

ON THE ANISOTROPY OF SCORE-BASED GENERATIVE MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

We investigate the role of network architecture in shaping the inductive biases of modern score-based generative models. To this end, we introduce the *Score Anisotropy Directions* (SADs), architecture-dependent directions that reveal how different networks preferentially capture data structure. Our analysis shows that SADs form adaptive bases aligned with the architecture’s output geometry, providing a principled way to predict generalization ability in score models prior to training. Through both synthetic data and standard image benchmarks, we demonstrate that SADs reliably capture fine-grained model behavior and correlate with downstream performance, as measured by Wasserstein metrics. Our work offers a new lens for explaining and predicting directional biases of generative models.¹

1 INTRODUCTION

Neural networks generalize through inductive biases, i.e., biases that guide learning beyond training data (Goyal & Bengio, 2020; Wilson & Izmailov, 2020). For discriminative tasks, they are partially characterized through the *Neural Anisotropy Directions* (NADs), which reveal the architecture’s directional preferences in the input space (Ortiz-Jimenez et al., 2020). However, generative modeling lacks a cohesive theory that explains how architectural geometry interacts with data manifolds (Kadkhodaie et al., 2024; An et al., 2025). In this work, we present a unified approach to explaining and interpreting inductive biases of score-based generative models by examining anisotropy in the output space, where networks exhibit preferential learning along certain directions. As motivated in Figure 1, we posit that generalization ability is largely characterized by the alignment of the data with the architecture’s “geometry” at initialization. Our contribution is making this notion of “geometry” precise and decomposing the output space in terms of the anisotropy directions induced by it, i.e., the *Score Anisotropy Directions* (SADs) that we introduce below:

Definition 1 (Score Anisotropy Directions). The *Score Anisotropy Directions* (SADs) of an architecture are the ordered set of orthonormal vectors of the output space, $\{\mathbf{u}_i\}_{i=1}^D$, that are ranked in terms of the preference (e.g., as measured via Wasserstein metrics) of the network to generate data along those particular directions via score-based generative modeling.

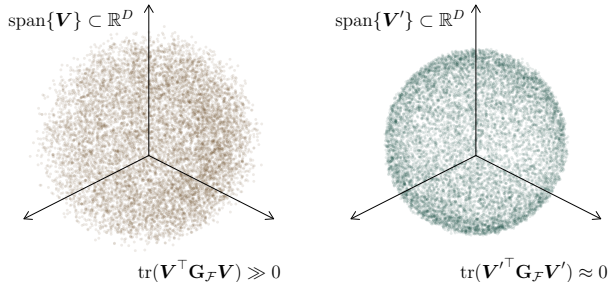


Figure 1: Sphere modeling in subspaces of \mathbb{R}^D ($D = 256$) via DiT (Peebles & Xie, 2023). The only difference is the choice of subspace: the left “sphere” lies in a subspace aligned with the network’s geometry, \mathbf{G}_F , while the right is in a non-aligned subspace. Despite identical setups, their quality differs consistently across repeated trials, suggesting that alignment with architectural geometry controls generalization. We formalize these ideas in Sections 3.2 and 3.3.

¹Code to reproduce our experiments is included in the supplementary material.

2 BACKGROUND

Here we give a brief overview of score-based generative models and provide some context on existing work that examines inductive biases of deep neural networks. We defer an extended discussion and comparison with related work to Section 4.1.

2.1 SCORE-BASED GENERATIVE MODELS

At the heart of score-based modeling is the *score function*, i.e., the gradient of the log-density of the underlying data distribution p . Given an estimate of the score, one can, in theory, sample from p via a gradient ascent procedure. Starting with an arbitrary prior, $\mathbf{x}_0 \sim \pi$, we have the following Langevin dynamics iterations:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\eta}{2} \overbrace{\nabla_{\mathbf{x}_k} \log p(\mathbf{x}_k)}^{\text{“score”}} + \sqrt{\eta} \mathbf{z}_k, \quad k = 0, 1, \dots, K \quad (1)$$

where $\mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian and $\eta > 0$. If $\eta \rightarrow 0$ and $K \rightarrow \infty$, under certain technical conditions, the iterates in Equation 1 converge to a sample from p (Welling & Teh, 2011). However, a practical limitation of the above setup is that estimated scores are inaccurate in low-density regions (e.g., in early iterations where learning is intractable). Moreover, score functions may be undefined in the case of data residing in low-dimensional manifolds. That is, the overall approach breaks down under the commonly adopted manifold hypothesis (Song & Ermon, 2019).

The research community has therefore largely moved on to Denoising Score Matching (DSM) (Vincent, 2011) and annealed Langevin dynamics, i.e., diffusion models, which are also the main focus of this work.² Specifically, consider noise scales, $\sigma \in [\sigma_{\min}, \sigma_{\max}]$, and associated densities, $p_\sigma = p * \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, where $*$ represents convolution. Here, σ_{\min} is small enough such that $p_{\sigma_{\min}} \approx p$ and σ_{\max} is large enough so we can write $p_{\sigma_{\max}} \approx \mathcal{N}(\mathbf{0}, \sigma_{\max}^2 \mathbf{I})$. With estimates of scores of the perturbed distributions, $\nabla_{\mathbf{x}_\sigma} \log p_\sigma(\mathbf{x}_\sigma)$, one samples from p by decaying σ from σ_{\max} to σ_{\min} (Ho et al., 2020; Song et al., 2021; Miyato et al., 2025; Song et al., 2025). This way, accurate modeling along sampling trajectories is tractable and the noise ensures support over the entirety of the ambient space, overcoming the limitations of naive Langevin dynamics. In particular, the scores are equivalent to minimum mean squared error Gaussian denoisers (Efron, 2011). That is, for neural networks, $\mathcal{F}_\theta : \mathbb{R}^D \times \mathbb{R} \rightarrow \mathbb{R}^D$, parameterized by θ , one can approximate the scores via the following DSM optimization:

$$\min_{\theta} \mathbb{E}_{\mathbf{x} \sim p, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \sigma} \left[\hat{\mathcal{J}}_{\text{DSM}} := \left\| \mathcal{F}_\theta(\mathbf{x} + \sigma \epsilon, \sigma) + \frac{\epsilon}{\sigma} \right\|_2^2 \right]. \quad (2)$$

2.2 INDUCTIVE BIASES IN DEEP LEARNING

There is a vast literature on understanding the inductive biases of deep neural networks (Goyal & Bengio, 2020; Wilson & Izmailov, 2020; Radhakrishnan et al., 2024). Of particular interest is the work of Ortiz-Jimenez et al. (2020), who identify directional biases in classifiers, i.e., the *Neural Anisotropy Directions* (NADs). More recently, Movahedi et al. (2025) have extended the NAD framework by formalizing the concept of input-space architectural geometry and exploring its evolution during training. Specifically, they propose the *geometric invariance hypothesis*, which posits that NADs persist through training.

To our knowledge, contrary to the discriminative case, there is no unified theory that explains the preferred modeling directions of diffusion. For example, a recent study by Kadkhodaie et al. (2024) argues that convolutional diffusion models are biased towards *Geometry-Adaptive Harmonic Bases* (GAHBs). However, they acknowledge that a mathematically precise definition of such bases remains an open question. More fundamentally, follow-up work by An et al. (2025) suggests that the theory of GAHBs does not extend to transformer-based diffusion models.

²Without loss of generality, we will standardize notation to the variance exploding formulation of Song & Ermon (2019).

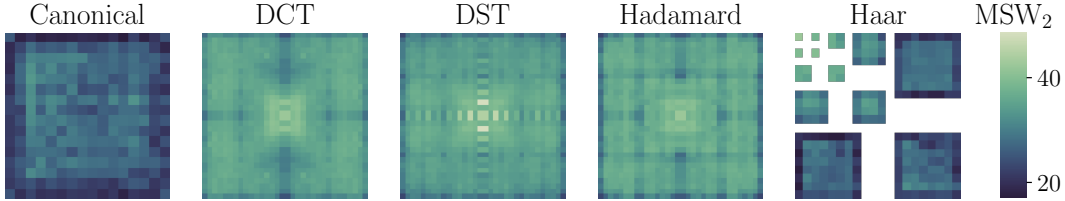


Figure 2: MSW_2 distance (computed over 10k test samples and 16384 projections) of iDDPM U-Net (Nichol & Dhariwal, 2021) architecture. Each pixel corresponds to a rank-one dataset of 16×16 images (with 10k training samples) that is aligned with a basis element of the canonical basis, DCT, DST, (ordered) Hadamard transform or Haar wavelet transform. That is, for a given location (canonical) or frequency / frequency / scale, channel and location (Haar), we visualize the performance on the corresponding dataset. For ease of visualization, in the case of DCT, DST and Hadamard, we center the zero frequency dataset and extend the images to the left and top regions while respecting the symmetries of the transforms. See Appendix A for details.

In contrast to existing works on inductive biases of diffusion (Kadkhodaie et al., 2024; An et al., 2025), we present a unified treatment of directional biases that accounts for network architecture. In particular, our key insight is recognizing that the NAD framework can be adapted and extended to score-based modeling by assuming an underlying data log-density that assigns high probability to “on-manifold” data and low probability otherwise. With this approach, our analysis amounts to understanding the anisotropy directions of such implicitly induced discriminative models.

3 DIRECTIONAL INDUCTIVE BIASES OF DIFFUSION MODELS

Notation Matrix entries are indexed as $(\cdot)_{(i,j)}$ for row i and column j . Unless otherwise specified, we write the i^{th} row of a matrix as $(\cdot)^{(i)}$. \otimes is the Kronecker product. $\delta_{(\cdot)}$ denotes the Kronecker delta or Dirac delta distribution, depending on the context. The term iid refers to independent and identically distributed random variables. Gaussian distributions with mean μ and covariance Σ are represented as $\mathcal{N}(\mu, \Sigma)$. $\mathcal{U}(S)$ is the uniform distribution over the set S . We abbreviate (Stochastic) Gradient Descent as (S)GD. Eigendecompositions of matrices are assumed to have the eigenvalues in descending order, so that the first eigenvector corresponds to the largest eigenvalue. We assume that all expectations appearing in our arguments are finite and well-defined.

Central to our exploration of directional inductive biases in diffusion is the following question:

Among equidimensional manifolds, which are preferred by diffusion-based generative modeling and how can such preferences be quantified?

To answer the above, we consider data manifolds aligned with a particular direction, $v \in \mathbb{S}^{D-1}$, and investigate generalization ability as a function of the direction. Concretely, we study distributions of the form $\mathcal{N}(\mathbf{0}, Dvv^\top)$ and are tasked with finding a suitable basis for \mathbb{R}^D from which we can draw v that reveal directional preferences. We include complete details of our experimental setup in Appendix A. In summary, for a given basis, we independently train diffusion models on datasets formed by each normalized element under identical settings. Performance is quantified via the (Max-)Sliced Wasserstein p -distance, $(M)SW_p$, between the distribution obtained by sampling from the trained model, μ , and the ground truth data, corresponding to $\nu = \mathcal{N}(\mathbf{0}, Dvv^\top)$. Notably, these are valid statistical distances and the slice distance, W_2 , is easily estimated via order statistics and Monte Carlo methods. Taking $p = 2$ and letting $F_{(\cdot)}^{-1}$, $(\cdot)_\#$ represent quantile functions, the push-forward respectively, we express the Wasserstein metrics theoretically as:

$$SW_2^2(\mu, \nu) = \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{D-1})} W_2^2(\theta_\#^\top \mu, \theta_\#^\top \nu), \quad MSW_2^2(\mu, \nu) = \sup_{\theta \in \mathbb{S}^{D-1}} W_2^2(\theta_\#^\top \mu, \theta_\#^\top \nu),$$

$$W_2^2(\theta_\#^\top \mu, \theta_\#^\top \nu) = \int_0^1 |F_{\theta_\#^\top \mu}^{-1}(q) - F_{\theta_\#^\top \nu}^{-1}(q)|^2 dq.$$

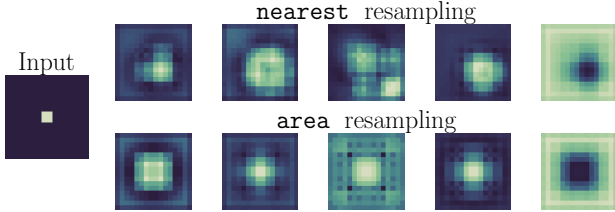


Figure 3: Responses of the iDDPM architecture (Nichol & Dhariwal, 2021) with a symmetrical initialization scheme. We show the default implementation, which uses nearest resampling layers, and a modified architecture that uses area resampling. We probe the models with a centered impulse input, shown on the left. Observe that the default resampling introduces asymmetry.

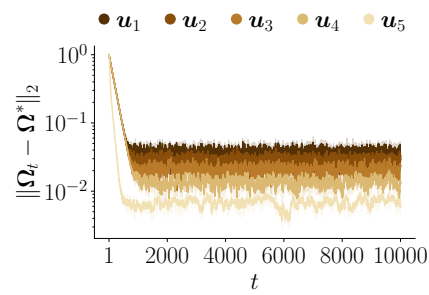


Figure 4: Setting of Theorem 1 ($\sigma = 1$) in \mathbb{R}^5 with SGD. We show error between $\Omega_t = \Phi \Theta_t$ and the optimal operator, Ω^* , defined in Lemma 1.

In Figure 2 we report our findings (averages over five runs) with the above-described approach for common bases found in signal processing literature. Interestingly, contrary to the conventional wisdom that neural networks better adapt to lower frequencies (Rahaman et al., 2019), the standard U-Net used in DDPMs (Nichol & Dhariwal, 2021) struggles with low-frequency data. This is evidenced by the center points of the DCT, DST, Hadamard images and the top-left corner of the Haar image. Also, comparing results of the canonical and Haar bases with the frequency / sequency transforms, we see that vectors localized in space are better modeled, especially around the borders.

3.1 ANISOTROPIC CONDITIONING OF THE OPTIMIZATION LANDSCAPE

Why would certain directions be preferred? While the learning process has a number of hyperparameters that can potentially induce asymmetry, here we specifically focus on analyzing the role of the architecture. Prior work on the inductive biases of discriminative models argues that preferences may emerge due to anisotropic loss of information or, more generally, conditioning of the optimization landscape (Ortiz-Jimenez et al., 2020). Indeed, such conditioning also naturally manifests in multiscale U-Nets, for example, in the form of asymmetric resampling layers. We verify this in Figure 3, where we observe that the default nearest interpolation leads to a clear directional bias. Similarly, one may hypothesize that the border effects observed in the canonical and Haar basis experiments of Figure 2 can be attributed to the standard zero padding strategy employed in the convolutional layers. Also, at the same time, one might intuitively expect that data aligned with such padding artifacts is poorly approximated by the network.

To more rigorously understand the effect of anisotropic conditioning, we now revisit the problem of learning rank-one distributions, introduced in the beginning of Section 3, from a theoretical perspective. Assuming a linear DSM architecture, where anisotropy is explicitly modeled via a fixed, linear transformation with decaying eigenvalues at the output, we make the following observation:

Theorem 1 (DSM under anisotropy, proof in Appendix D.2). *Consider the DSM problem with data drawn from $\mathcal{N}(\mathbf{0}, \mathbf{v}\mathbf{v}^\top)$ for a fixed noise level $\sigma > 0$ and $\|\mathbf{v}\|_2 = 1$. Let $\mathcal{F} : \mathbb{R}^D \rightarrow \mathbb{R}^D$ be linear networks expressed as $\Omega(\cdot)$, where $\Omega = \Phi \Theta$ with Φ fixed. Denote the sorted eigenvalues of $\Phi \Phi^\top$ as $\{\lambda_i\}_{i=1}^D$ with $\lambda_{D-1} > \lambda_D > 0$ and the corresponding normalized eigenvectors as $\{\mathbf{u}_i\}_{i=1}^D$. Assume a (S)GD procedure on initially zero-mean Θ , where the score is approximated by \mathcal{F} . After t steps, and with sufficiently small learning rate $\eta > 0$, the mean error, $\mathbb{E}[\Omega - \Omega_*]$, to the optimal solution, Ω_* , for $\mathbf{v} = \mathbf{u}_i$ decays as $\mathcal{O}[(1 - 2\eta\rho_i)^t]$ with $\rho_1 = \rho_D \forall i < D$. Moreover, near optimality, the SGD steps with respect to Θ for $\mathbf{v} = \mathbf{u}_i$ have covariance $\text{Cov}[\nabla_{\Theta} \hat{\mathcal{J}}_{\text{DSM}}] \propto \lambda_i$.*

Curiously, the effect of anisotropy is limited in deterministic GD, i.e., if $\lambda_D = \lambda_{D-1}$ then it completely disappears and the convergence rate is the same for all eigenvectors of $\Phi \Phi^\top$. However, our derivations suggest that anisotropy is greatly amplified in SGD. To confirm this, we design a small-scale experiment in \mathbb{R}^5 , with the results shown in Figure 4. Specifically, for small t , the optimization dynamics are well-predicted by the deterministic GD analysis of Theorem 1 since, intuitively, the iterates are far from the optimum and the deterministic drift, i.e., the gradient, dominates the stochastic fluctuation. Consequently, all vectors except \mathbf{u}_5 yield similar results. For large t , however, where

the gradient norm is sufficiently small, the noise dominates and it appears we have convergence to a stationary distribution. In such stochastic regimes, the error scales with the magnitude of the stochastic gradient covariance (Mandt et al., 2017), which we show to be $\propto \lambda_i$ near optimality.

The takeaway from this discussion is that the best performance is achieved when the data is *not* aligned with the “geometry” that is induced by the score network, but instead lives in the subspace defined by its *smallest* eigenvalues. Next, we will extend and formalize this intuition in a more general setting, for potentially non-linear architectures.

3.2 IDENTIFYING PREFERRED DIRECTIONS FOR GENERATIVE MODELING

Although the experiments of Figure 2 are insightful, it is unclear whether standard bases can reliably describe the biases of diffusion models given that they are completely decoupled from the underlying architecture. As the choice of bases was arbitrary in these initial experiments, we argue that, in general, we cannot hope to uncover the intricacies of directional biases in this manner. Moreover, note that each considered basis amounts to training hundreds of diffusion models, which becomes impractical even for relatively small-dimensional data. We now present a more principled approach to these experiments, by establishing theoretically motivated and architecture-dependent bases that predict directional biases prior to any training.

Concretely, for each noise level, σ , and assuming a (conservative and normalizable) parameterization via a neural network family, \mathcal{F} , we intuitively treat a realization, $\mathcal{F}_\theta : \mathbb{R}^D \times \mathbb{R} \rightarrow \mathbb{R}$, as implicitly defining a log-density function, $\mathbf{x}_\sigma \mapsto \log p_{\theta,\sigma}(\mathbf{x}_\sigma)$, that assigns high probability to \mathbf{x}_σ on the “data manifold” and low probability to “off-manifold” data, i.e., we can write $\mathcal{F}_\theta(\mathbf{x}_\sigma, \sigma) \approx \nabla_{\mathbf{x}_\sigma} \log p_{\theta,\sigma}(\mathbf{x}_\sigma)$. Here, we use the term “manifold” loosely, meaning regions of \mathbb{R}^D that are statistically likely to be sampled by score-based modeling via \mathcal{F}_θ , i.e., the spanning set of the top SADs that we aim to uncover. Now, intuitively, if one were to fix \mathbf{x}_σ and consider a small perturbation along some direction, \mathbf{v} , an abrupt change in the log-density indicates falling off or entering the “manifold”, that is, $\mathbf{v} = \mathbf{v}_\perp$ is perpendicular to the “manifold”. Similarly, if $\mathbf{v} = \mathbf{v}_\parallel$ is parallel to the “manifold”, then we expect minimal changes in the log-density, i.e., we are traveling along a contour line. We illustrate our argument in Figure 5. Now, taking the limit as the perturbation magnitude, τ , tends to zero, we summarize these dynamics via directional derivatives, i.e., we have $|\log p_{\theta,\sigma}(\mathbf{x}_\sigma + \tau \mathbf{v}) - \log p_{\theta,\sigma}(\mathbf{x}_\sigma)| \propto |\mathbf{v}^\top \nabla_{\mathbf{x}_\sigma} \log p_{\theta,\sigma}(\mathbf{x}_\sigma)|$. With this simplification, a straightforward application of Markov’s inequality bounds the *a priori* probability of crossing the “manifold” by moving along \mathbf{v} , suggesting that directions attaining the minimum upper bound are inherently easier to model. That is, such directions are aligned with the prior density induced by \mathcal{F}_θ . Specifically, for the family of networks, \mathcal{F} , parameterized by $\theta \sim \Theta$ over noise levels, σ , and probing with $(\mathbf{x}_\sigma, \sigma) \sim \mathcal{P}$, we write the bound:

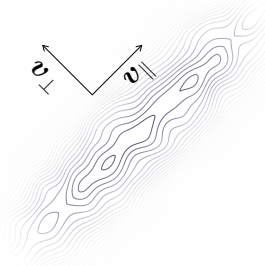


Figure 5: Visualization of our argument for uncovering anisotropy directions in \mathbb{R}^2 . We show contours of a hypothetical landscape, $\log p_{\theta,\sigma}(\mathbf{x}_\sigma)$, where \mathbf{v}_\parallel is parallel to the induced “manifold” and \mathbf{v}_\perp is orthogonal.

$$\mathbb{P}(|\mathbf{v}^\top \nabla_{\mathbf{x}_\sigma} \log p_{\theta,\sigma}(\mathbf{x}_\sigma)| \geq \eta) \leq \frac{\overbrace{\mathbf{v}^\top \left[\mathbb{E}_{(\mathbf{x}_\sigma, \sigma) \sim \mathcal{P}, \theta \sim \Theta} \mathcal{F}_\theta(\mathbf{x}_\sigma, \sigma) \mathcal{F}_\theta(\mathbf{x}_\sigma, \sigma)^\top \right] \mathbf{v}}^{\text{“geometry”}}}{\eta^2}. \quad (4)$$

We note that by applying this bound on the setting of Theorem 1 with θ iid and for any \mathcal{P} , the quantity under the overbrace, namely the *average geometry*, recovers the conditioning matrix, $\Phi \Phi^\top$, whose eigendecomposition defined the SADs in the case of linear networks. Moreover, the Markov bound correctly predicts that the small-eigenvalue vectors are easier to model compared to larger-eigenvalue eigenvectors. We therefore expect that this quantity also captures directional preferences of more general architectures. As the average geometry is central to our study of directional biases, let us formally introduce it below.

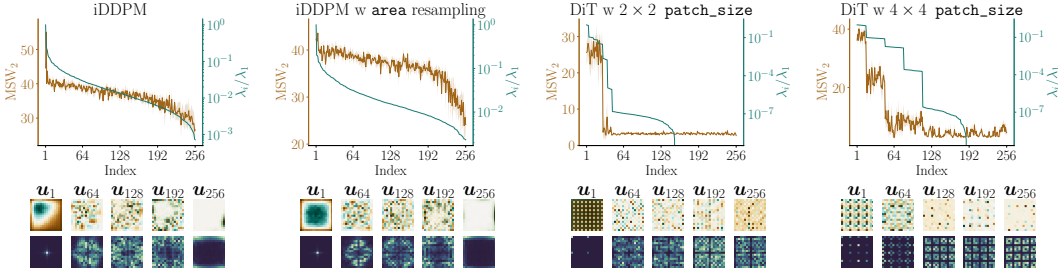


Figure 6: Test MSW_2 distance for different architectures on datasets aligned with the eigenvectors of their geometry at initialization, probing with $\mathcal{P} = \delta_0 \times \mathcal{U}(\{\sigma_{\min}, \dots, \sigma_{\max}\})$. We report the mean \pm the standard error over five independent runs. Corresponding normalized eigenvalues are on the right axes. The eigenvectors, with their energy in the DFT (zero frequency is centered), are shown below the plots (first, last row respectively). The experimental setup is identical to the one described in the beginning of Section 3 and Figure 2. We refer the reader to Appendix A for further details.

Definition 2 (Average Geometry). Let \mathcal{F} be a family of neural networks parameterized by $\theta \sim \Theta$ (e.g., at initialization) such that $\mathbb{R}^D \times \mathbb{R} \rightarrow \mathbb{R}^D$. In the context of diffusion-based generative modeling, we define the average geometry of \mathcal{F} , induced by a probing distribution, \mathcal{P} , and parameterized by Θ , as:

$$\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta) = \mathbb{E}_{(\mathbf{x}_{\sigma}, \sigma) \sim \mathcal{P}, \theta \sim \Theta} [\mathcal{F}_{\theta}(\mathbf{x}_{\sigma}, \sigma) \mathcal{F}_{\theta}(\mathbf{x}_{\sigma}, \sigma)^{\top}], \quad (5)$$

where we assume the networks are, roughly, aligned with some underlying data density, $p_{\theta, \sigma}(\mathbf{x}_{\sigma})$, via $\mathcal{F}_{\theta}(\mathbf{x}_{\sigma}, \sigma) \approx \nabla_{\mathbf{x}_{\sigma}} \log p_{\theta, \sigma}(\mathbf{x}_{\sigma})$. Note, the probing distribution, \mathcal{P} , is chosen independently and need not match this underlying density.

At this point it is important to note that, unlike in the linear networks investigated in Theorem 1, the notion of average geometry is, in general, local and adapts to the probing data distribution \mathcal{P} . To decouple the geometry from the data, one may consider an isotropic probe, e.g., $\mathcal{N}(\mathbf{0}, \sigma_{\mathcal{P}}^2 \mathbf{I})$. However, in practice, we find that the probe is a non-critical hyperparameter and not central to our analysis. For simplicity, we default to $\mathcal{P} = \delta_0 \times \mathcal{U}(\{\sigma_{\min}, \dots, \sigma_{\max}\})$, i.e., we fix $\mathbf{x}_{\sigma} = \mathbf{0}$ and draw σ uniformly from the noise levels of interest. We shall write the geometry under this default probe as $\mathbf{G}_{\mathcal{F}}$, where the parameters, θ , of \mathcal{F} are assumed to be drawn from the default initialization scheme for the architecture. Having formalized the concept of output-space geometry, we are now ready to state our main conjecture regarding the SADs, linking them to the average geometry at initialization:

Conjecture 1. Let \mathcal{F} be a family of networks with geometry $\mathbf{G}_{\mathcal{F}}$. We hypothesize that the eigenvectors of $\mathbf{G}_{\mathcal{F}}$, in *ascending* eigenvalue order, are the SADs. That is, we expect that data aligned with eigenvectors corresponding to small eigenvalues is better modeled compared to data that is aligned with large-eigenvalue eigenvectors.

In the remainder of the paper, we provide empirical evidence to justify Conjecture 1. First, we verify our claims on the rank-one datasets introduced in the beginning of Section 3 in Figure 6, where we draw the directions, $\mathbf{v} \in \mathbb{R}^D$, from the eigenvectors of $\mathbf{G}_{\mathcal{F}}$ and benchmark via Wasserstein metrics.

On the left in Figure 6, we focus on the iDDPM U-Net (Nichol & Dhariwal, 2021), which is representative of convolutional diffusion models. The experiments show a clear trend in support of the conjecture, where eigenvectors corresponding to small eigenvalues achieve the best performance and large-eigenvalue vectors have the worst performance. Interestingly, with reference to the visualizations below the plots, we also observe harmonic patterns in the eigenvectors, where large eigenvalues correspond to low-frequencies and small eigenvalues to high-frequencies. In this sense, our findings provide further evidence in support of the theory of GAHBs (Kadkhodaie et al., 2024), which posits that harmonic representations are fundamental in convolutional models.³ Crucially, beyond the harmonic structure that only loosely characterizes the inductive biases, our notion of geometry is flexible as it captures and adapts to architectural details. This is evident in the first eigenvectors, \mathbf{u}_1 , when we ablate the resampling method, also investigated previously in Figure 3.

³See Figure 9 in Appendix B for more visualizations that connect our work with Kadkhodaie et al. (2024).

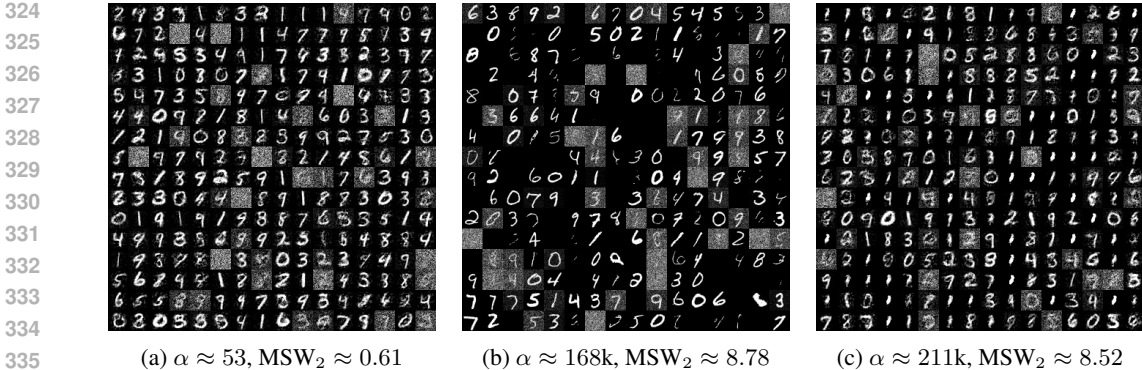


Figure 7: Uncurated samples of MNIST-trained iDDPMs. We vary the alignment of the data with the geometry, $\alpha := \mathbb{E}_{\mathbf{x} \sim p}[\mathbf{z}^\top \mathbf{G}_{\mathcal{F}} \mathbf{z}]$, by transforming the data via appropriate orthogonal matrices \mathbf{W} . On the left, we show the effect of minimizing α , where the model gives a reasonable approximation of the ground truth distribution. The middle shows the default alignment, i.e., we do not apply any transformation. It is evident that the data is not well-modeled in this case as a considerable fraction of samples do not contain a digit, as also quantified via the MSW_2 . The right shows samples from the model corresponding to maximizing α , where we see similar α and MSW_2 to the middle. We also observe signs of mode collapse, where the model is more likely to generate the digit “1”.

As further proof of the flexibility of our proposed average geometry, we now focus our analysis on the DiT architecture (Peebles & Xie, 2023), which is representative of transformer-based diffusion models. The results, shown on the right in Figure 6, are also in support of Conjecture 1. However, as also noted in (An et al., 2025), it is clear in this case that the transformer architecture does not induce harmonic representations and the theory of GAHBs does not apply. It is also interesting to observe that the DiT geometry exhibits considerable eigen-multiplicity, with larger patch sizes amplifying this, reflecting a looser structure and weaker inductive biases compared to convolutional networks. In fact, in Proposition 3 we show that the number of transformer tokens, T , is an upper bound on the number of distinct eigenvalues of $\mathbf{G}_{\mathcal{F}}$. More generally, we refer the reader to Appendix C for an analytical treatment of geometries of common architectures, noting that a surprising amount of structure can be inferred simply by inspecting the output layers.

3.3 SCORE ANISOTROPY DIRECTIONS IN THE WILD

Having identified the preferred modeling directions in our experiments on rank-one datasets, we now turn to more realistic data distributions that are encountered in practice. Based on our analysis in Section 3.2, we hypothesize that generalization is largely determined by the (mis)alignment of the data with the average geometry. For an arbitrary data distribution, p , we can extend our setup by defining the alignment with the network, α , as follows:

$$\alpha := \mathbb{E}_{\mathbf{x} \sim p}[\mathbf{z}^\top \mathbf{G}_{\mathcal{F}} \mathbf{z}], \quad \mathbf{z} = \mathbf{W}\mathbf{x}, \quad \mathbf{W}^\top \mathbf{W} = \mathbf{I}. \quad (6)$$

Note, in order to vary α in a way that preserves underlying structure, we have introduced the orthogonal matrix $\mathbf{W} \in \mathbb{R}^{D \times D}$, which models simple and lossless data transformations, effectively defining a linear autoencoder and a kind of “latent” diffusion model that operates on the transformed data. With this setup, Conjecture 1 amounts to the claim that the best performance is observed when α is minimized and the worst when α is maximized. The corresponding orthogonal matrices, \mathbf{W}_{\min} and \mathbf{W}_{\max} , that achieve such extreme (mis)alignment are given by Theorem 2:

Theorem 2 (Extreme alignment, proof in Appendix D.3). *Stacking the (descending by eigenvalue magnitude) eigenvectors of $\mathbf{G}_{\mathcal{F}}$ and $\mathbb{E}_{\mathbf{x} \sim p}[\mathbf{x}\mathbf{x}^\top]$ as columns in matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{D \times D}$ respectively, α is minimized for $\mathbf{W}_{\min} = \mathbf{U}\mathbf{R}\mathbf{V}^\top$ and maximized for $\mathbf{W}_{\max} = \mathbf{U}\mathbf{V}^\top$. Here, \mathbf{R} is the row-reversed identity matrix, i.e., it reverses the order of the columns in \mathbf{V} .*

Therefore, to test our hypothesis, for each \mathbf{W} , we will train such “latent” diffusion models under identical settings. To compare with the standard diffusion training, we also consider a baseline corresponding to $\mathbf{W} = \mathbf{I}$, i.e., the natural alignment of the data with the average geometry. In particular, we conduct experiments on the MNIST (LeCun et al., 2010) (28×28), CelebA-HQ (Karras et al., 2018) (56×56) and CIFAR-10 (Krizhevsky, 2009) (32×32) datasets, with implementation details included in Appendix A. Our findings, shown in Figures 7 and 8, agree with our central conjecture. For example, focusing on the MNIST samples obtained by our models, which are shown in Figure 7, we observe that large α results in artifacts on the distribution-level.⁴ Specifically for the default alignment, which is already significant, a considerable fraction of the samples do not contain any digit. Moreover, maximizing α leads to mode collapse, with the digit “1” being overrepresented. When we instead minimize α , substantial and consistent improvements are observed across all datasets, as quantified by the Wasserstein metrics in Figure 8. Interestingly, we find that, by default, the considered datasets are already well-aligned with $\mathbf{G}_{\mathcal{F}}$. Indeed, our analysis shows that explicitly maximizing α also yields very similar scores to the baseline $\mathbf{W} = \mathbf{I}$, which is consistent with our hypothesis that the optimization dynamics of diffusion models are tightly linked to the interaction of data and architecture via α . In particular, since both $\mathbf{W} = \mathbf{W}_{\max}$ and $\mathbf{W} = \mathbf{I}$ exhibit high alignment, their performance differences may not be statistically significant. The more meaningful comparison is with $\mathbf{W} = \mathbf{W}_{\min}$, where alignment is minimal and the Wasserstein metrics are also substantially lower in all cases.

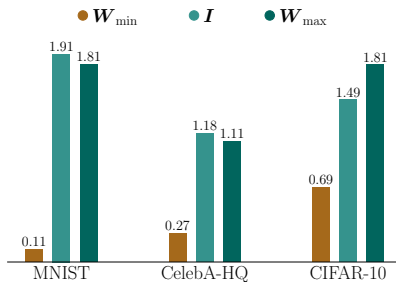


Figure 8: iDDPM average performance (SW_2) over five independent runs on standard image datasets as the alignment with the geometry, α , varies. We show the effect of minimizing, maximizing α as well as the default setting. These experiments correspond to matrices \mathbf{W}_{\min} , \mathbf{W}_{\max} and \mathbf{I} respectively.

4 DISCUSSION

We have presented and experimentally validated a theory for determining the directional inductive biases of diffusion models. Quite surprisingly, we find that, despite the highly non-linear nature of modern neural networks, these biases are well-described via fixed bases, that may be decoupled from the data and only depend on the architecture, making them useful for predicting generalization ability, as quantified via Wasserstein metrics. Specifically, we have shown that the directions defined by these bases impose strong priors during the training process and define the order by which features of the data are learned, i.e., we identify them as the SADs.

4.1 RELATED WORK

Kadkhodaie et al. (2024) observe that convolutional diffusion models are inductively biased toward GAHBs. Specifically, by assuming a homogeneous model (Mohan et al., 2020), they give a shrinkage-based interpretation of denoising with adaptive eigenbases given by the local Jacobian of the trained networks. A similar Jacobian-based analysis has also been explored in transformer-based diffusion (An et al., 2025). However, the observations of Kadkhodaie et al. (2024) are architecture-specific and largely empirical, therefore they cannot be used to predict preferred modeling directions in general. For example, transformer-based diffusion exhibits no obvious regularities in the eigenbases, such as harmonic structure, that could be systematically exploited in An et al. (2025). Despite using a fundamentally different approach, our theory of SADs appears to extend and is compatible with the findings of these works. In particular, we provide a more rigorous treatment of architectural geometry, decoupling it from the data and we demonstrate a method that can predict generalization ability *prior to any training*. We achieve this by decomposing and, crucially, *ordering the output space* via the SADs, which are defined explicitly. This ordering provides fine-grained information about inductive biases, rather than a loose characterization. Moreover, unlike prior Jacobian-based analyses, we make no assumptions regarding homogeneity, which may be suboptimal / not hold across domains or architectures, and our notion of geometry only requires forward passes through

⁴Note that, in general, we do not expect perceptual quality to correlate with the Wasserstein distances.

432 the networks. This makes our approach more broadly applicable and straightforward to implement.
433 As future work, we remark that it remains an open question whether our method has theoretical ties
434 with the analysis initially proposed by Kadkhodaie et al. (2024).

435 We note that, with some variations, a notion similar to our average geometry also appeared previ-
436 ously in Ortiz-Jimenez et al. (2020); Movahedi et al. (2025), who study the directional inductive
437 biases of discriminative networks. Interestingly, their analysis predicts that classifiers actually per-
438 form better for data aligned with the *largest* eigenvalues of the geometry. Though their setup is not
439 identical to ours, and therefore not directly comparable, we believe that these seemingly contrasting
440 conclusions may hint at a fundamental trade-off between discriminative and generative modeling.
441 However, such an investigation is out of the scope of this paper and we leave it as future work.

442 Given prior work that connects the NADs with Neural Tangent Kernel (NTK) theory (Jacot et al.,
443 2018; Ortiz-Jimenez et al., 2021), another interesting direction for future research is to investigate
444 potential links between the SADs and the NTK. While SADs capture architectural anisotropy by
445 exploiting the structure of score functions, relying only on forward passes and implicit modeling, the
446 NTK directly uses network Jacobians to describe optimization dynamics of wide neural networks.
447 Regardless of the connection with the NTK, the present findings are interesting on their own as they
448 provide insights on finite capacity networks, which are encountered in practice.

450 4.2 LIMITATIONS

451 Similarly to Kadkhodaie et al. (2024), our experiments, while thorough, focus on relatively small
452 scale settings, which may not reflect the current state-of-the-art. This choice was deliberate: a
453 rigorous validation of our claims at large scales is computationally prohibitive and would require
454 significantly more resources and time. Ultimately, we have prioritized insight as opposed to com-
455 pleteness. Moreover, it is important to stress that we have specifically studied directional inductive
456 biases imposed by the architecture. In principle, the overall dynamics of diffusion models may
457 also be influenced or dominated by different factors such as explicit regularization or other implicit
458 priors, e.g., the double descent phenomenon (Nakkiran et al., 2020).

461 4.3 BROADER IMPACT

462 We investigate the directional inductive biases inherent in score-based generative models and exam-
463 ine their influence on the approximation capabilities of neural networks. We propose a method to
464 systematically characterize these biases, offering insights into both the successes and potential limi-
465 tations of modern diffusion models. Although our focus is primarily theoretical, understanding these
466 mechanisms may guide the design of more effective and efficient generative modeling algorithms.

467 Specifically, deep learning is, at present, an empirical science that is enabled by scale and heavy
468 reliance on heuristics. We believe that our theoretical insights and further developments along this
469 line of research could allow for more cost-effective and principled development of generative tech-
470 nologies. For example, we see potential applications of our work in AutoML (Bergstra et al., 2011)
471 and neural architecture search (Zoph & Le, 2017).

472 Importantly, characterizing the inductive biases of generative models has implications for under-
473 standing and mitigating undesirable behaviors such as memorization of training data (Carlini et al.,
474 2023; Somepalli et al., 2023; Wen et al., 2024; Gu et al., 2025) and hallucination in generated outputs
475 (Aithal et al., 2024; Lu et al., 2025; Floros et al., 2025). By making explicit which patterns the model
476 is predisposed to reproduce, our approach could help identify circumstances under which models are
477 likely to overfit or generate spurious content. This, in turn, may inform the development of tech-
478 niques aimed at reducing privacy risks, improving reliability and enhancing the factual correctness
479 of generated outputs. In this sense, our work contributes not only to the theoretical understanding of
480 generative modeling but also to the broader goal of creating safer and more trustworthy systems.

483 REPRODUCIBILITY STATEMENT

484 Code to reproduce our experiments is in the supplementary material, with details in Appendix A.
485

REFERENCES

- 486
487
488 Sumukh K Aithal, Pratyush Maini, Zachary Chase Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=aNTnHBkw4T>.
489
490
491
- 492 Jie An, De Wang, Pengsheng Guo, Jiebo Luo, and Alex Schwing. On inductive biases that enable
493 generalization in diffusion transformers, 2025.
494
- 495 James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-
496 parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q.
497 Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran
498 Associates, Inc., 2011. URL [https://proceedings.neurips.cc/paper_files/
499 paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf).
- 500 Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr,
501 Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models.
502 In *Proceedings of the 32nd USENIX Conference on Security Symposium, SEC '23, USA, 2023*.
503 USENIX Association. ISBN 978-1-939133-37-3.
- 504 Bradley Efron. Tweedie’s formula and selection bias. *Journal of the American Statistical Association*,
505 106(496):1602–1614, 2011. ISSN 01621459.
506
- 507 Andreas Floros, Seyed-Mohsen Moosavi-Dezfooli, and Pier Luigi Dragotti. Trustworthy image
508 super-resolution via generative pseudoinverse. In *ICLR 2025 Workshop on Deep Generative
509 Model in Machine Learning: Theory, Principle and Efficacy*, 2025. URL <https://openreview.net/forum?id=UsYosqkokz>.
510
- 511 Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition.
512 *Proceedings of the Royal Society A*, 478, 2020.
513
- 514 Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization
515 in diffusion models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL
516 <https://openreview.net/forum?id=D3DBqvSDbj>.
- 517 Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In
518 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Infor-
519 mation Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.
520
- 521 Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and gen-
522 eralization in neural networks. In *Proceedings of the 32nd International Conference on Neural
523 Information Processing Systems, NIPS’18*, pp. 8580–8589, Red Hook, NY, USA, 2018. Curran
524 Associates Inc.
- 525 Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in
526 diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth Inter-
527 national Conference on Learning Representations*, 2024. URL [https://openreview.net/
528 forum?id=ANvmVS2Yr0](https://openreview.net/forum?id=ANvmVS2Yr0).
- 529 Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for im-
530 proved quality, stability, and variation. In *International Conference on Learning Representations*,
531 2018.
532
- 533 Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.
534
- 535 Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]*.
536 Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010.
- 537 Rui Lu, Runzhe Wang, Kaifeng Lyu, Xitai Jiang, Gao Huang, and Mengdi Wang. Towards un-
538 derstanding text hallucination of diffusion models via local generation bias. In *The Thirteenth
539 International Conference on Learning Representations*, 2025. URL [https://openreview.
net/forum?id=SKW10XJ1AI](https://openreview.net/forum?id=SKW10XJ1AI).

- 540 Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approx-
541 imate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017. URL
542 <http://jmlr.org/papers/v18/17-214.html>.
- 543 Takeru Miyato, Sindy Löwe, Andreas Geiger, and Max Welling. Artificial kuramoto oscillatory
544 neurons. In *The Thirteenth International Conference on Learning Representations*, 2025. URL
545 <https://openreview.net/forum?id=nwDRD4AMoN>.
- 546 Sreyas Mohan, Zahra Kadkhodaie, Eero P. Simoncelli, and Carlos Fernandez-Granda. Robust and
547 interpretable blind image denoising via bias-free convolutional neural networks. In *International*
548 *Conference on Learning Representations*, 2020. URL [https://openreview.net/forum?](https://openreview.net/forum?id=HJlSmC4FPS)
549 [id=HJlSmC4FPS](https://openreview.net/forum?id=HJlSmC4FPS).
- 550 Sajad Movahedi, Antonio Orvieto, and Seyed-Mohsen Moosavi-Dezfooli. Geometric inductive bi-
551 ases of deep networks: The role of data and architecture. In *The Thirteenth International Confer-*
552 *ence on Learning Representations*, 2025. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=cmXWYolrlo)
553 [cmXWYolrlo](https://openreview.net/forum?id=cmXWYolrlo).
- 554 Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep
555 double descent: Where bigger models and more data hurt. In *International Conference on Learn-*
556 *ing Representations*, 2020. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=Blg5sA4twr)
557 [Blg5sA4twr](https://openreview.net/forum?id=Blg5sA4twr).
- 558 Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models.
559 In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on*
560 *Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171.
561 PMLR, 18–24 Jul 2021.
- 562 Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi, and Pascal Frossard. Neural
563 anisotropy directions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.),
564 *Advances in Neural Information Processing Systems*, volume 33, pp. 17896–17906. Curran As-
565 sociates, Inc., 2020.
- 566 Guillermo Ortiz-Jimenez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. What
567 can linearized neural networks actually say about generalization? In M. Ranzato,
568 A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neu-*
569 *ral Information Processing Systems*, volume 34, pp. 8998–9010. Curran Associates, Inc.,
570 2021. URL [https://proceedings.neurips.cc/paper_files/paper/2021/](https://proceedings.neurips.cc/paper_files/paper/2021/file/4b5deb9a14d66ab0acc3b8a2360cde7c-Paper.pdf)
571 [file/4b5deb9a14d66ab0acc3b8a2360cde7c-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/4b5deb9a14d66ab0acc3b8a2360cde7c-Paper.pdf).
- 572 William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of*
573 *the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October
574 2023.
- 575 Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mecha-
576 nism for feature learning in neural networks and backpropagation-free machine learning mod-
577 els. *Science*, 383(6690):1461–1467, 2024. doi: 10.1126/science.adi5639. URL [https:](https://www.science.org/doi/abs/10.1126/science.adi5639)
578 [//www.science.org/doi/abs/10.1126/science.adi5639](https://www.science.org/doi/abs/10.1126/science.adi5639).
- 579 Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua
580 Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri
581 and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine*
582 *Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR,
583 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/rahaman19a.html>.
- 584 Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion
585 art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF*
586 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2023. doi:
587 [10.1109/CVPR52729.2023.00586](https://doi.org/10.1109/CVPR52729.2023.00586).
- 588 Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distri-
589 bution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and
590 R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran
591 Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper_files/](https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf)
592 [paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf).

594 Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben
595 Poole. Score-based generative modeling through stochastic differential equations. In *International
596 Conference on Learning Representations*, 2021.

597
598 Yue Song, T. Anderson Keller, Sevan Brodjian, Takeru Miyato, Yisong Yue, Pietro Perona, and
599 Max Welling. Kuramoto orientation diffusion models. In *The Thirty-ninth Annual Conference on
600 Neural Information Processing Systems*, 2025. URL [https://openreview.net/forum?
601 id=dxK2QgEKvz](https://openreview.net/forum?id=dxK2QgEKvz).

602 Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Compu-
603 tation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a.00142.

604
605 Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics.
606 In Lise Getoor and Tobias Scheffer (eds.), *Proceedings of the 28th International Conference on
607 Machine Learning (ICML-11)*, ICML ’11, pp. 681–688, New York, NY, USA, June 2011. ACM.
608 ISBN 978-1-4503-0619-5.

609 Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating
610 memorization in diffusion models. In *The Twelfth International Conference on Learning Repre-
611 sentations*, 2024.

612
613 Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspec-
614 tive of generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin
615 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4697–4708. Cur-
616 ran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper_files/
617 paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf).

618
619 Barret Zoph and Quoc Le. Neural architecture search with reinforcement learning. In *International
620 Conference on Learning Representations*, 2017. URL [https://openreview.net/forum?
621 id=r1Ue8Hcxg](https://openreview.net/forum?id=r1Ue8Hcxg).

622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

Table 1: Hyperparameters of iDDPM and DiT architectures used in the paper. *We train DiT/4 for double the iterations such that the running time is roughly the same as DiT/2 and iDDPM networks.

iDDPM	$\mathcal{N}(\mathbf{0}, Dvv^\top)$				DiT	$\mathcal{N}(\mathbf{0}, Dvv^\top)$	
	MNIST	CelebA-HQ	CIFAR-10		Sphere		
shape	$1 \times 16 \times 16$	$1 \times 28 \times 28$	$1 \times 56 \times 56$	$3 \times 32 \times 32$	$1 \times 16 \times 16$	$1 \times 16 \times 16$	
diffusion_steps	1000	1000	1000	1000	1000	1000	
noise_schedule	linear	linear	linear	linear	linear	linear	
channels	32	32	32	32	patch_size	2/4	2
channel_mults	1, 1	1, 1	1, 1, 1	1, 1, 1	hidden_size	48	48
depth	1	2	2	2	depth	8	8
attn_resolutions	-	-	-	-	mlp_ratio	2	2
num_heads	4	4	4	4	num_heads	4	4
batch_size	1000	500	500	500	batch_size	1000	1000
iterations	2k	100k	100k	200k	iterations	2/4k*	10k
learning_rate	1e-4	1e-3	1e-3	1e-3	learning_rate	1e-4	1e-4

A EXPERIMENTAL SETUP

All of our experiments were conducted on a Linux machine with 128GB of RAM and a NVIDIA RTX 4090 GPU. We train and evaluate diffusion models according to the DDPM framework (Ho et al., 2020), with our hyperparameters included in Table 1. To compute the sliced Wasserstein metrics, we use an overcomplete set of $L = 64D$ random directions. All experiments are repeated five times with different random seeds. Each dataset we consider consists of 10k samples. For sphere modeling, shown in Figure 1, the data is aligned with the first three and the last three SADs. When estimating the average geometry, $\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)$, we use 1M randomly initialized networks.

B REDISCOVERING GEOMETRY-ADAPTIVE HARMONIC BASES

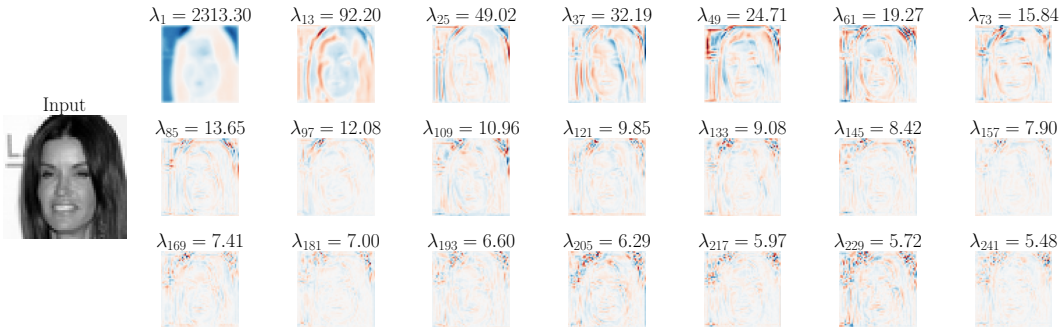


Figure 9: Eigendecomposition of iDDPM average geometry, $\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)$, at initialization with probing distribution $\mathcal{P} = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbf{I}) \times \mathcal{U}(\{\sigma_{\min}, \dots, \sigma_{\max}\})$, i.e., we probe along the standard forward diffusion process with the data sample \mathbf{x} . The input image, \mathbf{x} , is shown on the left and the first few eigenvectors, together with the corresponding unnormalized eigenvalues, are shown on the right. Despite our approach being fundamentally different from existing works, the results resemble the GAHBs of Kadkhodaie et al. (2024). Specifically, we see that our eigenvectors are also adaptive to the geometry of the input image. Moreover, they have harmonic structure as we also observe oscillating patterns whose frequency increases as the eigenvalue decreases.

C GEOMETRIES OF COMMON NEURAL NETWORK ARCHITECTURES

Proposition 1 (MLP geometry, proof in Appendix D.4). *Let \mathcal{F} be the family of networks of the form $\mathbf{z} = \phi(\mathbf{W}\mathbf{h} + \mathbf{b})$, where $\phi: \mathbb{R} \rightarrow \mathbb{R}$ is applied element-wise and $\mathbf{h} \in \mathbb{R}^L$ is an arbitrary function of the input. Assume that \mathbf{W}, \mathbf{b} are completely independent and parameters within each group are identically distributed. Letting $\mathbf{1}$ represent a vector of ones, the geometry takes the form:*

$$\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta) = \alpha_{\mathcal{F}}(\mathcal{P}, \Theta)\mathbf{I} + \beta_{\mathcal{F}}(\mathcal{P}, \Theta)\mathbf{1}\mathbf{1}^\top. \quad (7)$$

Proposition 2 (CNN geometry, proof in Appendix D.5). *Let \mathcal{F} be the family of networks of the form $\mathbf{z} = \phi(\mathbf{W}\mathbf{h} + \mathbf{b})$, where $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is applied element-wise and $\mathbf{h} \in \mathbb{R}^{C_{in} \times L}$ denotes an arbitrary function of the input. \mathbf{W} represents convolution with kernel size k , C_{in} input channels and C_{out} output channels. Assume \mathbf{W}, \mathbf{b} are completely independent and parameters within each group are identically distributed. Then, letting, \otimes denote the Kronecker product, the geometry is a block-diagonal matrix with a block-constant perturbation:*

$$\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta) = \mathbf{I}_{C_{out}} \otimes \mathbf{A}_{\mathcal{F}}(\mathcal{P}, \Theta) + (\mathbf{1}\mathbf{1}^\top)_{C_{out}} \otimes \mathbf{B}_{\mathcal{F}}(\mathcal{P}, \Theta). \quad (8)$$

Proposition 3 (Transformer geometry, proof in Appendix D.6). *Let \mathcal{F} be the family of networks of the form $\mathbf{z} = \mathbf{Q}(\mathbf{W}\mathbf{h} + \mathbf{b})$, where \mathbf{Q} is fixed, orthonormal (e.g., unpatchify) and $\mathbf{h} \in \mathbb{R}^{T \times L_{in}}$ denotes an arbitrary function of the input. $\mathbf{W} : \mathbb{R}^{T \times L_{in}} \rightarrow \mathbb{R}^{T \times L_{out}}$ and \mathbf{b} represent a transformer-style layer operating on T tokens separately. Assume \mathbf{W}, \mathbf{b} are completely independent and parameters within each group are zero-mean, identically distributed. Then, the geometry has at most T distinct eigenvalues and takes the form:*

$$\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta) = \mathbf{Q}[\mathbf{A}_{\mathcal{F}}(\mathcal{P}, \Theta) \otimes \mathbf{I}_{L_{out}}]\mathbf{Q}^\top. \quad (9)$$

D DEFERRED PROOFS

D.1 LEMMAS

Lemma 1. *Consider DSM with data drawn from $\mathcal{N}(\mathbf{0}, \mathbf{v}\mathbf{v}^\top)$ for a fixed noise level $\sigma > 0$ and $\|\mathbf{v}\|_2 = 1$. The associated score function is linear and of the form $\Omega(\cdot)$, with $\Omega = \frac{1}{\sigma^2} \left(\frac{\mathbf{v}\mathbf{v}^\top}{\sigma^2 + 1} - \mathbf{I} \right)$.*

Proof. We have the density $p_\sigma = \mathcal{N}(\mathbf{0}, \mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I})$. The log-density is therefore quadratic, resulting in a linear score function as follows:

$$\nabla_{\mathbf{x}_\sigma} \log p_\sigma(\mathbf{x}_\sigma) = -\frac{1}{2} \nabla_{\mathbf{x}_\sigma} [\mathbf{x}_\sigma^\top (\mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I})^{-1} \mathbf{x}_\sigma] = -(\mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I})^{-1} \mathbf{x}_\sigma. \quad (10)$$

We can further simplify this via the Sherman-Morrison formula:

$$\Omega = -(\mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I})^{-1} = \frac{1}{\sigma^2} \left(\frac{\mathbf{v}\mathbf{v}^\top}{\sigma^2 + 1} - \mathbf{I} \right). \quad (11)$$

□

Lemma 2. *Let \mathbf{W} be a random matrix with entries that are iid, with mean zero and variance σ^2 . For any compatible matrix \mathbf{X} , $\mathbb{E}_{\mathbf{W}}[\mathbf{W}^\top \mathbf{X} \mathbf{W}] = \sigma^2 \text{tr}(\mathbf{X})\mathbf{I}$, where $\text{tr}(\cdot)$ is the trace.*

Proof. Let $(\cdot)^{(i)}$ be column i . For entry (i, j) we can write the expectation as follows:

$$\sum_k \mathbb{E} \left[\mathbf{W}^{(i)\top} \mathbf{X}^{(k)} \mathbf{W}_{(k,j)} \right] = \sum_k \sigma^2 \delta_{i-j} \mathbf{X}_{(k,k)} = \sigma^2 \text{tr}(\mathbf{X}) \delta_{i-j}. \quad (12)$$

□

D.2 THEOREM 1

Proof. Let $\Omega = \Phi\Theta$ be a linear model. The optimization objective is:

$$\begin{aligned} \mathcal{J}_{\text{DSM}}(\Theta) &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{v}\mathbf{v}^\top), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\| \Phi\Theta(\mathbf{x} + \sigma\epsilon) + \frac{\epsilon}{\sigma} \right\|_2^2 \right] \\ &= \mathbb{E} \left[\left\| \Phi\Theta\mathbf{x} + \left(\sigma\Phi\Theta + \frac{\mathbf{I}}{\sigma} \right) \epsilon \right\|_2^2 \right] \\ &\stackrel{(*)}{=} \mathbb{E} \left[\|\Phi\Theta\mathbf{x}\|_2^2 \right] + \mathbb{E} \left[\left\| \left(\sigma\Phi\Theta + \frac{\mathbf{I}}{\sigma} \right) \epsilon \right\|_2^2 \right] = \|\Phi\Theta\mathbf{v}\|_2^2 + \left\| \sigma\Phi\Theta + \frac{\mathbf{I}}{\sigma} \right\|_F^2, \end{aligned} \quad (13)$$

where (*) follows by independence of \mathbf{x} and ϵ . The gradient with respect to Θ is then given by:

$$\nabla_{\Theta} \mathcal{J}_{\text{DSM}}(\Theta) = 2\Phi^\top [\Phi\Theta(\mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I}) + \mathbf{I}]. \quad (14)$$

With this, and for a suitable $\eta > 0$, we express the GD learning dynamics with respect to Ω as:

$$\Omega_t = \Omega_{t-1} - \eta \Phi \nabla_{\Theta} \mathcal{J}_{\text{DSM}}(\Theta) = \Omega_{t-1} - 2\eta \Phi \Phi^\top [\Omega_{t-1}(\mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I}) + \mathbf{I}]. \quad (15)$$

Moreover, we study the error dynamics defined by $\mathbf{E}_t = \Omega_t - \Omega^*$. Here, $\Omega^* = \frac{1}{\sigma^2} \left(\frac{\mathbf{v}\mathbf{v}^\top}{\sigma^2+1} - \mathbf{I} \right)$ is given by Lemma 1 and we can verify that $\Omega^*(\mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I}) + \mathbf{I} = \mathbf{0}$, so it is a stationary point. Therefore, we write:

$$\mathbf{E}_t = \mathbf{E}_{t-1} - 2\eta \Phi \Phi^\top \mathbf{E}_{t-1} (\mathbf{v}\mathbf{v}^\top + \sigma^2\mathbf{I}). \quad (16)$$

Let us focus on the sequence of expected errors, $\mathbb{E}[\mathbf{E}_t]$, where the randomness is over the initialization (and potentially stochastic gradients). Assuming $\mathbb{E}[\Theta_0] = \mathbf{0}$, $\mathbb{E}[\mathbf{E}_0] = -\Omega^*$. Additionally, if $\mathbf{v} \in \{\mathbf{u}_i\}_{i=1}^D$ is an eigenvector of $\Phi\Phi^\top$ with corresponding eigenvalues $\{\lambda_i\}_{i=1}^D$, matrices in Equation 16 commute since they share eigenvectors. This also forces the eigenspaces of all subsequent error terms. In particular, we have:

$$\mathbb{E}[\mathbf{E}_t] = \mathbb{E}[\mathbf{E}_{t-1}] - 2\eta \Phi \Phi^\top (\mathbf{u}_i \mathbf{u}_i^\top + \sigma^2 \mathbf{I}) \mathbb{E}[\mathbf{E}_{t-1}] = -[\mathbf{I} - 2\eta(\lambda_i \mathbf{u}_i \mathbf{u}_i^\top + \sigma^2 \Phi \Phi^\top)]^t \Omega^*, \quad (17)$$

where it is clear that the error decays exponentially. For sufficiently small η , the iterated matrix is positive semidefinite and therefore convergence depends on the minimum eigenvalue of $\lambda_i \mathbf{u}_i \mathbf{u}_i^\top + \sigma^2 \Phi \Phi^\top$, i.e., it is $\mathcal{O}[(1 - 2\eta\rho_i)^t]$ with $\rho_i = \min[(\sigma^2 + 1)\lambda_i, \sigma^2 \min_{j \neq i} \lambda_j]$. Suppose $\exists \lambda_j < \lambda_i$, then $\rho_i = \sigma^2 \lambda_D$, i.e., the convergence rate is fixed for $i < D$. However, for $i = D$ we have $\rho_D = \min[(\sigma^2 + 1)\lambda_D, \sigma^2 \lambda_{D-1}] > \sigma^2 \lambda_D$. That is, we converge faster for $i = D$.

Now, to complete the proof, we focus on the stochastic gradient at optimality:

$$\nabla_{\Theta} \widehat{\mathcal{J}}_{\text{DSM}}(\mathbf{x}, \epsilon; \Theta^*) = 2\mathbf{p}\mathbf{q}^\top, \quad \mathbf{p} = \Phi^\top \left(\Omega^* \mathbf{q} + \frac{\epsilon}{\sigma} \right), \quad \mathbf{q} = \mathbf{x} + \sigma\epsilon. \quad (18)$$

Note, by construction, all of the above quantities are zero-mean. In particular, this implies that \mathbf{p}, \mathbf{q} are uncorrelated and, since they are jointly Gaussian, we claim that they are independent. Therefore, by setting $\mathbf{v} = \mathbf{u}_i$ and vectorizing, we can write the stochastic gradient covariance as:

$$\begin{aligned} 4\mathbb{E} [\text{vec}(\mathbf{p}\mathbf{q}^\top) \text{vec}(\mathbf{p}\mathbf{q}^\top)^\top] &= 4\mathbb{E} [(\mathbf{q}\mathbf{q}^\top) \otimes (\mathbf{p}\mathbf{p}^\top)] \\ &= 4\mathbb{E} [\mathbf{q}\mathbf{q}^\top] \otimes \mathbb{E} [\mathbf{p}\mathbf{p}^\top] \\ &= \frac{4}{\sigma^2(\sigma^2 + 1)} (\mathbf{u}_i \mathbf{u}_i^\top + \sigma^2 \mathbf{I}) \otimes (\Phi^\top \mathbf{u}_i \mathbf{u}_i^\top \Phi) \propto \lambda_i. \end{aligned} \quad (19)$$

□

D.3 THEOREM 2

Proof. We first simplify Equation 6:

$$\alpha = \mathbb{E}_{\mathbf{x} \sim p}[\mathbf{x}^\top (\mathbf{W}^\top \mathbf{G}_{\mathcal{F}} \mathbf{W} \mathbf{x})] = \text{tr}(\mathbf{W}^\top \mathbf{G}_{\mathcal{F}} \mathbf{W} \mathbf{C}), \quad (20)$$

where $\mathbf{C} := \mathbb{E}_{\mathbf{x} \sim p}[\mathbf{x} \mathbf{x}^\top]$ is the second moment of the data. Let $\mathbf{G}_{\mathcal{F}} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top$ and $\mathbf{C} = \mathbf{V} \mathbf{\Sigma} \mathbf{V}^\top$ be eigendecompositions with eigenvalues $\{\lambda_i\}_{i=1}^D$ and $\{\sigma_i\}_{i=1}^D$ respectively. By defining the orthogonal matrix $\mathbf{Q} = \mathbf{U}^\top \mathbf{W} \mathbf{V}$, Equation 20 can be equivalently stated as follows:

$$\begin{aligned} \alpha &= \text{tr}(\mathbf{W}^\top \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top \mathbf{W} \mathbf{V} \mathbf{\Sigma} \mathbf{V}^\top) \\ &= \text{tr}(\mathbf{V}^\top \mathbf{W}^\top \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top \mathbf{W} \mathbf{V} \mathbf{\Sigma}) \\ &= \text{tr}(\mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q} \mathbf{\Sigma}) = \sum_{i=1}^D \sum_{j=1}^D \lambda_i \sigma_j [\mathbf{Q}_{(i,j)}]^2. \end{aligned} \quad (21)$$

Observe that, since \mathbf{Q} is orthogonal, the matrix defined by $\mathbf{P}_{(i,j)} = [\mathbf{Q}_{(i,j)}]^2$ is doubly stochastic. Moreover, Equation 21 shows that optimizing α is a linear problem in \mathbf{P} over the convex set of doubly stochastic matrices. In particular, by Birkhoff’s theorem, α is extremized when \mathbf{P} is a permutation matrix. The minimizing permutation is the one that misaligns the eigenvalues, i.e., $\mathbf{P} = \mathbf{R}$ is achieved if $\mathbf{Q} = \mathbf{R} \iff \mathbf{W} = \mathbf{U} \mathbf{R} \mathbf{V}^\top$. Similarly, α is maximized when the eigenvalues are aligned, i.e., $\mathbf{P} = \mathbf{I}$ if $\mathbf{Q} = \mathbf{I} \iff \mathbf{W} = \mathbf{U} \mathbf{V}^\top$. Indeed, one verifies that these permutations are optimal as they satisfy von Neumann’s trace inequalities with equality. \square

D.4 PROPOSITION 1

Proof. Let Θ denote the specified parameter distribution and consider a probe \mathcal{P} . The geometry is:

$$\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta) = \mathbb{E}_{\mathbf{h}} \mathbb{E}_{\mathbf{W}, \mathbf{b}} [\phi(\mathbf{W} \mathbf{h} + \mathbf{b}) \phi(\mathbf{W} \mathbf{h} + \mathbf{b})^\top]. \quad (22)$$

First, compute the inner expectation, i.e., $\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)|_{\mathbf{h}}$. For indices i, j , since the parameters are iid:

$$[\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)|_{\mathbf{h}}]_{(i,j)} = \begin{cases} \mathbb{E}_{\mathbf{W}, \mathbf{b}} [\phi(\mathbf{W}^{(1)} \mathbf{h} + \mathbf{b}^{(1)})^2] & \text{if } i = j \\ \mathbb{E}_{\mathbf{W}, \mathbf{b}} [\phi(\mathbf{W}^{(1)} \mathbf{h} + \mathbf{b}^{(1)})]^2 & \text{otherwise} \end{cases}. \quad (23)$$

Writing $\mathbf{z}^{(1)} = \phi(\mathbf{W}^{(1)} \mathbf{h} + \mathbf{b}^{(1)})$, we can express the above via the conditional mean, $\mu_{\mathbf{z}^{(1)}|\mathbf{h}}$, and conditional variance $\sigma_{\mathbf{z}^{(1)}|\mathbf{h}}^2$:

$$\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)|_{\mathbf{h}} = \sigma_{\mathbf{z}^{(1)}|\mathbf{h}}^2 \mathbf{I} + \mu_{\mathbf{z}^{(1)}|\mathbf{h}}^2 \mathbf{1} \mathbf{1}^\top, \quad (24)$$

where $\mathbf{1}$ is a vector of ones. Now, taking the outer expectation yields the desired result:

$$\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta) = \mathbb{E}_{\mathbf{h}} [\sigma_{\mathbf{z}^{(1)}|\mathbf{h}}^2] \mathbf{I} + \mathbb{E}_{\mathbf{h}} [\mu_{\mathbf{z}^{(1)}|\mathbf{h}}^2] \mathbf{1} \mathbf{1}^\top. \quad (25)$$

\square

D.5 PROPOSITION 2

Proof. Similarly to our treatment of MLPs in Appendix D.4, we first compute the geometry conditioned on \mathbf{h} , i.e., $\mathbf{G}(\mathcal{P}, \Theta)|_{\mathbf{h}}$. Vectorizing, write the i^{th} input channel as $\mathbf{h}^{(i)} \in \mathbb{R}^L$. Then, \mathbf{W} is a $C_{\text{out}} \times C_{\text{in}}$ block matrix with each $\mathbf{W}_{(i,j)}$ representing convolution with a filter of size k . Therefore, $\mathbf{z}^{(i)} = \phi(\sum_j \mathbf{W}_{(i,j)} \mathbf{h}^{(j)} + \mathbf{b}^{(i)})$. Since the parameters are iid, the conditional geometry is expressed as the following $C_{\text{out}} \times C_{\text{out}}$ block matrix:

$$[\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)|_{\mathbf{h}}]_{(m,n)} = \begin{cases} \mathbb{E}_{\mathbf{W}, \mathbf{b}}[\mathbf{z}^{(1)} \mathbf{z}^{(1)\top}] & \text{if } m = n \\ \mathbb{E}_{\mathbf{W}, \mathbf{b}}[\mathbf{z}^{(1)}] \mathbb{E}_{\mathbf{W}, \mathbf{b}}[\mathbf{z}^{(1)}]^\top & \text{otherwise} \end{cases}. \quad (26)$$

We can rephrase this result in terms of the conditional mean, $\boldsymbol{\mu}_{\mathbf{z}^{(1)}|\mathbf{h}}$, and the conditional covariance $\boldsymbol{\Sigma}_{\mathbf{z}^{(1)}|\mathbf{h}}$. The manipulation is identical to the one used to derive the MLP geometry. Averaging \mathbf{h} :

$$\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta) = \mathbf{I}_{C_{\text{out}}} \otimes \mathbb{E}_{\mathbf{h}}[\boldsymbol{\Sigma}_{\mathbf{z}^{(1)}|\mathbf{h}}] + (\mathbf{1}\mathbf{1}^\top)_{C_{\text{out}}} \otimes \mathbb{E}_{\mathbf{h}}[\boldsymbol{\mu}_{\mathbf{z}^{(1)}|\mathbf{h}} \boldsymbol{\mu}_{\mathbf{z}^{(1)}|\mathbf{h}}^\top]. \quad (27)$$

□

D.6 PROPOSITION 3

Proof. We first focus on $\mathbf{Q} = \mathbf{I}$. After vectorizing \mathbf{h} , \mathbf{W} is block-diagonal with $\mathbf{W}_{(i,i)} = \mathbf{W}_{(1,1)}$ operating on each token, which we write as $\mathbf{h}^{(i)} \in \mathbb{R}^{L_{\text{in}}}$. Conditioned on \mathbf{h} , the geometry is a block matrix with block (i, j) :

$$\begin{aligned} [\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)|_{\mathbf{h}}]_{(i,j)} &= \mathbb{E}_{\mathbf{W}, \mathbf{b}}[(\mathbf{W}_{(1,1)} \mathbf{h}^{(i)} + \mathbf{b}^{(1)})(\mathbf{W}_{(1,1)} \mathbf{h}^{(j)} + \mathbf{b}^{(1)})^\top] \\ &= (\sigma_{\mathbf{W}}^2 \mathbf{h}^{(i)\top} \mathbf{h}^{(j)} + \sigma_{\mathbf{b}}^2) \mathbf{I}, \end{aligned} \quad (28)$$

where the last equality is by Lemma 2, assuming parameters in \mathbf{W} , \mathbf{b} have variances $\sigma_{\mathbf{W}}^2$, $\sigma_{\mathbf{b}}^2$ respectively. Notice that every block is $\propto \mathbf{I}$ and depends on entries of $\mathbf{h}\mathbf{h}^\top \in \mathbb{R}^{T \times T}$, where, with a slight abuse of notation we have reverted to the matrix representation $\mathbf{h} \in \mathbb{R}^{T \times L_{\text{in}}}$. Writing this compactly as $(\sigma_{\mathbf{W}}^2 \mathbf{h}\mathbf{h}^\top + \sigma_{\mathbf{b}}^2 \mathbf{1}\mathbf{1}^\top) \otimes \mathbf{I}_{L_{\text{out}}}$, applying \mathbf{Q} and averaging over \mathbf{h} yields:

$$\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta) = \mathbf{Q} \mathbb{E}_{\mathbf{h}}[(\sigma_{\mathbf{W}}^2 \mathbf{h}\mathbf{h}^\top + \sigma_{\mathbf{b}}^2 \mathbf{1}\mathbf{1}^\top) \otimes \mathbf{I}_{L_{\text{out}}}] \mathbf{Q}^\top. \quad (29)$$

The eigenvalues of the above are products of the eigenvalues of the factors in the Kronecker product, i.e., they are the T eigenvalues of $\sigma_{\mathbf{W}}^2 \mathbb{E}_{\mathbf{h}}[\mathbf{h}\mathbf{h}^\top] + \sigma_{\mathbf{b}}^2 \mathbf{1}\mathbf{1}^\top$.

□

E MORE VISUALIZATIONS

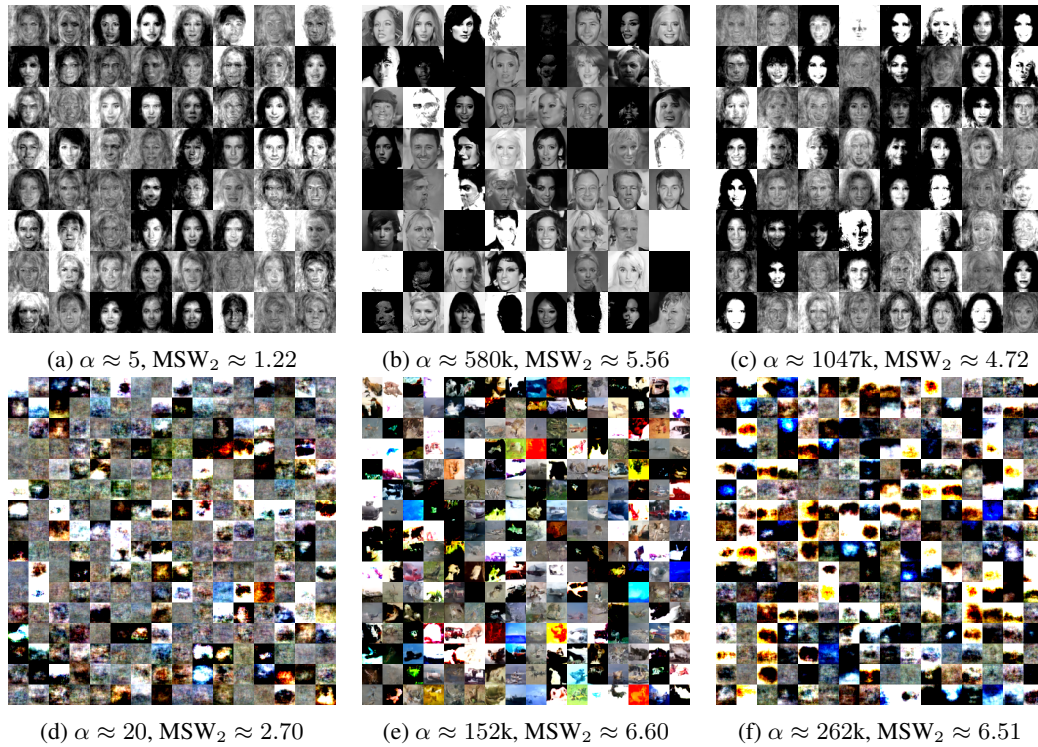


Figure 10: Uncurated samples of iDDPMs trained on CelebA-HQ (top) and CIFAR-10 (bottom). We vary the alignment of the data with the geometry, $\alpha := \mathbb{E}_{x \sim p}[z^\top \mathbf{G}_{\mathcal{F}} z]$, by transforming the data via appropriate matrices \mathbf{W} . The left column corresponds to minimal α , the middle to the default and the right to maximum α . Similarly with the samples of Figure 7, we see that the default alignment may lead to blank / saturated images. However, beyond this observation, it is difficult to draw conclusions based on perceptual quality since it is, by definition, subjective and, formally, does not imply an appropriate metric on distributions. A more meaningful, and statistically sound, comparison is via the Wasserstein metrics which validate our central hypothesis that low alignment leads to the best performance. Note that the networks are thoroughly trained. However, as our focus is on theoretical insight as opposed to completeness, there are several tricks (e.g., larger networks, more data, EMA, clipping) we have not employed that could lead to better quality results.

F SENSITIVITY OF $\mathbf{G}_{\mathcal{F}}$

In Section 3.2 we state that the probing distribution, \mathcal{P} , for the average geometry, $\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)$, is a non-critical hyperparameter and not central to our analysis. Here we explore how injecting anisotropy via \mathcal{P} might influence the underlying eigenvectors of $\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)$ and our estimated SADs. To simplify matters and notation, we will assume that σ is drawn uniformly from the noise levels of interest and the default initialization scheme of the architectures, as in Section 3.2. Under this setup, we will write the geometries for a given family of networks, \mathcal{F} , as $\mathbf{G}_{\mathcal{F}}(\mathcal{P})$ where \mathcal{P} will capture some desired data probe.

For a fixed family (e.g., iDDPM or DiT) we consider degenerate, anisotropic probes \mathcal{P}_i . The data is drawn from δ_{z_i} where $z_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. That is, we first draw a random Gaussian vector and then probe with this fixed input to get a set of geometries $\{\mathbf{G}_{\mathcal{F}}(\mathcal{P}_i)\}_{i=1}^N$. Given this set, we would like to measure the average similarity of eigenvectors. One approach to this problem would be to consider the standard (squared) dot product between corresponding eigenvectors. That is, if $\{\mathbf{u}_d\}_{d=1}^D$ and $\{\mathbf{v}_d\}_{d=1}^D$ are the eigenvectors of $\mathbf{G}_{\mathcal{F}}(\mathcal{P}_i)$ and $\mathbf{G}_{\mathcal{F}}(\mathcal{P}_j)$ respectively, one may compute $\{\mathbf{v}_d^\top \mathbf{u}_d\}_{d=1}^D$ and compare with a random baseline, e.g., the expected squared product between normalized Gaussian vectors is $1/D$. Note, however, that this method only gives a rough estimate of similarity as it does not account for eigen-multiplicity. However, it is straightforward to extend it to a window of size $2h + 1$. Concretely, denote $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_D] \in \mathbb{R}^{D \times D}$. For a given vector, \mathbf{v}_d , we propose to compute the projection energies along the subspace defined by $2h + 1$ consecutive columns of \mathbf{U} . Representing columns starting from a and ending at b as $\mathbf{U}_{a:b} \in \mathbb{R}^{D \times b-a}$, we write:

$$\|\mathbf{U}_{a:b} \mathbf{U}_{a:b}^\top \mathbf{v}_d\|_2^2, \quad a, b, d \in \{1, \dots, D\}, \quad b - a = 2h + 1, \quad a + b \approx 2d. \quad (30)$$

Note that these energies are $\in [0, 1]$ for orthonormal eigenbases and that high energy signals high similarity. Also note that this is a natural extension of cosine similarity, which is recovered by taking $h = 0$. In general, the random baseline for an arbitrary h gives a similarity of $(2h + 1)/D$. Although this is an improved similarity measure compared to the standard dot product, we caution that it is imperfect and may still underestimate similarity for high eigen-multiplicity subspaces when h is not properly calibrated. Nevertheless, our aim here is to show that despite such imperfections the underlying SADs (i.e., the eigenvectors of the geometry) are fairly robust even under extremely anisotropic probes.

In particular, we validate this claim in Figure 11 where we design our experiment for a space with $D = 256$ dimensions, pick $h = 1$ and quantify similarity over $N = 5$ different realizations of geometries. Interestingly, we find that, on average, the similarities are very high relative to the baseline (e.g. $\sim 100\%$ vs. 1.2%), especially for the first few (largest) eigenvectors. Even in the case of the last few (smallest) vectors, we see that the iDDPM U-Net still exhibits significant similarities. For DiT, as previously observed in Figure 6, the architecture shows considerable eigen-multiplicity for small eigenvalue vectors, i.e., $\{\mathbf{u}_d\}_{d=64}^{256}$ in Figure 11 are likely to belong to the same subspace. Therefore, although not explicitly shown in the figure due our choice of fixed, small h , given that that the first few vectors are similar, this forces high similarity over the rest of the spectrum as well.

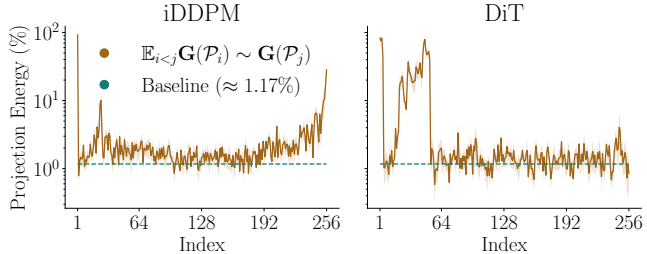


Figure 11: We investigate the robustness of the eigenvectors of $\mathbf{G}_{\mathcal{F}}(\mathcal{P}, \Theta)$ as a function of the probing distribution \mathcal{P} .

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025