# MAT1525 Project, Geometry Adaptive Representations in Diffusion Models

Anton Sugolov

April 20, 2025

## Contents



# 1 Introduction

Diffusion modeling has recently emerged as a powerful paradigm for generative models, where generation is implemented through iterative denoising. A central question in diffusion modeling is the exact mechanism of operation. In this project, we focus on the work of Kadkhodaie et al. [4], which addresses the implicit representations of UNet denoisers, the memorization phenomenon in generative models, and the optimality of denoising representations. We begin by reviewing related results in classical denoising (Section 4), present the work of Kadkhodaie et al. [4] (Section 3), and conclude with experiments on

recent single-step formulations of image generation (Section 4).

Denoising in the imaging context aims to recover an image $x \in \mathbb{R}^d$ that is corrupted by additive noise. In many settings, the corruption is modelled by $z \sim \mathcal{N}(0, \sigma^2 I)$ with density $g_\sigma(z) = \frac{1}{\sqrt{2\pi\sigma}^{n/2}} \exp(-z^T z/2\sigma^2)$. If the data $x$ has probability density $p_D$, the density of noisy images $y$ can be expressed by the convolution

$$p_\sigma(y) = \int_{\mathbb{R}^d} p_\sigma(y \mid x) p_D(x)\, dx = \int_{\mathbb{R}^d} g_\sigma(y - x) p_D(x)\, dx$$

The corrupted density $p_\sigma$ can be used to sample from $p_D$ through **score-based sampling** [9]. For $y_0 = z \sim \mathcal{N}(0, \sigma^2 I)$, steps of Langevin Monte-Carlo style sampling are implemented with the score $(1 \leq t \leq T)$.

$$y_{t+1} = y_t + \frac{\sigma^2}{2} \nabla \log_y p_\sigma(y_t) + z_i$$

as $T \to \infty$ and $\sigma \to 0$, it can be shown that $y_T \sim p_D$ under some regularity assumptions [9]. In practice, the score is approximated by a neural network $\nabla \log_y p_\sigma(y) \approx s_\theta(y)$, and the above sampling is accelerated by repeating the above sampling for $L$ noise levels $\sigma_1 < \sigma_2 < \cdots < \sigma_L$. These methods have shown remarkable generatlization abilities for generating images, and several interesting empirical questions arise in this setting.

1. The score $s_\theta(x) \approx \nabla \log_y p_\sigma(y)$ and the conditional expectation of the clean image $\mathbb{E}[x \mid y]$ have an exact relationship (Section 3). The score is typically learned via a U-Net (Section 2.5) neural network, which captures multi-scale image representations by design. The representations learned by the UNet provide insight into the structure of the train dataset that is represented within the a denoising process.

2. $p_D \sim \{x_i\}_{i=1}^N$ are discrete points which are uniformly sampled during training, therefore

$$p_\sigma(y) = \frac{1}{n} \sum_{i=1}^{n} g_\sigma(y - x_i)$$

   can be viewed as a uniform mixture of Gaussians. If the number of points is small, it is clear that score-based sampling leads to memorization of the training data. However, empirically, score-based models are able to generalize with extremely high accuracy as the number of data points $N$ becomes large.

These relationships motivate the understanding of generalization in score-based models through the structure present in trained denoisers. In this project, we review classical ideas from denoising, optimality and their relationship to generalization in score-based generative models, as presented by Kadkhodaie et al. [4].

# 2 Denoising

## 2.1 The structure of natural images

To motivate the discussion of possible representations that underly the diffusion model, we consider the structure present in images [14]. Low dimensional image decompositions enable denoisers to achieve optimality (Section 2.2) and this structure coincides with natural image data.

1. **Power law decay.** [13, 14] The amplitude ($f$) spectra (cf. Section 2.2) of natural images show a decay rate of $\sim f^{-\alpha}$, where typically $\alpha \approx 2$.

2. **Low-dimensional manifolds.** [1] Projections of a Lambertian surface (eg. a face with different lighting) form low-dimensional subspaces in pixel space, allowing for the adoption of a sparse basis that captures the underlying data structure.

3. **Average spectrum.** [14] The average frequency spectrum of various scenes has a distinctly identifiable structure. In 2003, Torralba and Oliva [14] proposed a way to exploit this structure to build spectrum-based filters for scene classification.

This structure motivates the discussion of the optimal PSNR slope under assumptions on the power spectrum (Sections 2.2, 2.3), the generalization and interpolation of sampling (Sections 1, 3), and the representations present in the UNet featuring data-dependent convolutional layers (Section 2.5).

## 2.2 Linear denoisers

The simplest denoisers, namely $f$, are linear on the corrupted images $f(\lambda y_1 + y_2) = \lambda f(y_1) + f(y_2)$. This setting ensures that the basis $(\psi_k)_k$ that $f$ operates in must optimally coincide with features of the dataset. In this section, we further discuss certain properties of a linear denoiser in an orthogonal basis.

### Oracle denoiser

Suppose that the ground truth $x$ is known and that $y = x + z$ is corrupted by Gaussian white noise. Consider the linear denoiser $f$, which is represented as a sum of projections onto some orthonormal basis $(\psi_k)_k$.

$$f(y) = \sum_k \lambda_k \langle y, \psi_k \rangle \psi_k$$

We denote $\hat{x} = f(y)$ as the denoised image. Assuming access to the ground truth $x$, we identify the optimal scaling $\lambda_k(x)$ such that $f$ minimizes the mean square error (MSE) between $x$ and $\hat{x}$.

$$\mathbb{E}_z \|\hat{x} - x\|^2 = \sum_k \mathbb{E}_z |\langle \hat{x} - x, \psi_k \rangle|^2 = \sum_k \mathbb{E}_z |\lambda_k(\langle x, \psi_k \rangle + \langle z, \psi_k \rangle) - \langle x, \psi_k \rangle|^2$$

Since $\mathbb{E}_z \langle z, \psi_k \rangle = 0$ by linearity of expectation, we expand the above by

$$= \sum_k \langle x, \psi_k \rangle^2 (\lambda_k - 1)^2 + \lambda_k^2 \sigma^2$$

The first order conditions for each quadratic term give

$$(\lambda_k - 1)\langle x, \psi_k \rangle^2 + \lambda_k \sigma^2 = 0 \implies \lambda_k(x) = \frac{\langle x, \psi_k \rangle^2}{\langle x, \psi_k \rangle^2 + \sigma^2}$$

The above optimal $\lambda_k(x)$ assumed knowledge of the ground truth $x$.

### Wiener filter

The **Wiener filter** accounts for $x$ following some distribution $x \sim X$ by accounting for its expectation in the MSE. For $\alpha_k = \mathbb{E}\,|\langle X, \psi_k\rangle|^2$, an identical calculation shows

$$\begin{aligned}\mathbb{E}_{X,z}\,\|f(X+z)-X\|^2 &= \sum_k \mathbb{E}_X\,|\langle X,\psi_k\rangle|^2\,(\lambda_k-1)^2+\lambda_k^2\sigma^2\\ &= \sum_k \alpha_k\,(\lambda_k-1)^2+\lambda_k^2\sigma^2\end{aligned}$$

Minimizing for the optimal $\lambda_k$ as before, we find that

$$\lambda_k = \frac{\alpha_k}{\alpha_k+\sigma^2}, \qquad \alpha_k = \mathbb{E}\,|\langle X,\psi_k\rangle|^2$$

The $\alpha_k$ are typically called the **power spectrum** of the underlying data distribution $x \sim X$. Under certain empirical assumptions, for example $\alpha_k$ decaying with a power law, it is possible to further derive some optimality results of the error with respect to $\psi_k$.

### Analysis of decay rate

We consider the $\lambda_k(x) = \frac{\langle x,\psi_k\rangle^2}{\langle x,\psi_k\rangle^2+\sigma^2}$ from Section 2.2. $\lambda_k(x)$ acts as a type of soft threshold for $\psi_k$, where projection gradually occurs as signal dominates the noise. To observe this, we find that the MSE of each term is of the order of $\min(|\langle x,\psi_k\rangle|^2, \sigma_k^2)$.

$$\mathbb{E}_z\,\|\hat{x}-x\|^2 = \sum_k \frac{\sigma^2\,|\langle x,\psi_k\rangle|^2}{|\langle x,\psi_k\rangle|^2+\sigma^2} \sim \sum_k \min(|\langle x,\psi_k\rangle|^2,\sigma_k^2)$$

The calculation follows due to the inequality $\frac{1}{2}\min(a,b) \le ab/(a+b) \le \min(a,b)$, which gives the order of decay of the coefficients. We may arrange basis elements by $|\langle x,\psi_k\rangle|^2 > \sigma^2$:

$$\sim \underbrace{\sum_{|\langle x,\psi_k\rangle|^2>\sigma^2} \sigma^2}_{M \text{ terms}} + \sum_{|\langle x,\psi_k\rangle|^2<\sigma^2} |\langle x,\psi_k\rangle|^2$$

Above, $\psi_1,\dots,\psi_K$ represent terms with signal dominating the noise $\langle x,\psi_k\rangle^2 < \sigma^2$. This motivates denoising by using the truncated $M$-term representation $x_M = \sum_{k\le M}\langle x,\psi_k\rangle\psi_k$, giving $\|x-x_M\|^2 = \sum_{|\langle x,\psi_k\rangle|^2<\sigma^2}|\langle x,\psi_k\rangle|^2$. The MSE is then of the order of

$$\mathbb{E}_z\,\|\hat{x}-x\|^2 \sim M\sigma^2 + \|x-x_M\|^2 \sim \sigma^{2\alpha/(\alpha+1)}$$

In order for $(\psi_k)_k$ to minimize the MSE, it is advantageous for $x$ to have an accurate sparse representation in $(\psi_k)_k$. A typical empirical assumption is that $\langle x,\psi_k\rangle^2 \sim k^{-(\alpha+1)}$ which is true for natural images (Section 2.1). We next consider a general result about the decay of the MSE when $\langle x,\psi_k\rangle^2 \sim k^{-(\alpha+1)}$ for some basis $(\psi_k)_k$.

**Theorem 1.** *If $f$ is a linear denoiser represented in the basis $(\psi_k)_k$ and there exist $c, c'$ independent of $x, k$ such that $c\, k^{-(\alpha+1)} \leq \langle x, \psi_k \rangle^2 \leq c'\, k^{-(\alpha+1)}$, (denoted $\langle x, \psi_k \rangle^2 \sim k^{-(\alpha+1)}$) then*

$$\mathbb{E}\, \|x - \hat{x}\|^2 \sim \sigma^{2\alpha/(\alpha+1)}$$

**Note:** The **optimal PSNR slope** is $\frac{\alpha}{\alpha+1}$. That is, the lower bound of the decay rate of $\mathbb{E}\, \|x - \hat{x}\|^2$ with respect to $\sigma^2$ is of order $\frac{\alpha}{\alpha+1}$.

*Proof.* We begin by labelling the $\psi_k$ such that $\langle x, \psi_i \rangle^2 \geq \cdots \geq \langle x, \psi_{i+1} \rangle^2$. Note that the ordering depends on $x$. Let $M$ be the largest index with $\langle x, \psi_k \rangle^2 > \sigma^2$.

$$\langle x, \psi_M \rangle^2 > \sigma^2 \geq \langle x, \psi_{M+1} \rangle^2,$$

so that

$$c'\, M^{-(\alpha+1)} > \sigma^2 \geq c\, (M+1)^{-(\alpha+1)}.$$

We then have $M^{-(\alpha+1)} \sim \sigma^2$, i.e., $M \sim \sigma^{-2/(\alpha+1)}$, and thus $M\sigma^2 \sim \sigma^{2\alpha/(\alpha+1)}$. We also have

$$\sum_{k>M} \langle x, \psi_k \rangle^2 \leq c' \sum_{k>M} k^{-(\alpha+1)} \leq c' \int_M^{+\infty} t^{-(\alpha+1)}\, dt = \frac{c'}{\alpha} M^{-\alpha},$$

$$\sum_{k>M} \langle x, \psi_k \rangle^2 \geq c \sum_{k>M} k^{-(\alpha+1)} \geq c \int_{M+1}^{+\infty} t^{-(\alpha+1)}\, dt = \frac{c}{\alpha} (M+1)^{-\alpha},$$

so that $\|x - x_M\|^2 \sim M^{-\alpha} \sim \sigma^{2\alpha/(\alpha+1)}$. Previously, we showed that the terms in the expansion of the MSE are of the same order $\min(|\langle x, \psi_k \rangle|^2, \sigma_k^2)$, and it follows that $\mathbb{E}\, \|x - \hat{x}\|^2 \sim \sigma^{2\alpha/(\alpha+1)}$. □

## 2.3 Optimal $C^\alpha$ bases

A particular class of images that admit a basis $(\psi_k)_k \sim k^{-(\alpha+1)}$ are the $C^\alpha$ images. Heuristically, this class of images is highly regular except possibly at the edges, which are regular as curves. The optimal bases for representing such images are known [7]. In Section 3 we discuss the optimality of U-Net denoisers trained on $C^\alpha$ images, demonstrating the robustness of the representations learned by these models. We begin by surveying the optimality of $C^\alpha$ images in the bandlet basis.

**Definition 1.** A function f is **uniformly $\alpha$-Lipschitz** over a domain $\Omega$ if there exists a constant $C$ such that for all $x \in \Omega$, there exists a polynomial $q_x$ of degree $\lfloor \alpha \rfloor$ such that for all $y \in \Omega$, $|f(y) - q_x(y)| \leq C|x - y|^\alpha$.

**Definition 2.** The image $x : [0,1]^2 \to \mathbb{R}$ is **$C^\alpha$-geometrically-regular** if it is uniformly $\alpha$-Lipschitz over $[0,1]^2 \setminus \cup_i \gamma_i$ where the $\{\gamma_i\}_{i=1}^d$ are $\alpha$-Lipschitz curves in $[0,1]^2$.

Peyré and Mallat [7] show that an optimal approximation of a $C^\alpha$ image $x$, can be constructed by a wavelet basis adapted to the geometry of the image. A wavelet basis is taken with respect to the image, which gives a transform with many coefficients that can be threshholded. To accelerate computation, a segmentation is applied to capture particular geometric features, and a localized **bandelet basis** is

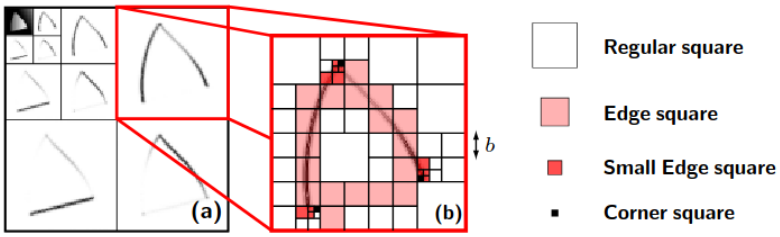



Figure 1: (from Peyré and Mallat [7]) (a) wavelet transform of the $C^\alpha$ image (b) segmentation of image into dyadic geometry-adapted regions on which the bandelet basis is supported.

introduced on each new subdivision. The bandelet basis is parametrizable, allowing each step to be computable in practice.

**Proposition 1** (Peyré and Mallat [7], Lemma 6.2.). *Let $f$ be a $C^\alpha$-geometrically-regular image. There exists $C$ such that for all $T > 0$ there exists a bandelet basis $\mathcal{B}(\Gamma)$ in which the truncated approximation $x_M$ satisfies*

$$\|x - x_M\|_{L^2}^2 \leq C T^{\frac{2\alpha}{\alpha+1}} \quad \textit{and} \quad M \leq C T^{-\frac{2}{\alpha+1}}.$$

Remarkably, the image features that are implicitly represented in UNet denoisers show similar optimality (cf. Section 3).

## 2.4 Score-based denoising

We relate the ideas of denoising to the score, which will be key elements of the observations of Section 3. We show that in denoising, the MSE optimal denoiser is $\mathbb{E}[x \mid y]$ (Proposition 2). Additionally, when $z \sim N(0, \sigma^2 I)$ is Gaussian, there is an exact relationship between the score and the conditional expectation (Proposition 3).

**Proposition 2** (Optimal Denoiser). *The conditional expectation minimizes the mean-square error among all $f \in {}^2(\mathbb{R}^n)$:*

$$\mathbb{E}[x \mid y] = \arg \min_{f \in L^2(\mathbb{R}^n)} \|x - g(y)\|^2$$

*Proof.* By adding and subtracting $\mathbb{E}[x \mid y]$ and using its properties, we may establish

$$\mathbb{E}_x \left[(x - f(y))^2\right] = \mathbb{E}_x \left[(x - \mathbb{E}[x \mid y])\right]^2 + \mathbb{E}\left[(\mathbb{E}[x \mid y] - f(y))\right]^2$$

Note that the left hand term is a constant and that the right hand term is minimized exactly when $\mathbb{E}[x \mid y] = f(y)$. □

**Proposition 3** (Miyasawa relationships, 1962)**.** [a] *For $y = x + z$ where $z \sim N(0, \sigma^2 I)$ and $x \sim p_D(x)$, the score and conditional covariance can be prepresented by*

$$f^\star(y) = \mathbb{E}[x \mid y] = y + \sigma^2 \nabla \log p(y)$$
$$Df^\star(y) = Cov[x \mid y] = \sigma^2 (I + \sigma^2 \nabla^2 \log p(y))$$

[a] These relationships have been established for a long time!

*Proof.* We begin by recalling that the noisy density is given by

$$p(y) = \int p(x)\, p(y \mid x)\, dx$$

For any function $h$, the logarithmic derivative yields $\nabla \log h(y) = \frac{1}{h(y)} \nabla h(y)$. Applying this above twice,

$$\begin{aligned}
\nabla \log p(y) &= \frac{1}{p(y)} \int p(x)\, \nabla_y p(y \mid x)\, dx \\
&= \frac{1}{p(y)} \int p(x)\, \nabla_y p(y \mid x)\, dx \\
&= \frac{1}{p(y)} \int p(x)\, p(y \mid x)\, \nabla_y \log p(y \mid x)\, dx \\
&= \int p(x \mid y)\, \nabla_y \log p(y \mid x)\, dx \\
&= \mathbb{E}[\nabla_y \log p(y \mid x) \mid y]
\end{aligned}$$

which can be interpreted as a chain rule on scores. To compute the second derivative, we find

$$\nabla^2 \log p(y) = \int p(x \mid y) \left( \nabla_y \log p(x \mid y) \nabla_y \log p(y \mid x)^T + \nabla_y^2 \log p(y \mid x) \right) dx$$

By logarithmically differentiating Bayes rule,

$$\nabla_y \log p(x \mid y) = \nabla_y \log p(y \mid x) - \nabla_y \log p(y)$$

Applying this to the earlier equality, we find

$$\begin{aligned}
\nabla^2 \log p(y) &= \int p(x \mid y) \left( (\nabla_y \log p(y \mid x) - \nabla_y \log p(y)) \nabla_y \log p(y \mid x)^T + \nabla_y^2 \log p(y \mid x) \right) dx \\
&= \mathrm{Cov}[\nabla_y \log p(y \mid x) \mid y] + \mathbb{E}[\nabla^2 p(x \mid y) \mid y]
\end{aligned}$$

Recall that $y$ is obtained from $x$ by adding noise of variance $\sigma^2$. Therefore

$$\begin{aligned}
\log p(y \mid x) &= -\frac{1}{2\sigma^2} \|y - x\|^2 + C \\
\nabla_y \log p(y \mid x) &= -\frac{1}{\sigma^2}(y - x) \\
\nabla_y^2 \log p(y \mid x) &= -\frac{1}{\sigma^2} I
\end{aligned}$$

From the previous equations, we conclude

$$\begin{aligned}\nabla \log p(y) &= \mathbb{E}\left[\nabla_y \log p(y \mid x) \mid y\right] \\ &= \frac{1}{\sigma^2}(\mathbb{E}\left[x \mid y\right] - y) \\ \nabla^2 \log p(y) &= \mathrm{Cov}\left[\nabla_y \log p(y \mid x) \mid y\right] + \mathbb{E}\left[\nabla^2 p(x \mid y) \mid y\right] \\ &= \frac{1}{\sigma^4}\mathrm{Cov}[x \mid y] - \frac{1}{\sigma^2} I\end{aligned}$$

We therefore find

$$\begin{aligned}\mathbb{E}\left[x \mid y\right] &= y + \sigma^2 \nabla \log p(y) \\ \mathrm{Cov}[x \mid y] &= \sigma^2 (I + \sigma^2 \nabla^2 \log p(y))\end{aligned}$$

which gives the desired equalities. □

**Consequences**

There are several natural consequences that follow from the above lemmas.

1. A denoiser $f$ trained to minimize the mean square error $\mathbb{E}\left\|x - f(y)\right\|^2$ optimally achieves $f(y) = \mathbb{E}\left[x \mid y\right]$. In the setting of Proposition 3, we can re-express the denoiser as

$$f(y) = y + \sigma^2 \nabla \log p(y) \implies \nabla \log p(y) = \frac{1}{\sigma^2}(f(y) - y) \tag{1}$$

2. Differentiating $f(y)$ above, we find

$$Df(y) = I + \sigma^2 \nabla^2 \log p(y) \implies Df(y) \propto \mathrm{Cov}[x \mid y]$$

   The optimal Jacobian therefore captures the conditional covariance structure of the data given $y$. We therefore expect the eigendecomposition of $Df(y)$ to represent some local data-dependent structure (cf. Section 2.2).

The following examples explicitly demonstrate the Miyasawa relationships.

**Example 1.** If the training set consists of $\{x\}$, then $p(y) \sim N(x, \sigma^2 I)$, and it is easy to see that

$$x = y + \sigma^2 \nabla \log p(y)$$

A denoiser would simply memorize $x$. Sampling from the denoiser residual of Equation 1 would correspond to the memorization phenomenon in diffusion models.

**Example 2.** For training examples $\{x_i\}_{i=1}^n$ the uniform sampling measure of the train data can be written as $p_D(x) = \frac{1}{n}\sum \delta_{x_i}$. In this case,

$$\log p(y) = \log\left(\frac{1}{n}\sum_{i=1}^{n} g_\sigma(y - x_i)\right)$$

$$\nabla \log p(y) = -\frac{1}{\sigma^2 \sum_{i=1}^n g_\sigma(y - x_i)} \cdot \sum_{i=1}^{n}(y - x_i)$$

We see that this corresponds to a weighted sum in the direction of the nearest data points. Areas of low probability generating a high score due to $(\sum_{i=1}^n g_\sigma(y - x_i))^{-1}$ and $\|\nabla \log p(y)\|$ scales with $\sigma^{-2}$.

The above implies that as the number of datapoints increases, and as the score varies, a trained denoiser must learn to interpolate nearby data points in a data-dependent way. A natural question is about the bounds that control the approximation error of a learned denoiser $f_\theta$. The next result shows that the error in estimating the true density $p(x)$ is controlled by minimizing the denoising error for all noise levels $\sigma^2$.

**Proposition 4.** *Let $f_\theta(x)$ be a learned denoiser and $s_\theta(x) = \frac{1}{\sigma^2}(y - f_\theta(x))$. For underlying data density $p(x)$, noisy density $p_\sigma(y)$, and learned density $p_\theta(x)$, $D_{KL}(p(x)\|p_\theta(x))$ is controlled by the MSE of denoiser $f_\theta$ at all noise levels*

$$D_{KL}(p(x)\|p_\theta(x)) \leq \int_0^\infty \left(MSE(f_\theta, \sigma^2) - MSE(f^*, \sigma^2)\right)\sigma^{-3}\, d\sigma,$$

*where $f^* = \mathbb{E}[x \mid y]$ is the optimal denoiser. In particular, $D_{KL}$ is minimized when $f_\theta = f^*$.*

*Proof.* From the results of Song et al. [11], the KL divergence is controlled by the score-matching objective by the inequality

$$D_{\mathrm{KL}}(p(x)\|p_\theta(x)) \leq \int_0^\infty \mathbb{E}_y\left[\|\nabla \log p_\sigma(y) - s_\theta(y)\|^2\right]\sigma\, d\sigma$$

By inserting the results of Propositions 2 3, we find that

$$\begin{aligned}\|\nabla \log p_\sigma(y) - s_\theta(y)\|^2 &= \frac{1}{\sigma^4}\mathbb{E}\left[\|\mathbb{E}[x|y] - f_\theta(y)\|^2\right] \\ &= \mathbb{E}\left[\|x - f_\theta(y)\|^2\right] = \mathbb{E}\left[\|x - \mathbb{E}[x|y]\|^2\right] + \mathbb{E}\left[\|\mathbb{E}[x|y] - f_\theta(y)\|^2\right]\end{aligned}$$

We therefore have

$$\mathbb{E}\left[\|\nabla \log p_\sigma(y) - s_\theta(y)\|^2\right] = \frac{1}{\sigma^4}\left(\mathbb{E}\left[\|x - f_\theta(y)\|^2\right] - \mathbb{E}\left[\|x - \mathbb{E}[x|y]\|^2\right]\right)$$

$$= \frac{1}{\sigma^4}\left(\mathrm{MSE}(f_\theta, \sigma^2) - \mathrm{MSE}(f^*, \sigma^2)\right).$$

Combining with the bound given by the score-matching objective, this proves the proposition. □

**Proposition 5** (Stein's unbiased risk estimate)**.** *The MSE of a denoiser $f$ may be written by the statistical estimator*

$$\mathbb{E}\left[\|x - f(y)\|^2\right] = \mathbb{E}\left[\|y - f(y)\|^2\right] + 2\sigma^2\mathbb{E}[tr\ \nabla f(y)] - \sigma^2 d$$

*Proof.* The MSE may be written as

$$\begin{aligned}\mathbb{E}\left[\|x - f(y)\|^2\right] &= \mathbb{E}\left[\|(y - f(y)) - (y - x)\|^2\right] \\ &= \mathbb{E}\left[\|y - f(y)\|^2\right] - 2\mathbb{E}[(y - x, y - f(y))] + \mathbb{E}\left[\|y - x\|^2\right]\end{aligned}$$

The last term is the total variance of the noise and equals $\sigma^2 d$. The middle term can be rewritten with an integration by parts since $y - x = -\sigma^2\nabla_y \log p(y|x)$.

$$\begin{aligned}\mathbb{E}[(y - x, y - f(y))] &= -\sigma^2 \iint \langle \nabla_y \log p(y|x), y - f(y)\rangle p(x)p(y|x)\, dx\, dy, \\ &= -\sigma^2 \iint \langle \nabla_y p(y|x), y - f(y)\rangle p(x)\, dx\, dy, \\ &= \sigma^2 \iint \mathrm{tr}(\mathrm{Id} - \nabla f(y))p(x)p(y|x)\, dx\, dy \\ &= \sigma^2\mathbb{E}[d - \mathrm{tr}\ \nabla f(y)]\end{aligned}$$

Inserting into the original equation, we arrive at the desired equality, which proves the proposition. □

## 2.5 UNet Architecture

The U-Net is a widely adopted convolutional architecture for segmentation, denoising, and generation tasks [2, 8, 10]. The model performs both downscaling and upscaling operations, which form a U-like shape, displayed in Figure 2.
There are several notable facts about the UNet.

1. If each convolutional layer does not include any biases ($\mathrm{conv}(x) = Wx + b$ with $b = 0$), then the UNet is a composition of linear and piecewise linear operations. That is $\hat{x} = f_\theta(y)$ can be written as
$$\hat{x} = W(y)y$$
where $W(y) \in \mathbb{R}^{d\times d}$ is a $y$-dependent matrix that implements the denoising step.

2. The down-sampled representations are also copied to the representations at each up-sampling step, suggesting that the layers learn scale-dependent features of the underlying data. This coincides with the construction of the $C^\alpha$ optimal of bandelet basis, which use scale-dependent wavelets adapted to the geometry of the particular image (Section 2.3).

These observation and results of Section 2.4 motivate considering the linearization of a score-based generative model that is implemented as a UNet.

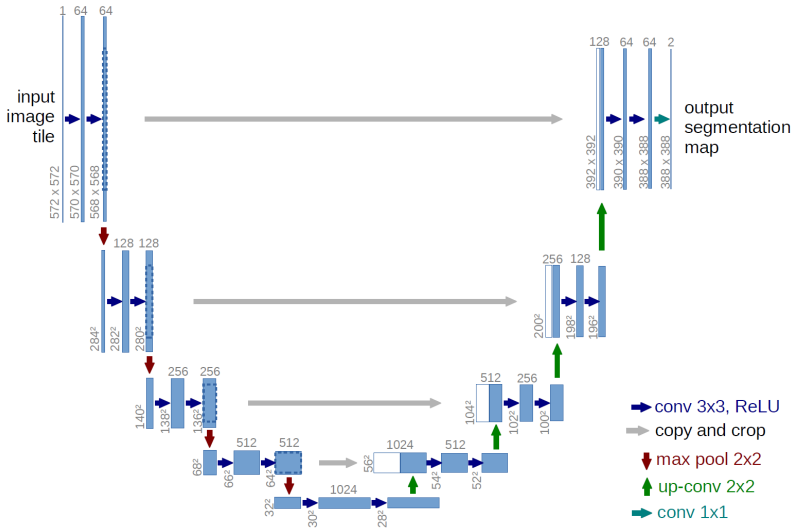



Figure 2: (from Ronneberger, Fischer, and Brox [8].) The UNet architecture. In the downsampling stage, a series of $3 \times 3$ convolutions with ReLU nonlinearities and $2 \times 2$ max-pooling to generate the next scale. During the up-sampling stage, $2 \times 2$ max-pooling and up-convolutions are used in addition to the concatenation of features features from the earlier down-sampling stage. In this example, a smaller output resolution is generated. The same output resolution can be achieved by setting the same filter size during downsampling and upsampling.

# 3 Kadkhodaie et al. [4], *Generalization in diffusion models arises from geometry-adaptive harmonic representations*.

*Generalization in diffusion models arises from geometry-adaptive harmonic representations* by Kadkhodaie et al. [4] is a recent empirical work that unites the ideas of Section 2 to analyze the representations that are learned by a UNet denoiser. For a trained UNet denoiser $f_\theta$, the authors demonstrate that (1) the eigenvectors of $Df_\theta(y)$ acquire a learned data-dependent structure, (2) sampling from the denoiser residual $f_\theta$ transitions from memorizing training examples to interpolation, and (3) the learned basis vectors achieve optimal PSNR decay (cf. Section 2.2)

## 3.1 Overview

A piece-wise linear UNet denoiser $f_\theta$ is trained to minimize MSE of train samples $\{x_i\}_{i=1}^n$. Corrupted samples $y = x + \sigma z$, $z \sim N(0, I)$ are generated over various noise levels $\sigma^2$ (in light of Proposition 4) and trained to recover the clean image. The learned parameters $\theta$ minimize

$$\hat{\theta} = \arg\min_\theta \mathbb{E}_{x \sim p_D,\, z,\, \sigma} \left[ \| f_\theta(x + \sigma z) - x \|^2 \right] \tag{2}$$

In Section 2.4, we showed $Df_\theta(y) \approx \text{Cov}[x \mid y]$. By piecewise linearity, we may compute the eigendecomposition of $Df_\theta$ at the particular $y$, namely $\{e_k(y)\}_{k=1}^n$ with eigenvalues $\{\lambda_k(y)\}_{k=1}^d$. In particular,

$$Df_\theta(y) = \sum_{k=1}^{d} \lambda_k(y) \langle y, e_k(y) \rangle \, e_k(y)$$

The approximated score may be represented by $s_\theta(y) = \frac{1}{\sigma^2}(f_\theta(y) - y)$ and sampling with the UNet denoiser expoits this fact in this way. The Stein unbiased risk estimate (Proposition 5) yields that

$$\text{MSE}(f_\theta, \sigma^2) = \mathbb{E}_y \left[ 2\sigma^2 \underbrace{\text{tr } Df_\theta(y)}_{\text{rank penalty}} + \underbrace{\|y - f_\theta(y)\|^2}_{\text{distance penalty}} - \sigma^2 d \right]$$

The SURE suggests that to minimize the MSE, tr $Df_\theta(y)$ must be minimized, and therefore the basis $\{e_k(y)\}$ must sparsely capture the structure of the underlying image. This further motivates the question of memorization and generalization. For few examples, memorization may lead to a favourable MSE. However, with many datapoints, some learned interpolation of training images is necessary to achieve the optimum in the above estimator. This is in addition to the observations of Section 2.2, which suggest optimality of sparse data representations in the simple setting of oracle denoisers.

## 3.2 Contributions

The authors present several striking results demonstrating the transition from memorization to generalization in diffusion models.

1. **Data-dependent basis.** The eigenvectors of $Df_\theta(y)$, $\{e_k(y)\}_{k=1}^d$ are adapted to the features of the clean image. In UNets that are trained on sufficient training data, $e_k(y)$ interpolate the training data while the clean image is sparsely represented in this basis (Figures 5).

   This behaviour fits with the discussion of linear denoisers (Section 2.2). We heuristically demonstrated that the image data must be sparse in the optimal basis.

2. **Memorization to generalization.** For $f_\theta$ trained with increasing amounts of training data, the PSNR curves demonstrate a clear transition; with a small training dataset the PSNR is constant on test data, while it is highly accurate on training data. As the amount of train data increases, the test and train PSNR become essentially identical, suggesting that the denoiser begins to interpolate the training data (Figure 3).

   This interpolation of training data is rather strikingly demonstrated by distinct samplers converging to the same samples despite being trained on non-overlapping subsets of the training data (Figure 4). The authors partitions a training set $S = S_1 \sqcup S_2$ and fit distinct denoisers for sizes $|S_1| = |S_2| = N \in \{1, 10, \ldots, 10^5\}$. For small $N$, both models effectively memorize the training datapoints. As $N$ becomes large, the same noise initialization for distinct $f_\theta^1, f_\theta^2$ produce almost identical samples. As the generated samples from both models become similar, their similarity begins to deviate from their training set (bottom row of Figure 4).

   This corresponds to the discussion of 2.4, where we observe that the fitted score must learn some interpolation of the training data values. As the number of training data point grows, the score of the aforementioned mixture of Gaussians transitions to give an accurate represtation of the data.

3. $C^\alpha$ **optimality.** For synethetic $C^\alpha$ image datasets, and denoisers trained at fixed $\sigma^2$, the learned bases of the denoiser are almost exactly optimal. This is demonstrated by achieved the theoretical lower bound in the PSNR decay (Figure 6), which was derived in Section 2. Similarly to the optimal $C^\alpha$ basis constructel by Peyré and Mallat [7], the optimal $C^\alpha$ basis represented in the diffusion model is adapted to the edge contours. Remarkably, these features are implicitly represented in a UNet, without careful construction of a basis for the problem.

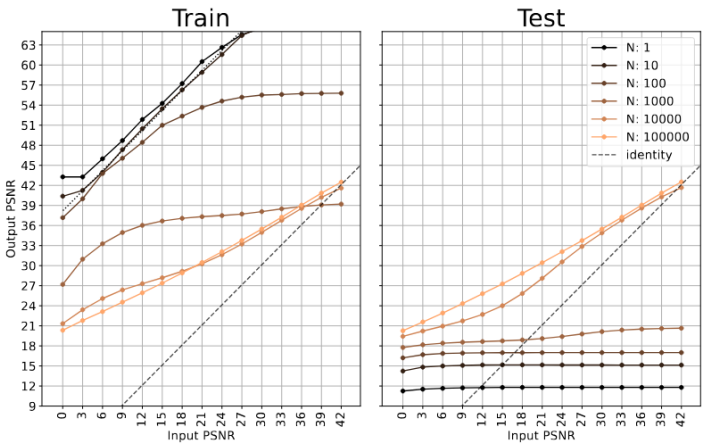



Figure 3: (from Kadkhodaie et al. [4]) Transition from memorization to generalization for a UNet denoiser trained on face images. Each curve shows the denoising error (output PSNR, ten times $\log_{10}$ ratio of squared dynamic range to MSE) as a function of noise level (input PSNR), for a training set of sizes $N \in \{1, \ldots, 10^5\}$. The increase in test performance on small noise levels at $N = 1000$ is indicative of the transition phase from memorization to generalization. At N = $10^5$, test and train PSNR are essentially identical, and the model has moved to the interpolation regime.

## 3.3 Experimental details

### Training

A standard UNet, of the type described in Section 2.5 in the majority of the experiments. Training of $f_\theta$ occurs by minimizing the MSE objective (Equation 2) where $\sigma^2 \sim U[0, 1]$ are uniformly drawn from $[0, 1]$ to match image pixel intensities. Figures presented in this project feature models trained on on synthetic and CelebA datasets [5].

### Sampling

Sampling from denoisers follows the algorithm of Kadkhodaie and Simoncelli [3], and is described below. For initial noise levels $(\sigma_0, \sigma_\infty)$, this method uses two additional hyperparameters $h \in [0, 1]$ and $\beta \in (0, 1]$, which control the step size and injected noise respectively (to accelerate the Langevin style sampling presented in Section 1). The authors set $h = 0.01$, $\beta = 0.1$, $\sigma_0 = 1$, and $\sigma_\infty = 0.05$.

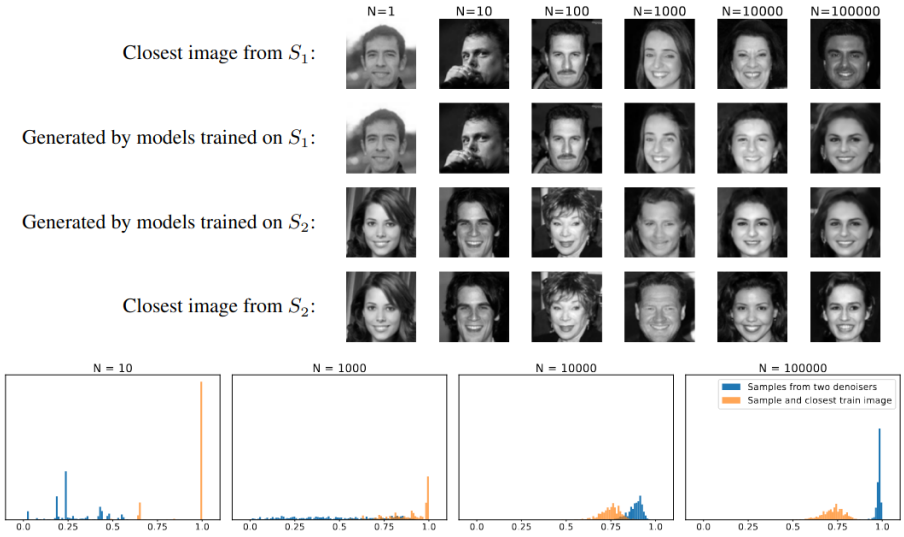



Figure 4: (from Kadkhodaie et al. [4]) Denoisers are trained on non-overlapping subsets $S_1$ and $S_2$ of a face dataset of size $N \in \{1, \ldots, 10^5\}$. After training, samples are compared in their similarity with the training data, visual (top row) and with cosine similarity across the training data (bottom row). The networks memorize for small $N$, and a distinct transition at $N \approx 1000$ where the generated samples are roughly interpolated. For large $N = 10^5$, the models generate almost identical samples from the same noise intialization. The cosine similarities betwen samples generated by the models becoms greater than with their respective training data.

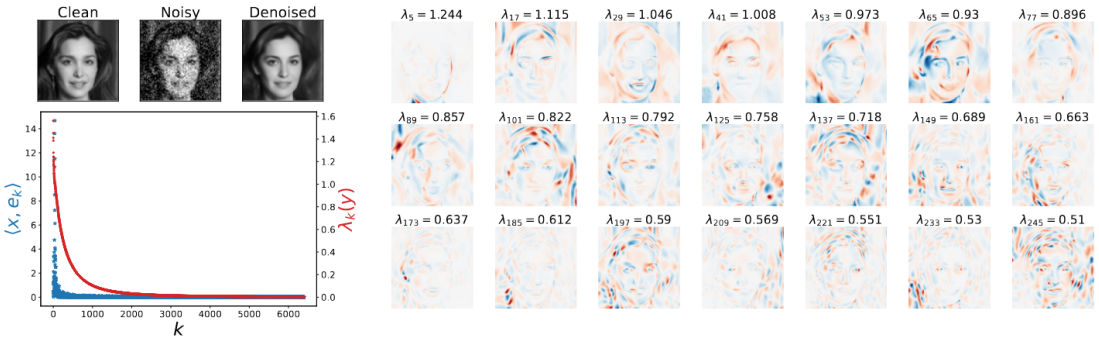



Figure 5: (from Kadkhodaie et al. [4]) Eigenvectors $\{e_k(y)\}$ of $Df_\theta(y)$ of a denoiser trained on $10^5$ face images, evaluated on a noisy test image. On the left, the decay of the eigenvalues $\lambda_k(y)$ is rapid, and and corresponding coefficients $\langle x, e_k(y) \rangle$ suggest that the clean image is sparse in $\{e_k(y)\}$. The adaptive basis vectors contain oscillating patterns, adapted to lie along the contours and within smooth regions of the image, whose frequency increases as $\lambda_k(y)$ decreases.

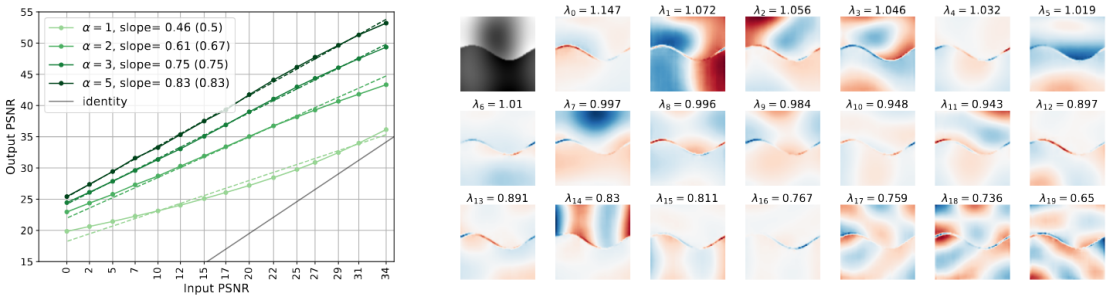



Figure 6: (from Kadkhodaie et al. [4]) UNet denoisers trained on $10^5$ $C^\alpha$ images achieve near-optimal PSNR slope decay. For varying $\alpha \in \{1, 2, 3, 5\}$, the slopes follow the theoretical lower bound $\frac{\alpha}{\alpha+1}$. On the right, the original $C^\alpha$ image with $(\alpha = 4)$ and the top eigenvectors of $Df_\theta(y)$. The vectors capture the contour of the images, and harmonic structure along the regular region, with oscillating patterns increasing for lower $\lambda_k(y)$.

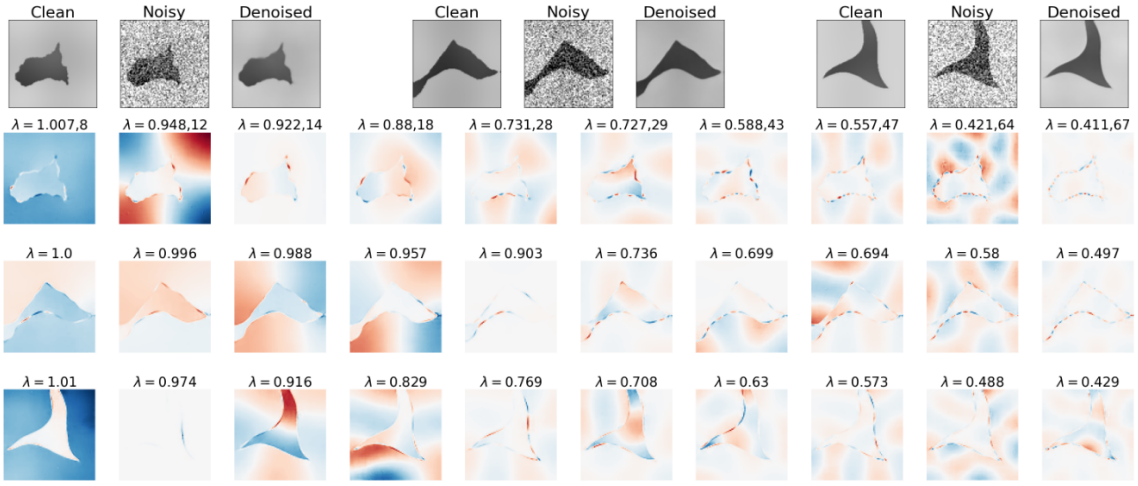



Figure 7: (from Kadkhodaie et al. [4]) Learned bases $\{e_k(y_i)\}$ basis shown for three test images $y_1, \ldots, y_3$ that are $C^\alpha$ geometrically regular. The regularity of the contours decreases from left to right, while the regularity of the background is fixed. The model adapts to edge and interior features of the data. As $\lambda_k(y)$ increases, the more regular contours show increased oscillating patterns. For the irregular (left) contours, different oscillatiing patterns are obesrved in the background (top row).

# 4 Consistency Models

Diffusion and score-based generative models use an iterative approach to generate data through a sampling process. In practice, the iterative approach often requires many steps at several noise levels for convergence (Section 1), also reflected in Algorithm 1 [3]. Song et al. [10] suggest using a single-step denoiser, which is trained to solve the ODE satisfied by the optimal score.

**Algorithm 1** Sampling via ascent of the log-likelihood gradient from a denoiser residual

**Require:** denoiser $f$, step size $h$, stochasticity from injected noise $\beta$, initial noise level $\sigma_0$, final noise level $\sigma_\infty$, distribution mean $m$

$t = 0$
Draw $x_0 \sim \mathcal{N}(m, \sigma_0^2 \mathrm{Id})$
**while** $\sigma_t \geq \sigma_\infty$ **do**
    $t \leftarrow t + 1$
    $s_t \leftarrow f(x_{t-1}) - x_{t-1}$ ▷ Compute the score from the denoiser residual
    $\sigma_t^2 \leftarrow \|s_t\|^2/d$ ▷ Compute the current noise level for stopping criterion
    $\gamma_t^2 = \left((1-\beta h)^2 - (1-h)^2\right)\sigma_t^2$
    Draw $z_t \sim \mathcal{N}(0, I)$
    $x_t \leftarrow x_{t-1} + h s_t + \gamma_t z_t$ ▷ Perform a partial denoiser step and add noise
**end while**
**return** $x_t$

## 4.1 Background

In the forward process, diffusion models slowly add noise to the initial data density $p_0(x)$, in which samples $(x_t)_{0 \leq t \leq T}$ evolve by

$$dx_t = \mu(x_t, t) + \sigma(t) dw_t$$

where $w_t$ is standard Brownian motion, $\sigma(t)$ is a time dependent noise level, and $\mu(x_t, t)$ represents the drift of these trajectories. Remarkably, the trajectories at time $t$ are distributed according to $x_t \sim p_t(x)$ and satisfy the **probability flow ODE** [12]

$$dx_t = (\mu(x_t, t) - \frac{1}{2}\sigma(t)^2 \nabla \log p_t(x_t))\, dt$$

Setting $\mu(x_t, t)$, $\sigma(t) = \sqrt{2t}$, and training a network to learn the score function $s_\theta(x_t) \approx \nabla \log p_t(x_t)$, the above can empirically be set rearranged to the ODE

$$\frac{dx_t}{dt} = -t s_\theta(x_t)$$

Sampling a particular $x_T \sim N(0, \sigma^2 I)$ and solving in backwards time therefore generates a sample $x_0$. In theory, for various time points the same trajectory $x_t, x_{t'}$, the solver theoretically arrives at similar values $x_0$ independent of $x_t, x_{t'}$. To encapsulate this process, we denote

$$x_0 = f_\theta(x_t, t)$$

where $f_\theta$ is the probability flow ODE solver. The above may be interpreted as a type of non-linear denoiser with additional time conditioning. In practice, $f_\theta$ is implemented by a UNet, matching Section 2.5 and the methods of Section 3. In the above, the conditioning value of $t$ depends nonlinearly on the noise level $\sigma$, which we will denote $t(\sigma)$ in the following sections.

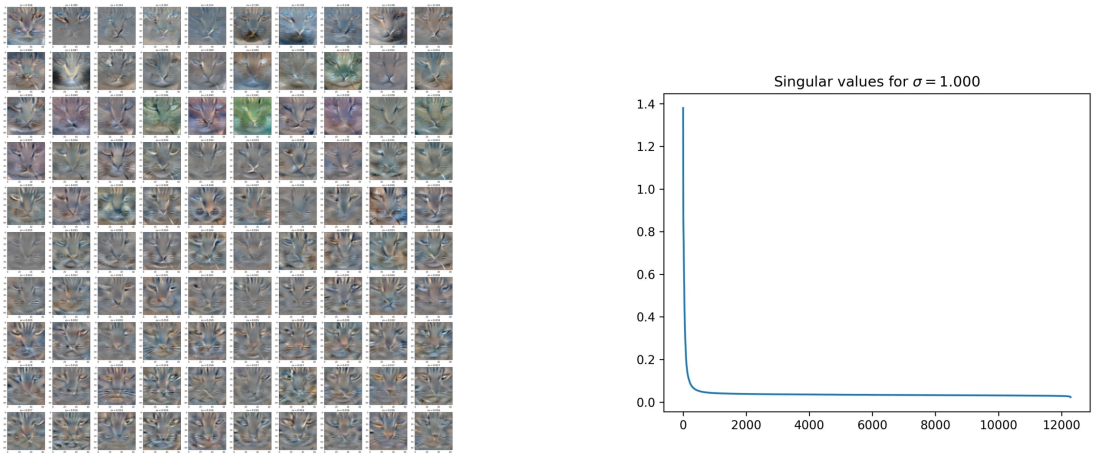
## 4.2 Experiments

We perform several small experiments to test the similarity to the results of Kadkhodaie et al. [4]. The models differ significantly in the method of training, and consistency models feature full non-linearities

throughout the model depth. Consistency models additionally condition on noise level, which is a significant difference from the previous results.

### SVD of Denoiser Jacobian

We test whether similar data-dependent representations occur in the singular vectors of the Jacobian $D_x f_\theta(x_t, t(\sigma))$ for a consistency model $f_\theta$.




(a) (b)

Figure 8: For $\sigma = 1$ the Jacobian $L = Df_\theta(y, t(\sigma))$ for noisy image $y$ is decomposed into its SVD $J = U\sigma V^T$. (a) The first 100 singular vectors, ie. columns of $U$ the singular vectors show similar oscillating structure in the later singular vectors, following the observations of Section 3. (b) The singular values (the diagonal of $\Sigma$). The singular vectors show a sharp decay rate in the dimension, demonstrating low-dimensional structure in the representations of the denoiser, in analogy with Section 3.

### Basis Conditioning

The consistency models are single-step denoisers and we may view them as directly mapping noise $z \sim N(0, \sigma^2 I)$ to the image manifold $x = f(z, t(\sigma))$. A natural question is whether a fixed noise realization $y^* = x + z^*$ may be perturbed by some basis vector $\phi$ such that

$$f_\theta(x + z^* + \phi, t(\sigma))$$

provides a slightly altered image. We consider the sinusoidal perturbation

$$\phi_{n,m} = \sin\left(\frac{2\pi n}{d}x\right)\sin\left(\frac{2\pi n}{d}y\right)$$

We apply $f$ the input renormalized to the original pixel intensity (Figure 9).

$$y^* = \frac{\|z^*\|}{\|z + \phi_{n,m}\|}(z + \phi_{n,m})$$

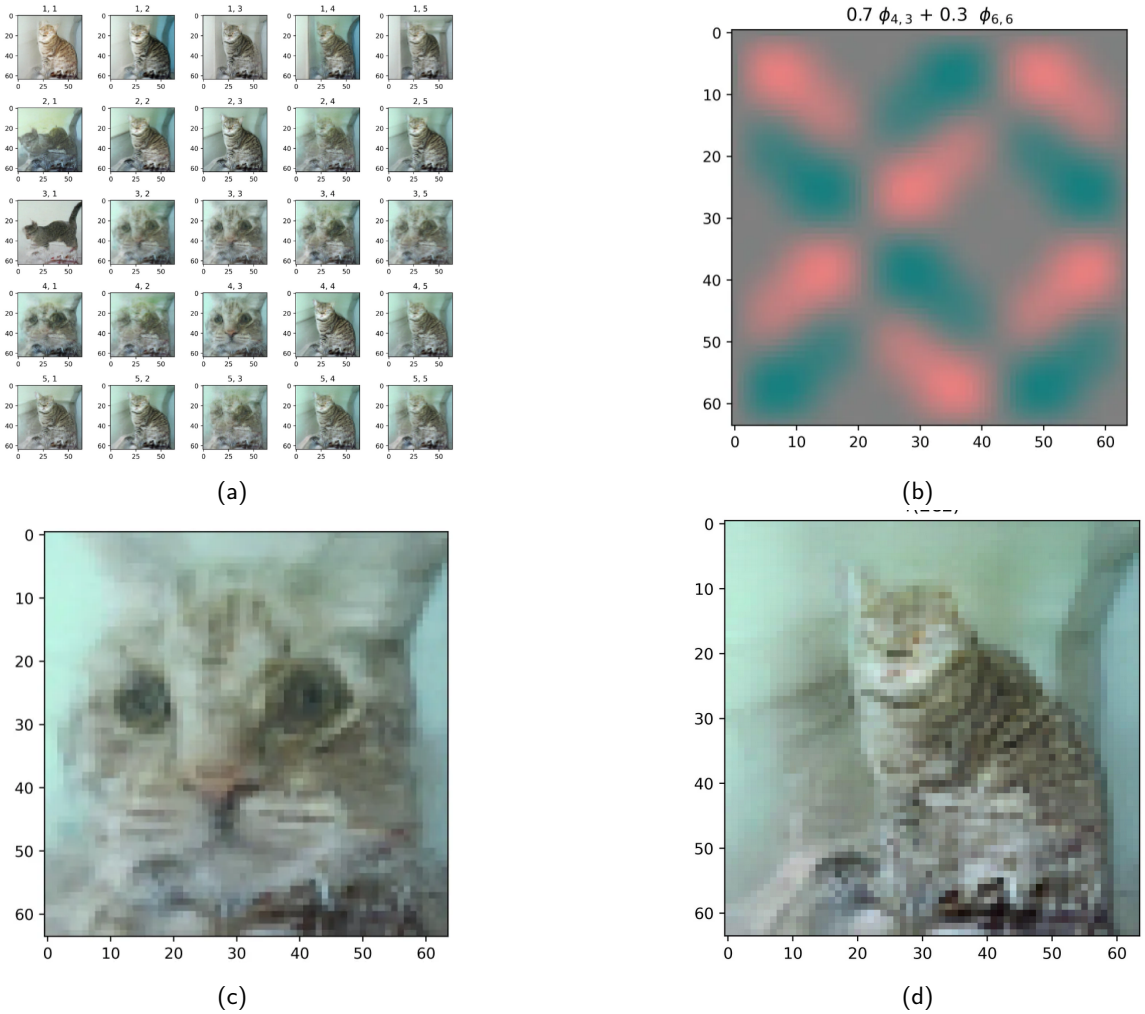



Figure 9: We fix $z^* \sim N(0, \sigma^2 I)$ where $\sigma^2 = 28$ (it is later renormalized by $f_\theta$.) (a) For $1 \leq n, m \leq 5$ we construct $y^*$ with $\phi_{n,m}$ as we previously described. This perturbation results in slight changes in the output. In the above, $\phi_{4,3}$ perturbs sufficiently perturbs the input to create a realistic yet significantly altered output. (b) An example of a linear combination of perturbations, $\phi = 0.7\phi_{4,3} + 0.3\phi_{6,6}$ which is designed perturb the input towards the desired output of (c). (c) The output when perturbing $z^*$ by the new $\phi$. (d) The output when performing no perturbations, ie. $f(z^*, t(\sigma))$.

We find that by slightly modifying the structure of the noise, we are able to modify the output of $f_\theta$. This suggests that $f_\theta$ *interpolates between images* in a steerable way; the geometry of the noise structure determines the exact features present in the output images. This motivates questions about which exact features of the noise influence generation in single-step consistency models, the underlying data-dependent representations of $f_\theta$, and their relationship to generalization.

# 5 Conclusion

In this report, we investigated the relationship between classical ideas in denoising and the properties of modern score-based models, as presented by Kadkhodaie and Simoncelli [3]. The work surveys how UNet architectures learn geometry-adaptive representations, and how these structures coincide with the features present in natural images. The main focus on the work is the transition from memorization to generalization in sampling, which is a central empirical question in score-based generative models. In an appropriate sense, the representations of these samplers are PSNR optimal, which is a remarkable fact. Motivated by these ideas, we test similar insights in consistency models, showing that feature adaptivity and low rank structure also emerge. We experiment with noise perturbations and reveal that these models learn steerable interpolations between images. This opens further questions, in particular an exact formulation of the features that UNets attend to, and how they impact the output. These ideas are important for interpretable diffusion and for precise mathematical formulations of generative models, which we leave as motivation for future work.

# References


[1] R. Basri and D.W. Jacobs. "Lambertian reflectance and linear subspaces". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.2 (2003), pp. 218–233. DOI: 10.1109/TPAMI.2003.1177153.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.

[3] Zahra Kadkhodaie and Eero P. Simoncelli. *Solving Linear Inverse Problems Using the Prior Implicit in a Denoiser*. 2021. arXiv: 2007.13640 [cs.CV]. URL: https://arxiv.org/abs/2007.13640.

[4] Zahra Kadkhodaie et al. *Generalization in diffusion models arises from geometry-adaptive harmonic representations*. 2024. arXiv: 2310.02557 [cs.CV]. URL: https://arxiv.org/abs/2310.02557.

[5] Ziwei Liu et al. "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. 2015.

[6] Gabriel Peyré. *Mathematical Foundations - Mathematical Tours of Data Sciences — mathematical-tours.github.io*. https://mathematical-tours.github.io/book/. 2024.

[7] Gabriel Peyré and Stéphane Mallat. "Orthogonal bandelet bases for geometric images approximation". In: *Communications on Pure and Applied Mathematics* 61.9 (2008), pp. 1173–1212. DOI: https://doi.org/10.1002/cpa.20242. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20242. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20242.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[9] Yang Song and Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. 2020. arXiv: 1907.05600 [cs.LG]. URL: https://arxiv.org/abs/1907.05600.

[10] Yang Song et al. *Consistency Models*. 2023. arXiv: 2303.01469 [cs.LG]. URL: https://arxiv.org/abs/2303.01469.

[11] Yang Song et al. *Maximum Likelihood Training of Score-Based Diffusion Models*. 2021. arXiv: 2101.09258 [stat.ML]. URL: https://arxiv.org/abs/2101.09258.

[12] Yang Song et al. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456 [cs.LG]. URL: https://arxiv.org/abs/2011.13456.

[13] D. J. Tolhurst, Y. Tadmor, and Tang Chao. "Amplitude spectra of natural images". In: *Ophthalmic and Physiological Optics* 12.2 (1992), pp. 229–232. DOI: https://doi.org/10.1111/j.1475-1313.1992.tb00296.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-1313.1992.tb00296.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-1313.1992.tb00296.x.

[14] A. Torralba and A. Oliva. "Statistics of natura image categories". In: *Network: Computation in Neural Systems* 14.3 (2003). PMID: 12938764, pp. 391–412. DOI: 10.1088/0954-898X_14_3_302. eprint: https://doi.org/10.1088/0954-898X_14_3_302. URL: https://doi.org/10.1088/0954-898X_14_3_302.