}_\gamma)\gamma^{d/2})\right)] && \text{(Multiply by 1)} \\
&= \mathbb{E}_{p(\boldsymbol{\epsilon})}[\log\left((p(\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon})\gamma^{d/2})/(p_G(\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon})\gamma^{d/2})\right)] && \text{(Reparametrize)}
\end{aligned}$$

Set $Z = (2\pi)^{d/2}$ for $\boldsymbol{x} \in \mathbb{R}^d$ and re-arrange.

$$\begin{aligned}
p(\boldsymbol{z}_\gamma)\gamma^{d/2} &= \int d\bar{\boldsymbol{x}}\; p(\bar{\boldsymbol{x}}) 1/Z\; e^{-{}^1\!/\!_2 (\boldsymbol{z}_\gamma - \sqrt{\gamma}\bar{\boldsymbol{x}})^2} \gamma^{d/2} \\
&= \int d\bar{\boldsymbol{x}}\; p(\bar{\boldsymbol{x}}) 1/Z\; e^{-{}^1\!/\!_2 (\boldsymbol{z}_\gamma/\sqrt{\gamma} - \bar{\boldsymbol{x}})^2 \gamma} \gamma^{d/2}
\end{aligned}$$

For large γ, we recognize a limit representation of the Dirac delta function in blue.

$$\begin{aligned}
p(\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon})\gamma^{d/2} &= \int d\bar{\boldsymbol{x}}\; p(\bar{\boldsymbol{x}}) 1/Z\; e^{-{}^1\!/\!_2 (\boldsymbol{x} + \boldsymbol{\epsilon}/\sqrt{\gamma} - \bar{\boldsymbol{x}})^2 \gamma} \gamma^{d/2} \\
\lim_{\gamma \to \infty} p(\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon})\gamma^{d/2} &= \int d\bar{\boldsymbol{x}}\; p(\bar{\boldsymbol{x}}) \delta(\boldsymbol{x} - \bar{\boldsymbol{x}}) \\
&= p(\boldsymbol{x})
\end{aligned}$$

Using this result leads to the desired result.

$$\begin{aligned}
\lim_{\gamma \to \infty} f(\boldsymbol{x}, \gamma) &= \lim_{\gamma \to \infty} \mathbb{E}_{p(\boldsymbol{\epsilon})}[\log p(\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon})\gamma^{d/2} - \log p(\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon})\gamma^{d/2}] \\
&= \log p(\boldsymbol{x})/p_G(\boldsymbol{x}) \quad \square
\end{aligned}$$

This informal proof using delta functions could be made more rigorous in a measure-theoretic setting. Once again, we have derived a point-wise generalization of Eq. 177 from Guo et al. (2005), which can be recovered by taking the expectation of our result.

A.3 Simplifying Density with an Integral Identity

Our goal is to simplify the expression of the density coming from Eq. (8).

$$\begin{aligned} -\log p(\boldsymbol{x}) &= -\log p_G(\boldsymbol{x}) - {}^1\!/\!_2 \int_0^\infty d\gamma \, (\mathrm{mmse}_G(\boldsymbol{x}, \gamma) - \mathrm{mmse}(\boldsymbol{x}, \gamma)) \\ &= d/2 \log(2\pi) + x^2/2 - {}^1\!/\!_2 \int_0^\infty d\gamma \left(\frac{x^2 + \gamma\, d}{(1+\gamma)^2} - \mathrm{mmse}(\boldsymbol{x}, \gamma) \right) \\ &= d/2 \log(2\pi e) - {}^1\!/\!_2 \int_0^\infty d\gamma \left(\frac{d}{1+\gamma} - \mathrm{mmse}(\boldsymbol{x}, \gamma) \right) \end{aligned}$$

In the second line we write out the expressions. Note the the pointwise MMSE for a standard normal input is derived in more detail in App. A.5. In the third line we make use of the following integral identity.

$$\int_0^\infty \left(\frac{x^2 + \gamma\, d}{(1+\gamma)^2} - \frac{d}{1+\gamma} \right) d\gamma = x^2 - d$$

This identity can be verified via elementary manipulations (multiply second term in integrand by $(1+\gamma)/(1+\gamma)$), but we state it explicitly because of its counter-intuitive form.

A.4 Non-Gaussian Density Representation

Instead of using $p_G(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; 0, \mathbb{I})$, we can take the base measure to be a Gaussian, $p_G(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean and covariance that match the data. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of the covariance matrix. We could start with the derivation in App. A.3 and try to simplify in a similar way with a more complex integral identity. However, a more straightforward derivation proceeds as follows.

Start with the simple density estimator from Eq. (9) assuming a standard normal as the base measure, but replacing the factor of d with a sum over d terms.

$$-\log p(\boldsymbol{x}) = d/2 \log(2\pi e) - {}^1\!/\!_2 \int_0^\infty d\gamma \left(\sum_{i=1}^d \frac{1}{1+\gamma} - \mathrm{mmse}(\boldsymbol{x}, \gamma) \right)$$

Now, note the following integral identity:

$$\int_0^\infty \left(\frac{1}{\gamma + 1/\lambda_i} - \frac{1}{\gamma + 1} \right) d\gamma = \log \lambda_i.$$

Using this to replace $1/(\gamma + 1)$ terms in the previous expression gives the following.

$$-\log p(\boldsymbol{x}) = d/2 \log(2\pi e) + {}^1\!/\!_2 \sum_{i=1}^d \log \lambda_i - {}^1\!/\!_2 \int_0^\infty d\gamma \left(\sum_{i=1}^d \frac{1}{\gamma + 1/\lambda_i} - \mathrm{mmse}(\boldsymbol{x}, \gamma) \right)$$

We recognize the sum of the log eigenvalues as equivalent to the log determinant. The constants can also be pulled inside the determinant.

$$-\log p(\boldsymbol{x}) = \underbrace{1/2 \log \det(2\pi e \boldsymbol{\Sigma})}_{\text{Gaussian entropy}} \underbrace{- {}^1\!/\!_2 \int_0^\infty d\gamma \left(\sum_{i=1}^d \frac{1}{\gamma + 1/\lambda_i} - \mathrm{mmse}(\boldsymbol{x}, \gamma) \right)}_{\text{Deviation from Gaussianity} \geq 0}$$

Note that the deviation from Gaussianity need only be non-negative in expectation.

When we change variables of integration to $\alpha = \log \gamma$ we get the following.

$$\int_0^\infty d\gamma \sum_{i=1}^d \frac{1}{\gamma + 1/\lambda_i} = \int_{-\infty}^\infty d\alpha \sum_{i=1}^d \sigma(\alpha + \log \lambda_i) \qquad (17)$$

Here the traditional sigmoid function is used $\sigma(t) = 1/(1 + e^{-t})$.

A.5 Gaussian properties

Consider $p(\boldsymbol{x}) = \mathcal{N}(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, for some covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ be the SVD, with eigenvalues $\Lambda_{ii} = \lambda_i$. For this distribution we would like to derive the ground truth decoder and MMSE, both for testing purposes and to use as a fallback estimator in regions where our discovered decoder is clearly sub-optimal.

We have already mentioned in Eq. (3) that the ideal decoder is:

$$\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma) \equiv \arg \operatorname{mmse}(\gamma) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}[\boldsymbol{x}].$$

For $p(\boldsymbol{x})$ a Gaussian distribution, we can simply look up this conditional mean in a textbook to find that the MMSE estimator is:

$$\begin{aligned}\hat{\boldsymbol{x}}^*_G(\boldsymbol{z}_\gamma, \gamma) &= \boldsymbol{\mu} + \sqrt{\gamma}(\gamma\mathbb{I} + \boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{z}_\gamma - \sqrt{\gamma}\boldsymbol{\mu}) \\ &= \boldsymbol{\mu} + \sqrt{\gamma}\boldsymbol{U}(\gamma + \boldsymbol{\Lambda}^{-1})^{-1}\boldsymbol{U}^T(\boldsymbol{z}_\gamma - \sqrt{\gamma}\boldsymbol{\mu})\end{aligned}$$

Taking the expectation, $\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}_\gamma}[(\boldsymbol{x} - \hat{\boldsymbol{x}}^*_G(\boldsymbol{z}_\gamma, \gamma))^2]$, gives the MMSE with some manipulation:

$$\operatorname{mmse}_G(\gamma) = Tr\left((\boldsymbol{\Sigma}^{-1} + \gamma\mathbb{I})^{-1}\right) = \sum_i \frac{1}{1/\lambda_i + \gamma}.$$

Note that we used the cyclic property of the trace to write the final expression in terms of the eigenvalues only. The case for the standard normal distribution (Guo et al., 2005) follows from setting the eigenvalues to 1.

The negative log likelihood is:

$$-\log p(\boldsymbol{x}) = 1/2\ (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) + 1/2 \log \det(2\pi\boldsymbol{\Sigma})$$

It can be verified using the integral identities in App. A.4 and App. A.3 that using the Gaussian MMSE in Eq. (9) recovers this equation.

A.6 MMSE Upper Bounds via Bregman Divergence Interpretation

In this section, we derive the upper bound in the objective using a suboptimal denoising function $\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)$. The Bregman divergence generated from the strictly convex function $\phi(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{x}, \boldsymbol{x}\rangle = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ is

$$d_\phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\langle \boldsymbol{x}, \boldsymbol{x}\rangle - \frac{1}{2}\langle \boldsymbol{y}, \boldsymbol{y}\rangle - \langle \boldsymbol{x} - \boldsymbol{y}, \nabla\phi(\boldsymbol{y})\rangle = \frac{1}{2}\langle \boldsymbol{x}, \boldsymbol{x}\rangle - \frac{1}{2}\langle \boldsymbol{y}, \boldsymbol{y}\rangle - \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{y}\rangle = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

Note that the Legendre dual of ϕ, or $\psi(\boldsymbol{y}) = \sup_{\boldsymbol{x}}\langle \boldsymbol{x}, \boldsymbol{y}\rangle - \frac{1}{2}\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}\|_2^2$, matches $\phi(\boldsymbol{x})$. Although, in general, we only know that $d_\phi(\boldsymbol{x}, \boldsymbol{y}) = d_\psi(\boldsymbol{y}, \boldsymbol{x})$, in this case the Bregman divergence is symmetric with $d_\phi(\boldsymbol{x}, \boldsymbol{y}) = d_\psi(\boldsymbol{x}, \boldsymbol{y})$.

For the Gaussian noise channel $\boldsymbol{z}_\gamma = \sqrt{\gamma}\boldsymbol{x} + \boldsymbol{\epsilon}$ with source $\boldsymbol{x} \sim p(\boldsymbol{x})$, we treat the Bregman divergence as a distortion loss for reconstructing $\boldsymbol{x}$ from observed samples $\boldsymbol{z}_\gamma$ using the denoising function $\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)$. In contrast to the developments in the main text, here we consider the mmse as a function of each noisy sample $\boldsymbol{z}_\gamma \sim p(\boldsymbol{z}_\gamma)$ instead of input sample $\boldsymbol{x} \sim p(\boldsymbol{x})$.

Performing minimization in the second argument, Banerjee et al. (2005) show that the arithmetic mean over inputs minimizes the expected Bregman divergence (regardless of the convex generator ϕ)

$$\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma) = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}[\boldsymbol{x}] = \arg \min_{\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)} \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}\left[d_\phi(\boldsymbol{x}, \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma))\right] \tag{18}$$

At this minimizing argument, the expected divergence is the MMSE and corresponds to the conditional variance (see (Banerjee et al., 2005) Ex. 5)

$$\frac{1}{2}\operatorname{mmse}(\boldsymbol{z}_\gamma, \gamma) = \min_{\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)} \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}\left[d_\phi(\boldsymbol{x}, \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma))\right] = \frac{1}{2}\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}\left[\|\boldsymbol{x} - \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}[\boldsymbol{x}]\|_2^2\right] = \frac{1}{2}\mathrm{Var}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}[\boldsymbol{x}]$$

To prove that the mean provides the minimizing argument in Eq. (18), (Banerjee et al., 2005) Prop. 1 consider the gap in the expected divergence for a suboptimal representative $\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)$. This can be shown to yield another Bregman divergence, which in our case becomes

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}\left[d_\phi(\boldsymbol{x}, \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma))\right] - \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}\left[d_\phi(\boldsymbol{x}, \hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma))\right] &= d_\phi(\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma), \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)) \\ &= \frac{1}{2}\|\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma) - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)\|_2^2 \geq 0.\end{aligned}$$

This allows us to derive the gap in the MMSE upper bounds in Sec. 4, which arise from using suboptimal neural network denoisers $\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)$ instead of the true conditional expectation.

$$\mathbb{E}_{p(\boldsymbol{z}_\gamma,\boldsymbol{x})}\left[\|\boldsymbol{x}-\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma)\|_2^2\right] = \underbrace{\mathbb{E}_{p(\boldsymbol{z}_\gamma,\boldsymbol{x})}\left[\|\boldsymbol{x}-\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma)\|_2^2\right]}_{\mathrm{mmse}(\gamma)} + \underbrace{\mathbb{E}_{p(\boldsymbol{z}_\gamma,\boldsymbol{x})}\left[\|\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma)-\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma)\|_2^2\right]}_{\text{estimation gap}}. \quad (19)$$

Finally, Banerjee et al. (2005) Sec. 4 proves a bijection between regular Bregman divergences and exponential familes, so that minimizing a Bregman divergence loss corresponds to maximum likelihood estimation within a corresponding exponential family. See their Ex. 9 demonstrating the case of the mean-only, spherical Gaussian family. Notably, in this case, the natural parameters θ and expectation parameters η are equivalent. For decoding in our Gaussian noise channel using the squared error as a (Bregman) loss, we obtain a probabilistic interpretation of the MMSE optimization in Eq. (18) (Banerjee et al. (2005))

$$\min_{\eta=\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma)} \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}\left[d_\phi(\boldsymbol{x},\eta)\right] = \max_{\theta=\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma)} \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}[\log\mathcal{N}(\boldsymbol{x};\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma),\sigma^2\mathbb{I})] \quad (20)$$

where the Normal family is the appropriate exponential family $p(\boldsymbol{x};\theta)$ and the optimum is $\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma) = \eta^* = \theta^* = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}[\boldsymbol{x}]$. Finally, we can view the equality in Eq. (19) as an expression of the Pythagorean relation or m-projection (Amari (2016) Sec. 1.6, 2.8, Nielsen (2018)) in information geometry. In particular, $\theta^* = \hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma)$ is the projection of $p(\boldsymbol{x}|\boldsymbol{z}_\gamma)$ onto the submanifold of fixed-variance, diagonal Gaussian distributions, and for a suboptimal $\theta = \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma)$ we have

$$\begin{aligned} D_{KL}[p(\boldsymbol{x}|\boldsymbol{z}_\gamma)\|\mathcal{N}(\boldsymbol{x};\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma),\sigma^2\mathbb{I})] =& D_{KL}[p(\boldsymbol{x}|\boldsymbol{z}_\gamma)\|\mathcal{N}(\boldsymbol{x};\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma),\sigma^2\mathbb{I})] \\ &+ D_{KL}[\mathcal{N}(\boldsymbol{x};\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma),\sigma^2\mathbb{I})\|\mathcal{N}(\boldsymbol{x};\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma),\sigma^2\mathbb{I})]. \end{aligned}$$

In future work, it would be interesting to consider using the I-MMSE relations for general Bregman divergences as in Wang et al. (2013; 2014).

A.7 Comparison with Variational Bounds

In this section, we aim to relate the mmse terms in our bound to terms in the standard variational lower bound for diffusion models.

Forward and Reverse Processes We first review the standard forward and reverse processes used to define denoising diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Kingma et al., 2021). Using the notation of Kingma et al. (2021) in terms of SNR values $\gamma_t = \alpha_t^2/\sigma_t^2$, we set $\sigma_t^2 = 1$ for simplicity and without loss of generality, since the eventual objective will depend only on γ_t.

$$\begin{aligned} q(\boldsymbol{z}_{\gamma_{0:T}}|\boldsymbol{x}) &= q(\boldsymbol{z}_{\gamma_0}|\boldsymbol{x})\prod_{t=1}^{T} q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{z}_{\gamma_{t-1}}) \\ \text{where} \quad q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{z}_{\gamma_{t-1}}) &:= \mathcal{N}\left(\boldsymbol{z}_{\gamma_t}; \sqrt{\frac{\gamma_t}{\gamma_{t-1}}}\boldsymbol{z}_{\gamma_{t-1}}, \frac{\gamma_{t-1}-\gamma_t}{\gamma_{t-1}}\mathbb{I}\right) \end{aligned} \quad (21)$$

The Markovian time-reversal $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ of the forward process is only Gaussian in the limit of $T \to \infty$ (Anderson, 1982; Feller, 2015), in which case we interpret both processes as stochastic differential equations (Song et al., 2020). However, the conditional $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x})$ is always Gaussian. Using Bayes rule,

$$q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x}) = \frac{q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{z}_{\gamma_{t-1}})q(\boldsymbol{z}_{\gamma_{t-1}}|x)}{q(\boldsymbol{z}_{\gamma_t}|x)} = \mathcal{N}\left(\boldsymbol{z}_{\gamma_{t-1}}; \sqrt{\frac{\gamma_t}{\gamma_{t-1}}}\boldsymbol{z}_{\gamma_t} + \frac{\gamma_{t-1}-\gamma_t}{\sqrt{\gamma_{t-1}}}\boldsymbol{x}\,,\ \frac{\gamma_{t-1}-\gamma_t}{\gamma_{t-1}}\mathbb{I}\right) \quad (22)$$

since each the forward process is Markovian $q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{z}_{\gamma_{t-1}},\boldsymbol{x}) = q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{z}_{\gamma_{t-1}})$ and each $q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{x})$ is Gaussian by construction of the noise channel

$$q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{x}) = \sqrt{\gamma_t}\,\boldsymbol{x} + \boldsymbol{\epsilon}. \quad (23)$$

See Kingma et al. (2021) App. A2 for example derivations of the Gaussian parameters.

Finally, consider defining a generative model using a variational reverse process. Sohl-Dickstein et al. (2015) (Sec. 2.2) choose to parameterize each conditional $p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ as a Gaussian, which is inspired by the fact that $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ is Gaussian in the limit as $T \to \infty$,

$$\begin{aligned} p(\boldsymbol{z}_{\gamma_{0:T}}, \boldsymbol{x}) &= p(\boldsymbol{z}_{\gamma_T}) \prod_{t=1}^{T} p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) \cdot p(\boldsymbol{x}|\boldsymbol{z}_{\gamma_0}) \qquad (24) \\ \text{where} \quad p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) &:= q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}, \hat{\boldsymbol{x}}(\boldsymbol{z}_{\gamma_t}, \gamma_t)) \\ &= \mathcal{N}\left(\boldsymbol{z}_{\gamma_{t-1}}; \sqrt{\frac{\gamma_t}{\gamma_{t-1}}}\boldsymbol{z}_{\gamma_t} + \frac{\gamma_{t-1}-\gamma_t}{\sqrt{\gamma_{t-1}}}\hat{\boldsymbol{x}}(\boldsymbol{z}_{\gamma_t}, \gamma_t)\,,\ \frac{\gamma_{t-1}-\gamma_t}{\gamma_{t-1}}\mathbb{I}\right). \end{aligned}$$

Note that we have written $p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ in terms of a denoising function $\hat{\boldsymbol{x}}(\boldsymbol{z}_{\gamma_t}, \gamma_t)$. While other forms are possible (e.g. Kingma et al. (2021) Eq. 28), this expression will be useful to draw connections with the optimal denoising function $\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma, \gamma)$ found in the mmse expression.

Discrete Variational Lower Bound We can now express the discrete variational lower bound using extended state space importance sampling (Neal, 2001; Sohl-Dickstein et al., 2015)

$$\begin{aligned} \log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{z}_{\gamma_{0:T}}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z}_{\gamma_{0:T}}, \boldsymbol{x})}{q(\boldsymbol{z}_{\gamma_{0:T}}|\boldsymbol{x})}\right] &= \mathbb{E}_{q(\boldsymbol{z}_{\gamma_{0:T}}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z}_{\gamma_T}) \prod_{t=1}^{T} p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) \cdot p(\boldsymbol{x}|\boldsymbol{z}_{\gamma_0})}{q(\boldsymbol{z}_{\gamma_0}|\boldsymbol{x}) \prod_{t=1}^{T} q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{z}_{\gamma_{t-1}})}\right] \\ &= \mathbb{E}_{q(\boldsymbol{z}_{\gamma_{0:T}}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z}_{\gamma_T}) \prod_{t=1}^{T} p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) \cdot p(\boldsymbol{x}|\boldsymbol{z}_{\gamma_0})}{q(\boldsymbol{z}_{\gamma_T}|\boldsymbol{x}) \prod_{t=1}^{T} q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}, \boldsymbol{x})}\right] \end{aligned}$$

$$= \mathbb{E}_{q(\boldsymbol{z}_{\gamma_0}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z}_{\gamma_0})\right] - D_{KL}[q(\boldsymbol{z}_{\gamma_T}|\boldsymbol{x}) \| p(\boldsymbol{z}_{\gamma_T})] + \sum_{t=1}^{T} D_{KL}[q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}, \boldsymbol{x}) \| p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})]$$

Compared to the notation in the main text, note that the forward process $q(\boldsymbol{z}_{\gamma_0}|\boldsymbol{x})$ and $q(\boldsymbol{z}_{\gamma_T}|\boldsymbol{x})$ represent the Gaussian noise channels instead of $p(\boldsymbol{z}_\gamma|\boldsymbol{x})$. Instead of the true denoising posterior $p(\boldsymbol{x}|\boldsymbol{z}_{\gamma_0})$ is a variational distribution in the above expression. Our goal is now to relate the KL divergence terms, particularly the true time-reversal process $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ or its Gaussian approximation $p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$, to the MMSE.

A.7.1 Relation to Incremental Noise Channel Proof of I-MMSE Relation

The proof of the I-MMSE relation in Guo et al. (2005) relies on the construction of an *incremental channel* which successively adds Gaussian noise to the data. We recognize this channel as being identical to the conditional $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}, \boldsymbol{x})$ in the limit as $\delta = \gamma_{t-1} - \gamma_t \to 0$, and restate the proof using the notation above to highlight connections with terms in the variational lower bound.

Recall that the following distributions are Gaussian: each noise channel $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{x})$ and $q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{x})$ (Eq. (23)), the forward conditional $q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{z}_{\gamma_{t-1}})$ (Eq. (21)), and the reverse data-conditional $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}, \boldsymbol{x})$ (Eq. (22)). Using subscripts to distinguish different standard Gaussian variables, e.g. $\boldsymbol{\epsilon}_{\gamma_t} \sim \mathcal{N}(0, \mathbb{I})$, we have the following relationships

$$\begin{aligned} \boldsymbol{z}_{\gamma_{t-1}} &= \sqrt{\gamma_{t-1}}\,\boldsymbol{x} + \boldsymbol{\epsilon}_{\gamma_{t-1}} && \text{(using Eq. (23))} \qquad (25) \\ &= \sqrt{\frac{\gamma_t}{\gamma_{t-1}}}\boldsymbol{z}_{\gamma_t} + \frac{\gamma_{t-1}-\gamma_t}{\sqrt{\gamma_{t-1}}}\boldsymbol{x} + \sqrt{\frac{\gamma_{t-1}-\gamma_t}{\gamma_{t-1}}}\boldsymbol{\epsilon} && \text{(using Eq. (22))} \qquad (26) \\ \boldsymbol{z}_{\gamma_t} &= \sqrt{\gamma_t}\,\boldsymbol{x} + \boldsymbol{\epsilon}_{\gamma_t} && \text{(using Eq. (23))} \\ &= \frac{\sqrt{\gamma_t}}{\sqrt{\gamma_{t-1}}}\,\boldsymbol{z}_{\gamma_{t-1}} + \sqrt{\frac{\gamma_{t-1}-\gamma_t}{\gamma_{t-1}}}\,\boldsymbol{\epsilon}_\delta && \text{(using Eq. (21)).} \qquad (27) \end{aligned}$$

Up to change in notation, Eq. (25) and (27) match the construction of the incremental noise channel in Eq. (30)-(31) of Guo et al. (2005), while Eq. (26) matches their Eq. (37).

Proof of I-MMSE Relation using Incremental Noise Channel Our goal is to take the limit as $T \to \infty$ or $\delta = \gamma_{t-1} - \gamma_t \to 0$, in order to recover the MMSE relation $\frac{d}{d\gamma_t} I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_t}) =$

$1/2 \operatorname{mmse}(\boldsymbol{x}, \gamma_t)$. The MMSE relation for the derivative is equivalent to $I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_t}) - I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}}) = \delta/2 \operatorname{mmse}(\boldsymbol{x}, \gamma_t) + o(\delta)$ for small δ. Thus, we focus on the difference $I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}}) - I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_t})$.

Using the chain rule for mutual information and the Markov property $\boldsymbol{z}_{\gamma_t} \perp \boldsymbol{x} \,|\, \boldsymbol{z}_{\gamma_{t-1}}$ (since $\boldsymbol{x} \to \boldsymbol{z}_{\gamma_{t-1}} \to \boldsymbol{z}_{\gamma_t}$), we have

$$\begin{aligned} I(\boldsymbol{x}; \{\boldsymbol{z}_{\gamma_{t-1}}, \boldsymbol{z}_{\gamma_t}\}) = I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}}) + \cancel{I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_t} | \boldsymbol{z}_{\gamma_{t-1}})} &= I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_t}) + I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t}) \\ \implies I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}}) - I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_t}) &= I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t}) \end{aligned} \tag{28}$$

which allows us to restrict attention to $I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t})$. Using the forward or target noise process, note that this mutual information compares the ratio of the time-reversed data-conditional $q(\boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t}, \boldsymbol{x})$ and the time-reversed Markov process $q(\boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t})$,

$$\begin{aligned} I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t}) &= \mathbb{E}_{q(\boldsymbol{z}_{\gamma_{t-1}}, \boldsymbol{x} | \boldsymbol{z}_{\gamma_t})} \left[\log \frac{q(\boldsymbol{z}_{\gamma_{t-1}}, \boldsymbol{x} | \boldsymbol{z}_{\gamma_t})}{q(\boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t}) q(\boldsymbol{x} | \boldsymbol{z}_{\gamma_t})} \right] \\ &= \mathbb{E}_{q(\boldsymbol{z}_{\gamma_{t-1}}, \boldsymbol{x} | \boldsymbol{z}_{\gamma_t})} \left[\log \frac{q(\boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t}, \boldsymbol{x})}{q(\boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t})} \right]. \end{aligned} \tag{29}$$

Recalling Eq. (26),

$$\boldsymbol{z}_{\gamma_{t-1}} = \sqrt{\frac{\gamma_t}{\gamma_{t-1}}} \boldsymbol{z}_{\gamma_t} + \frac{\gamma_{t-1} - \gamma_t}{\sqrt{\gamma_{t-1}}} \boldsymbol{x} + \sqrt{\frac{\gamma_{t-1} - \gamma_t}{\gamma_{t-1}}} \boldsymbol{\epsilon}, \tag{30}$$

we can see that $I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t})$ is the mutual information over a Gaussian noise channel $q(\boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t}, \boldsymbol{x})$, where $\boldsymbol{z}_{\gamma_t}$ is given and the input is drawn from the conditional distribution $q(\boldsymbol{x} | \boldsymbol{z}_{\gamma_t})$. The marginal output distribution $q(\boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t})$, which is analogous to $p(\boldsymbol{z}_{\gamma_{t-1}})$ for the unconditional channel marginal in Eq. (4), is not Gaussian in general.

Noting that the mutual information $I(\boldsymbol{x}; \boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t}) = I(\boldsymbol{x}; c\boldsymbol{z}_{\gamma_{t-1}} | \boldsymbol{z}_{\gamma_t})$ is invariant to rescaling of $\boldsymbol{z}_{\gamma_{t-1}}$ by c, we consider

$$\begin{aligned} \sqrt{\frac{\gamma_{t-1}}{\gamma_{t-1} - \gamma_t}} \boldsymbol{z}_{\gamma_{t-1}} &= \sqrt{\frac{\gamma_t}{\gamma_{t-1}}} \sqrt{\frac{\gamma_{t-1}}{\gamma_{t-1} - \gamma_t}} \boldsymbol{z}_{\gamma_t} + \frac{\gamma_{t-1} - \gamma_t}{\sqrt{\gamma_{t-1}}} \sqrt{\frac{\gamma_{t-1}}{\gamma_{t-1} - \gamma_t}} \boldsymbol{x} + \boldsymbol{\epsilon} \\ &= \sqrt{\frac{\gamma_t}{\gamma_{t-1} - \gamma_t}} \boldsymbol{z}_{\gamma_t} + \sqrt{\gamma_{t-1} - \gamma_t}\, \boldsymbol{x} + \boldsymbol{\epsilon} \end{aligned} \tag{31}$$

in the limit as the SNR, $\delta = \gamma_{t-1} - \gamma_t$, approaches 0.

Lemma A.1 (Guo et al. (2005) Lemma 1)**.** *For a Gaussian noise channel*

$$\boldsymbol{z} = \sqrt{\delta} \boldsymbol{x} + \boldsymbol{\epsilon} \tag{32}$$

where $\boldsymbol{x} \sim p(\boldsymbol{x})$, $\mathbb{E}[\boldsymbol{x}^2] < \infty$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$, the input-output mutual information in the limit as $\delta \to 0$ is given by

$$I(\boldsymbol{z}; \boldsymbol{x}) = \frac{\delta}{2} \mathbb{E}_{p(\boldsymbol{x})} \left[\left(\boldsymbol{x} - \mathbb{E}_{p(\boldsymbol{x})}[\boldsymbol{x}] \right)^2 \right] + o(\delta) = \frac{\delta}{2} \operatorname{Var}(\boldsymbol{x}) + o(\delta) \tag{33}$$

In particular, the mutual information is independent of shape of the channel input distribution $p(\boldsymbol{x})$.

Proof. The proof proceeds by constructing a upper bound on the channel mutual information $D_{KL}[p(\boldsymbol{z}|\boldsymbol{x}) \| g(\boldsymbol{z})] = I(\boldsymbol{z}; \boldsymbol{x}) + D_{KL}[p(\boldsymbol{z}) \| g(\boldsymbol{z})]$ where $g(\boldsymbol{z}) := \mathcal{N}(\boldsymbol{z}; \mathbb{E}_{p(\boldsymbol{x})}[\sqrt{\delta} \boldsymbol{x}], (\delta \operatorname{Var}[\boldsymbol{x}] + 1)\mathbb{I})$ is a Gaussian distribution with the same mean and variance as $p(\boldsymbol{z}) = \int p(\boldsymbol{x}) p(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{x}$. As $\delta \to 0$, it can be shown that $D_{KL}[p(\boldsymbol{z}) \| g(\boldsymbol{z})] = o(\delta)$. For each $\boldsymbol{x}$, the divergence between Gaussians $\mathbb{E}_{p(\boldsymbol{x})} \left[D_{KL}[p(\boldsymbol{z}|\boldsymbol{x}) \| g(\boldsymbol{z})] \right]$ is tractable, and the terms involving the mean cancel in expectation. The $1/2 \log \frac{\operatorname{Var}_{g(\boldsymbol{z})}[\boldsymbol{z}]}{\operatorname{Var}_{p(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{z}]}$ term in the KL divergence contributes the variance term in Eq. (33), where $\operatorname{Var}_{p(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{z}] = 1$ and $\operatorname{Var}_{g(\boldsymbol{z})}[\boldsymbol{z}] = \delta \operatorname{Var}[\boldsymbol{x}] + 1$ is chosen to be the same as the variance of the output marginal $p(\boldsymbol{z})$. See Guo et al. (2005) App. II for detailed proof. □

Applying Lemma A.1 for the Gaussian channel in Eq.(30), where the input distribution is $q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})$ and the mutual information is $I(\boldsymbol{x};\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) = I(\boldsymbol{x};c\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ (see Eq.(31)), yields the desired MMSE relation. Summarizing the reasoning above, we have

$$I(\boldsymbol{x};\boldsymbol{z}_{\gamma_t}) - I(\boldsymbol{x};\boldsymbol{z}_{\gamma_{t-1}}) \overset{(28)}{=} I(\boldsymbol{x};\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) \overset{(33)}{=} \frac{\delta}{2}\operatorname{Var}_{q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})}[\boldsymbol{x}] + o(\delta) \tag{34}$$

$$\implies \quad \frac{d}{d\gamma} I(\boldsymbol{x};\boldsymbol{z}_{\gamma_t}) = \frac{1}{2}\mathbb{E}_{q(\boldsymbol{x},\boldsymbol{z}_{\gamma_t})}\left[\left(\boldsymbol{x} - \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})}[\boldsymbol{x}]\right)^2\right] := \frac{1}{2}\operatorname{mmse}(\gamma_t), \tag{35}$$

which proves the MMSE relation in Eq. (2).

Relation to Variational Lower Bound Compare our expression in Eq. (6)

$$\mathcal{L}^{\text{diff}}_{\infty}(\boldsymbol{x}) := \underbrace{\mathbb{E}_{p(\boldsymbol{z}_{\gamma_0}|\boldsymbol{x})}[-\log p(\boldsymbol{x}|\boldsymbol{z}_{\gamma_0})]}_{\text{Reconstruction loss}} \underbrace{-1/2\int_{\gamma_0}^{\gamma_1} \operatorname{mmse}(\boldsymbol{x},\gamma)d\gamma}_{\text{Diffusion loss}}.$$

to Eq. (11)-(12) in Kingma et al. (2021)

$$\mathcal{L}^{\text{diff}}_{T}(\boldsymbol{x}) := \underbrace{-\sum_{t=1}^{T} \mathbb{E}_{q(\boldsymbol{z}_{\gamma_t}|\boldsymbol{x})} D_{KL}\left[q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x}) \| p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})\right]}_{\text{Diffusion loss}}. \tag{36}$$

The conditional Gaussian parameterization of $p(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ is often justified by the fact that $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x})$ is Gaussian (Ho et al., 2020; Kingma et al., 2021).

However, from the information theoretic perspective, our goal should be to estimate the conditional mutual information $I(\boldsymbol{x};\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ in Eq. (29). Let $p^*_G(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) = \mathcal{N}(\boldsymbol{z}_{\gamma_{t-1}};\mu(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}),\sigma^2(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}))$ be the maximum likelihood Gaussian approximation to the channel output marginal $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) = \int q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x})q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})d\boldsymbol{x}$ (i.e. with the same mean and variance). Rewriting the mutual information in terms of an upper bound using this Gaussian marginal,

$$I(\boldsymbol{x};\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) = \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})}\left[D_{KL}[q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x})\|q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})]\right] \tag{37}$$

$$= \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})}\left[D_{KL}[q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x})\|p^*_G(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})]\right] - D_{KL}[q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})\|p^*_G(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})].$$

In the continuous time limit as $\delta \to 0$ or $T \to \infty$, we have that $D_{KL}[q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})\|p^*_G(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}))] = o(\delta)$ (as in the proof of Lemma A.1, where the marginal divergence $D_{KL}[p(\boldsymbol{z})\|g(\boldsymbol{z})] = o(\delta)$). Thus, we have

$$I(\boldsymbol{x};\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t}) = \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})}\left[D_{KL}[q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x})\|p^*_G(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})]\right] - \underbrace{D_{KL}[q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})\|p^*_G(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})]}_{o(\delta) \text{ as } \delta\to 0}$$

$$\overset{\delta\to 0}{\to} \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})}\left[D_{KL}[q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x})\|p^*_G(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})]\right] \tag{38}$$

To summarize, in the continuous time limit, the proof of the I-MMSE relation shows that estimating the reverse process $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ only requires a Gaussian variational family. The optimal Gaussian approximation $p^*_G(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$ requires the variance of $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t})$, which depends on the variance of the input $q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})$ to the noise channel $q(\boldsymbol{z}_{\gamma_{t-1}}|\boldsymbol{z}_{\gamma_t},\boldsymbol{x})$ (as in the proof of Lemma A.1). Evaluating this variance involves (learning) the conditional expectation or optimal denoiser $\hat{\boldsymbol{x}}^*(\boldsymbol{z}_{\gamma_t},\gamma_t) = \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{z}_{\gamma_t})}[\boldsymbol{x}]$ at each SNR, which matches Eq. (3) and leads to the optimization in Sec. 4.

B Implementation Details

B.1 Discrete probability estimator bound

We derive some results that are useful for estimating upper bounds on discrete negative log likelihood.

$$\begin{aligned}
\mathbb{E}[-\log P(\boldsymbol{x})] &= 1/2 \int_0^\infty \mathrm{mmse}(\gamma) d\gamma \\
&= 1/2 \int_{\gamma_0}^{\gamma_1} \mathrm{mmse}(\gamma) d\gamma + 1/2 \left(\int_0^{\gamma_0} + \int_{\gamma_1}^{\infty} \right) \mathrm{mmse}(\gamma) d\gamma \\
&\leq 1/2 \int_{\gamma_0}^{\gamma_1} \mathbb{E}_{\boldsymbol{z}_\gamma, \boldsymbol{x}} [\| \boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma) \|_2^2] d\gamma + c(\gamma_0, \gamma_1)
\end{aligned}$$

First, consider the left tail integral, $I_L = 1/2 \int_0^{\gamma_0} \mathrm{mmse}(\gamma) d\gamma$. When the SNR is low, the distribution is approximately Gaussian, and we can use the Gaussian MMSE to get an upper bound. The Gaussian MMSE should be for a Gaussian with either the same variance or same covariance matrix as the data, to ensure that it is an upper bound on the true MMSE. The eigenvalues of the covariance matrix are denoted with λ_i. The MMSE results for Gaussians are discussed and derived in App. A.5.

$$\begin{aligned}
I_L &= 1/2 \int_0^{\gamma_0} \mathrm{mmse}(\gamma) d\gamma \\
&\leq 1/2 \int_0^{\gamma_0} \mathrm{mmse}_G(\gamma) d\gamma \\
&= 1/2 \int_0^{\gamma_0} \sum_i 1/(\gamma + 1/\lambda_i) d\gamma \\
&= 1/2 \sum_i \log(1 + \gamma_0 \lambda_i)
\end{aligned}$$

Next we consider the right tail integral, $I_R = 1/2 \int_{\gamma_1}^{\infty} \mathrm{mmse}(\gamma) d\gamma$. In this regime, the noise is extremely small. Because the data is discrete, we can get a good estimate by simply rounding to the nearest discrete value. Let the distance between discrete values be Δ.

$$\begin{aligned}
I_R &= 1/2 \int_{\gamma_1}^{\infty} \mathrm{mmse}(\gamma) d\gamma \\
&\leq 1/2 \int_{\gamma_1}^{\infty} \mathbb{E}_{\boldsymbol{x}, \epsilon} [\| \boldsymbol{x} - \hat{\boldsymbol{x}}(\sqrt{\gamma} \boldsymbol{x} + \boldsymbol{\epsilon}, \gamma) \|_2^2] d\gamma \\
&= 1/2 \sum_{i=1}^{d} \int_{\gamma_1}^{\infty} \mathbb{E}_{\boldsymbol{x}, \epsilon} [\| \boldsymbol{x}_i - \hat{\boldsymbol{x}}_i(\sqrt{\gamma} \boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \gamma) \|_2^2] d\gamma
\end{aligned}$$

We separate the analysis into a contribution from each vector component. Next, the possible errors per components will be multiples of Δ, whose appearance depends only on the size of the noise.

$$\begin{aligned}
I_R &\leq 1/2 \sum_{i=1}^{d} \int_{\gamma_1}^{\infty} \sum_{j=1}^{j_{\max}} (\Delta\, j)^2 2P((j-1/2)\Delta \leq \boldsymbol{\epsilon}_i/\sqrt{\gamma} \leq (j+1/2)\Delta) d\gamma \\
&\leq d/2 \int_{\gamma_1}^{\infty} \sum_{j=1}^{j_{\max}} (\Delta\, j)^2 2P((j-1/2)\Delta \leq \boldsymbol{\epsilon}_i/\sqrt{\gamma}) d\gamma \\
&\leq d \int_{\gamma_1}^{\infty} \sum_{j=1}^{j_{\max}} (\Delta\, j)^2 e^{-(j-1/2)^2\Delta^2\gamma/2} d\gamma && \text{(Use Gaussian Chernoff bound)} \\
&= d \int_{\gamma_1}^{\infty} \sum_{j=1}^{j_{\max}} \frac{(\Delta\, j)^2}{(j-1/2)^2\Delta^2/2} e^{-(j-1/2)^2\Delta^2\gamma_1/2} \\
&= 2d \sum_{j=1}^{j_{\max}} \frac{j^2}{(j-1/2)^2} e^{-(j-1/2)^2\Delta^2\gamma_1/2} && \text{(Do integral)} \\
&\leq\ 4\, d \sum_{j=1}^{j_{\max}} e^{-(j-1/2)^2\Delta^2\gamma_1/2} && \text{(Bound } j \text{ term)}
\end{aligned}$$

To summarize, the tail bound constants for our upper bound on discrete likelihood are as follows.

$$\mathbb{E}[-\log P(\boldsymbol{x})] \leq 1/2 \int_{\gamma_0}^{\gamma_1} \mathbb{E}_{\boldsymbol{z}_\gamma, \boldsymbol{x}}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, \gamma)\|_2^2] d\gamma + c(\gamma_0, \gamma_1)$$

$$c(\gamma_0, \gamma_1) \equiv 1/2 \sum_{i=1}^{d} \log(1 + \gamma_0 \lambda_i) + 4\, d \sum_{j=1}^{j_{\max}} e^{-(j-1/2)^2\Delta^2\gamma_1/2}$$

For practical integration ranges, these terms are nearly zero.

B.2 Relationship Between log(SNR) and Timesteps

To use pre-trained models in the literature with our estimator, we need to translate "t", a parameter representing time in a Markov chain that progressively adds noise to data, to a SNR. Referring to Eq. (9) and Eq. (17) in Nichol & Dhariwal (2021), it is easy to build up the mapping from $\alpha = \log(\text{SNR})$ to timestep t in IDDPM. The following derivation shows the relationship:

$$\begin{aligned}
\text{sigmoid}(\alpha) &= \cos\left(\frac{\frac{t}{T}+s}{1+s} \cdot \frac{\pi}{2}\right)^2 / \cos\left(\frac{s}{1+s} \cdot \frac{\pi}{2}\right)^2 \\
\Longrightarrow \cos\left(\frac{\frac{t}{T}+s}{1+s} \cdot \frac{\pi}{2}\right) &= \cos\left(\frac{s}{1+s} \cdot \frac{\pi}{2}\right) \sqrt{\text{sigmoid}(\alpha)} \\
\Longrightarrow t &= T\left(\arccos\left(\cos\left(\frac{s}{1+s} \cdot \frac{\pi}{2}\right) \sqrt{\text{sigmoid}(\alpha)}\right) \frac{2(1+s)}{\pi} - s\right)
\end{aligned}$$

For DDPM, the mapping between α and t is a little bit tricky because it's discrete. Firstly, we construct a one-to-one mapping between the variance β_t and t. In (Ho et al., 2020), the β_t is scheduled in a linear way. Denote $\eta_t = 1 - \beta_t$ and then we have a mapping between $\bar{\eta}_t = \prod_{s=1}^{t} \eta_s$ and t. We match the closest value $\bar{\eta}_t$ with sigmoid(α), and then find the corresponding t of α. Specifically, the mapping between $\log(\text{SNR})$ and t/T (scaled from 0 to 1) is shown in Fig. 4. Our schedule is generated from CIFAR-10 dataset, which gives more diffusion steps for high $\log(\text{SNR})$. Moreover, the schedule adapts along with the dataset by computing scale and mean from it.

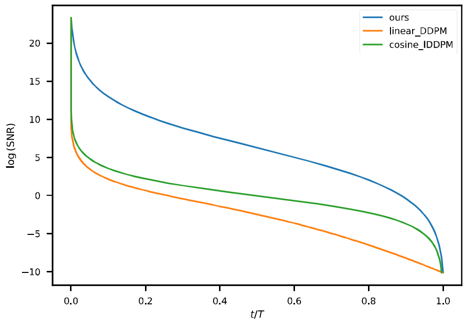
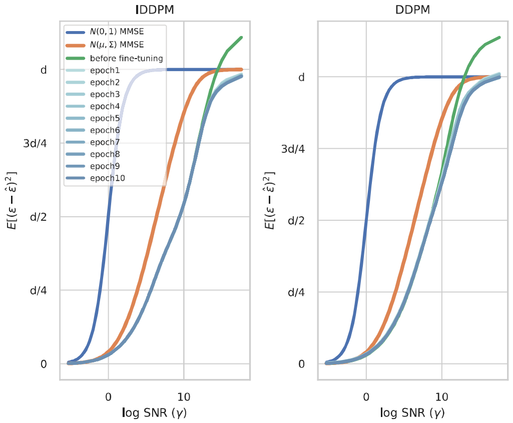


Figure 4: log(SNR) v.s. timestep t/T



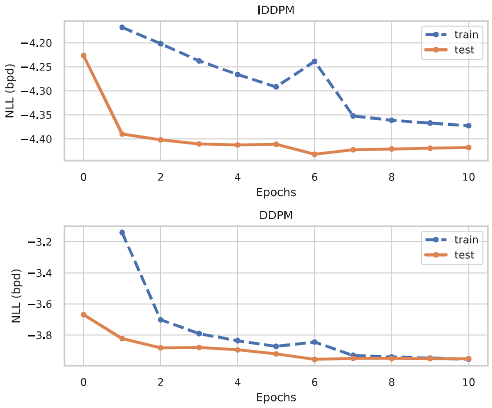


Figure 5: Change of denoising MSE curve during fine-tuning process.

Figure 6: Training and testing NLL during fine-tuning process.

B.3 Details on Model Training

In our fine-tuning experiments, we adopted a IDDPM model provided by (Nichol & Dhariwal, 2021), pre-trained on the unconditional CIFAR-10 dataset with the L_{vlb} objective and cosine schedule. We also consider a pre-trained DDPM model from Hugging Face (`https://huggingface.co/google/ddpm-cifar10-32`), which is provided by (Ho et al., 2020). We process CIFAR-10 32 × 32 dataset in the same way as that in `https://github.com/openai/improved-diffusion/tree/main/datasets`. Since both models are trained with timesteps, we have to convert our log(SNR) values to timesteps before passing them into models. The dataset is scaled to [-1, 1] for each pixel.

For fine-tuning, we train each model for 10 epochs with the same 'learning rate / batch size' ratio in the (Ho et al., 2020) and (Nichol & Dhariwal, 2021), e.g., '$10^{-4}/64$' for DDPM, and '$2.5 \times 10^{-5}/32$' for IDDPM. During training, we reduce the learning rate by a factor of 10 and keep the same batch size after 5 epochs where the training loss starts to be flat. The optimizer for both models is Adam. In testing, we clip denoised images to range [-1, 1], but not during training. The fine-tuning results are displayed in Fig. 5 and Fig. 6. It shows that the fine-tuning improves the NLL value by pushing the MSE curve down when the log(SNR) is high. The NLL results are continuous NLL comparable to Table 1.

B.4 Soft Discretization Function

We found in our experiments that existing denoising architectures were not able to learn the discreteness of the underlying data for the CIFAR dataset, which takes values $g_i = -1 + \Delta i$, with

$\Delta = 2/255, i = 0, \ldots, 255$. At high values of SNR, the noisy input, z, is very close to the true value of x, so zero prediction error can be achieved with high probability just by predicting the nearest discrete value. Therefore, we needed to add a discretization function to the denoiser value at high SNR to improve results. At first we tried a simple function that just rounded to the nearest discrete value. This function was overly aggressive in rounding near borderline values, which sometimes caused an increase in the mean square error. Therefore, we devised a "soft discretization" function. Besides improving the MSE, a soft discretization is also more amenable to being used as a differentiable nonlinearity in a trainable neural network, compared to the hard discretization function.

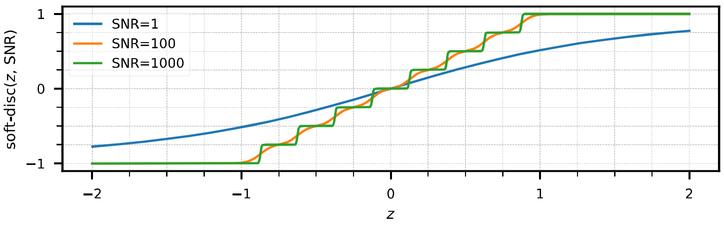


Figure 7: Illustrating the soft discretization function for various values of SNR (γ), for regularly spaced discrete data, $x = -1, -0.75, -0.5, \ldots, 1$.

Consider the noisy channel for scalar random variables, $z = \sqrt{\gamma/(1+\gamma)}x + \sqrt{1/(1+\gamma)}\epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ and $x \sim P(x)$. Let $P(x)$ be a distribution over N discrete grid points, $g_1, \ldots, g_N$. Then we define the soft-discretization function as follows.

$$q(i) = \text{softmax}_i(-0.5(z - g_i)^2(1+\gamma))$$

$$\text{soft-disc}(z; \gamma) = \sum_i g_i q(i)$$

The softmax can be interpreted as a distribution over nearby grid points, where the closest values are most likely. Then we simply take the expected grid point value as the output. The SNR plays the role of a temperature parameter that leads to a more strongly discretizing function at high SNR, and a more linear function at low SNR, as shown in Fig. 7. The form of this nonlinearity was inspired by looking at the optimal case where x is uniform, $P(x = g_i) = 1/N$.

In Fig. 3, we see that this function is not optimal, leading to higher MSEs at low SNR values. However, it performs well at high SNR, so when we ensemble different denoisers we get the best results in Table 2. A more effective strategy would be to replace $(1 + \gamma)$ in the soft discretization function with some learnable function, $f(\gamma)$, then train the neural network accordingly. In this work, we wanted to see that we could improve density estimates over variational bounds by fine-tuning existing architectures with a new objective. Therefore, we used a fixed discretization function instead of a learnable one to avoid adding new parameters or complexity to the model.

B.5 Variance Estimates

We now study two different types of variance estimates for our discrete and continuous log likelihood estimators. We use stochastic estimators for our upper bounds on the negative log likelihood in both cases. We would like to study the variance of these estimators to be sure that the estimates in our results do not look artificially low simply due to noise in the estimators.

Table 3: The standard deviation of NLL (bpd) estimates with information-theoretic bounds

	Continuous	Discrete	Continuous + Dequantization
DDPM	0.00035	0.00112	0.00033
IDDPM	0.00051	0.00088	0.00046
DDPM(tune)	0.00049	0.00100	0.00061
IDDPM(tune)	0.00063	0.00087	0.00059

Our estimators for the discrete and continuous case are simple Monte Carlo estimates, rewritten in the following expressions with irrelevant constants discarded.

$$\mathcal{L}_{cont} \propto \mathbb{E}_{\alpha,\boldsymbol{x},\boldsymbol{\epsilon}} \left[{}^1\!/\!_2 \, 1/q(\alpha) (\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{z}_\gamma, \gamma)\|_2^2 - f_\Sigma(\alpha)) \right]$$

$$\mathcal{L}_{disc} \propto \mathbb{E}_{\alpha,\boldsymbol{x},\boldsymbol{\epsilon}} \left[{}^1\!/\!_2 \, 1/q(\alpha) \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{z}_\gamma, \gamma)\|_2^2 \right]$$

The random variables are $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I}), \boldsymbol{x} \sim p(\boldsymbol{x}), \alpha \sim q(\alpha)$, where $q(\alpha)$ is the importance sampling distribution described in Sec. 4. Empirical estimates are taken by drawing samples from these distributions and computing an empirical mean. This produces an unbiased estimate that converges to the true expectation as we include more samples. Due to the central limit theorem, we know that the distribution of this estimator will converge to a Gaussian whose mean is the true expectation and whose variance is equal to the variance of samples divided by the total number of samples, N. We used this result to estimate the standard deviation of the estimators in Sec. 5, where we used $n = 100$ samples of $\alpha \sim q(\alpha)$ each with 10000 points from $\boldsymbol{x} \sim p(\boldsymbol{x}), \boldsymbol{\epsilon} \sim N(0, 1)$, so that $N = 10^6$. For the discrete case, we used $n = 1000$ samples from $q(\alpha)$ per point in results in Sec. 5, but we report the result here using $n = 100$ samples so that numbers are directly comparable.

The results are shown in Table 3, and we see that variance estimates are small compared to the values reported in Fig. 2. Note that the continuous density estimators have smaller variance than the discrete estimator. This makes sense because the importance sampling distribution that we used was chosen to match the integrand for the continuous density estimator. The more closely the importance sampling distribution matches the target, the lower the variance. The higher variance for the discrete estimator is due to the mismatch between the curves in Fig. 3 and the logistic distribution, and we observe a similar phenomenon using the bootstrap estimators studied next. Therefore, we used more samples from $q(\alpha)$ when using the discrete probability estimator.

This analysis has an important caveat. These variance estimates reflect the case where all samples are drawn IID. However, for the purposes of plotting and ensembling, it is more convenient to have errors on a common set of $\log$(SNR) values, so that we can estimate average MSE's across samples per $\log$(SNR) ($\mathbb{E}_{p(\boldsymbol{x})}[\text{mmse}(\boldsymbol{x}, \gamma)] = \text{mmse}(\gamma)$). Because we use the same set of $\log$(SNR) samples (drawn IID from $q(\alpha)$) across *all* samples, these samples are not fully IID. Hence, the standard error estimates in Table 3, which assume fully IID samples, are imperfect. This prompted us to include an alternate analysis based on bootstrap sampling.

Table 4: Bootstrap variance estimates for continuous estimator

	$n = 1000$	Bootstrap samples with $n = 100$
DDPM	-3.59	-3.56 ± 0.04
IDDPM	-4.14	-4.09 ± 0.10
DDPM(tune)	-3.87	-3.84 ± 0.07
IDDPM(tune)	-4.31	-4.28 ± 0.11
Ensemble	-4.31	-4.29 ± 0.11

Table 5: Bootstrap variance estimates for discrete estimator

	$n = 1000$	Boot. $n = 100$
DDPM	3.68	3.56 ± 0.32
IDDPM	3.16	3.05 ± 0.24
DDPM(tune)	3.48	3.36 ± 0.28
IDDPM(tune)	3.15	3.04 ± 0.24
Ensemble	2.90	2.79 ± 0.27

Table 6: Bootstrap variance estimates using uniform dequantization

	$n = 1000$	Boot. $n = 100$
DDPM	3.47	3.51 ± 0.04
IDDPM	3.12	3.17 ± 0.07
DDPM(tune)	3.37	3.41 ± 0.05
IDDPM(tune)	3.13	3.18 ± 0.07
Ensemble	3.11	3.16 ± 0.07

Bootstrap sampling analysis To get a sense for the variance of test time estimates with IID random log SNR values using our data consisting of a fixed set of random $\log$(SNR) values in common across samples, we did a bootstrap sampling analysis. For this analysis, we first constructed estimates using the full fixed set of $n = 1000$ samples of $\alpha \sim q(\alpha)$ across all samples. Then, we took 10 random subsets of $n = 100$ $\log(\text{SNR})$ values and again estimated NLL values. We show the mean and standard deviation across the bootstrap samples in each case. The results for different estimators are shown in Tables 4, 5, 6.

The use of non-IID $\log(\text{SNR})$s does lead to higher variance. In principle, this could be avoided, even in the case when we are ensembling. To do so, we could use a validation set to determine which denoising model to use in each SNR range. Then, at test time, we could use fully IID samples for each $\log(\text{SNR})$ value.

Note that the continuous estimators (including uniform dequantization, which uses a continuous estimator to give a discrete probability estimate) have far lower variance. This is why we used fewer $\log(\text{SNR})$ samples for continuous versus discrete estimators in the main results.

Finally, note that the average bootstrap NLL using $n = 100$ in the discrete case, for example 2.79 for the ensembling (Tab. 5), is actually lower than the best result reported in the main experiment section, NLL = 2.90 using $n = 1000$. We chose to report the larger but more reliable result using $n = 1000$ since the variance of the discrete estimator using $n = 100$ is large.

Diffusion models are computationally expensive and each point requires many calls per sample at test time to estimate log likelihood. Our variance analysis shows that our continuous density estimator (which can be used for discrete estimation also using uniform dequantization) can significantly reduce variance, leading to reliable estimates with far fewer model evaluations.