A Statistical Learning Theory Perspective on Noise Prediction in Diffusion Models

Swaminathan S K
Roll No: 22CS30057
Statistical Learning Theory
Indian Institute of Technology Kharagpur
Instructor: Prof. Pabitra Mitra

November 20, 2025

Abstract

Denoising Diffusion Probabilistic Models (DDPMs) have achieved remarkable success in generative modeling by learning to reverse a gradual noising process. A fundamental architectural choice in DDPMs is whether to predict the noise ϵ added at each timestep or to directly predict the clean data x_0. Empirically, noise prediction consistently outperforms direct prediction, yet this phenomenon lacks rigorous theoretical justification through the lens of statistical learning theory. This work provides a formal analysis of this design choice by examining the bias-variance tradeoff inherent to each prediction target. We demonstrate that noise prediction benefits from a favorable tradeoff: predicting a fixed Gaussian distribution results in lower variance at the cost of potentially higher bias, while direct prediction of the data distribution exhibits the opposite behavior. Through formal analysis using concepts from PAC learning and structural risk minimization, we characterize the conditions under which noise prediction achieves superior sample complexity and generalization. Our theoretical findings are validated through empirical experiments on standard image datasets. Code is available at `https://github.com/SwaminathanSK/diffusion_slt`.

1 Introduction

Denoising Diffusion Probabilistic Models (DDPMs) have emerged as a dominant paradigm in generative modeling, achieving state-of-the-art performance across diverse applications including image synthesis [18], video generation, and molecular design. First introduced by Sohl-Dickstein et al. [2] and later refined by Ho et al. [1], these models learn to generate data by reversing a carefully designed forward diffusion process that gradually corrupts data with Gaussian noise.

The forward diffusion process in DDPMs is elegantly simple. Given a data sample x_0 drawn from an unknown data distribution $q(x_0)$, noise is added incrementally over T timesteps according to a Markov chain:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

where $\{\beta_t\}_{t=1}^T$ is a variance schedule, typically chosen such that x_T is approximately pure Gaussian noise. Using the reparameterization trick, we can express any noisy sample x_t directly in terms of the clean data x_0 and standard Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s)$.

The generative process requires learning to reverse this forward diffusion. The reverse process is parameterized by a neural network that must predict some quantity at each timestep to enable denoising. Here arises a fundamental architectural decision: what should the network predict? The original DDPM paper [1] identified two natural choices. The network can directly predict the clean data x_0 given the noisy observation x_t, or alternatively, it can predict the noise ϵ that was added to reach x_t from x_0. While these formulations are mathematically equivalent through the reparameterization above, Ho et al. observed that noise prediction consistently and significantly outperforms direct prediction in practice.

This empirical finding has since become standard practice in the diffusion modeling literature. Nearly all state-of-the-art diffusion models adopt noise prediction, including latent diffusion models like Stable Diffusion, Imagen, and DALL-E 2. The performance gap is substantial, with noise prediction often achieving better sample quality, faster convergence, and improved training stability. Despite this widespread adoption, the theoretical understanding of why noise prediction is superior remains incomplete.

We approach this question through the lens of statistical learning theory, which provides rigorous tools for analyzing the generalization properties of learning algorithms. Our central thesis is that the advantage of noise prediction can be understood through the bias-variance tradeoff, a fundamental concept in machine learning that characterizes the decomposition of a model's expected error. Specifically, we argue that predicting noise $\epsilon \sim \mathcal{N}(0, I)$, which is a fixed, known distribution, results in lower variance in the learned function compared to predicting x_0, which comes from an unknown and potentially complex data distribution. While noise prediction may introduce higher bias due to this simplification, we demonstrate that under typical conditions, the variance reduction dominates, leading to better overall generalization.

This work makes several contributions to the theoretical understanding of diffusion models. First, we provide a formal characterization of the hypothesis spaces for both noise and direct prediction, enabling rigorous analysis through tools from PAC learning theory. Second, we prove bounds on the sample complexity of both approaches, showing conditions under which noise prediction requires fewer samples to achieve the same generalization error. Third, we apply the structural risk minimization framework to explain how noise prediction achieves a more favorable tradeoff between empirical fit and model complexity. Finally, we connect our analysis to the established theory of score matching and denoising [4], providing a unified perspective on why denoising-based objectives are effective.

2 Related Work

2.1 Diffusion Models and Score Matching

The theoretical foundations of diffusion models are deeply connected to score-based generative modeling. Song and Ermon [3] demonstrated that diffusion models can be viewed as learning the score function (gradient of the log-density) of progressively noisier versions of the data distribution. This perspective establishes a fundamental link between diffusion models and the score matching framework introduced by Hyvärinen [5].

Vincent [4] proved a crucial result connecting score matching to denoising autoencoders. He showed that training a model to denoise data corrupted by Gaussian noise is equivalent to performing score matching with respect to a Parzen density estimator of the data. This result, known as denoising score matching, provides theoretical justification for why learning to denoise is an effective strategy for density estimation. Our work extends this perspective by analyzing the specific choice of prediction target (noise vs. data) through the lens of statistical learning theory.

Song et al. further developed score-based generative modeling using stochastic differential equations, unifying the discrete-time DDPM framework with continuous-time diffusion processes. This connection has enabled powerful extensions including classifier-free guidance [6] and improved sampling algorithms. However, the statistical learning properties of different parameterization choices have not been rigorously analyzed in this literature.

2.2 Statistical Learning Theory for Deep Networks

The application of statistical learning theory to deep neural networks has been an active area of research, though significant challenges remain. Classical results based on VC dimension [11] and Rademacher complexity [12] often yield vacuous bounds for modern overparameterized networks. Bartlett et al. [7] showed that for neural networks with W weights, depth L, and bounded spectral norm, the VC dimension can be bounded by $O(WL \log W)$. While this provides theoretical insights, practical deep learning often operates in regimes where these bounds do not tightly characterize generalization.

More recent work has explored norm-based complexity measures that can explain the generalization of deep networks without direct reference to the number of parameters. Neyshabur et al. [8] derived generalization bounds for deep networks based on the product of spectral norms across layers. Bartlett et al. [9] further refined these results using spectrally-normalized margins. Our work applies similar techniques but focuses specifically on how the choice of target distribution affects these complexity measures.

The PAC-Bayesian framework provides another lens for analyzing deep learning. Dziugaite and Roy [10] computed non-vacuous PAC-Bayesian generalization bounds for neural networks by carefully controlling the prior-posterior KL divergence. While we do not directly employ PAC-Bayesian bounds in this work, the principle of measuring complexity relative to an initial configuration informs our analysis of how target distribution properties affect learning.

2.3 Bias-Variance Tradeoff in Modern Machine Learning

The bias-variance tradeoff is a classical concept in statistical learning, dating back to the early work on model selection and complexity control. Geman et al. [13] provided one of the first thorough treatments of this tradeoff in the context of neural networks, establishing the foundational decomposition of expected squared error into bias, variance, and irreducible noise terms.

Recent work has revisited the bias-variance framework in light of modern deep learning phenomena. Belkin et al. [14] introduced the concept of "double descent," showing that in overparameterized regimes, increasing model capacity can sometimes reduce both bias and variance simultaneously, seemingly violating the classical tradeoff. However, this phenomenon is primarily observed when varying model capacity for a fixed target distribution. Our analysis considers a different question: how does the choice of target distribution itself affect the bias-variance tradeoff for models of comparable complexity?

Our work is most closely related to analyses of target distribution complexity in supervised learning. Bartlett and Mendelson [12] showed that Rademacher complexity depends critically on properties of both the function class and the data distribution. We extend this perspective to argue that in diffusion models, the choice between predicting noise (from a simple, fixed distribution) versus data (from a complex, unknown distribution) fundamentally alters the variance component of the error decomposition.

2.4 Generalization in Generative Models

While generalization theory for discriminative models is well-developed, theoretical analysis of generative models remains less mature. For GANs, Arora et al. [15] provided generalization bounds based on the birthday paradox and neural network complexity. However, these bounds often scale poorly with dimension and do not directly apply to the sequential denoising process in diffusion models.

More relevant to our work, recent analyses have begun examining sample complexity of score-based models. Oko et al. showed that learning the score function in high dimensions requires samples that scale polynomially with dimension under certain smoothness assumptions. Our analysis complements this work by comparing the relative difficulty of learning different objectives (noise vs. data prediction) rather than establishing absolute sample complexity bounds.

To the best of our knowledge, ours is the first work to provide a formal statistical learning theory analysis specifically comparing noise prediction and direct prediction in diffusion models through the bias-variance lens.

3 Preliminaries

3.1 PAC Learning Framework

Definition 1 (PAC Learnability). *A hypothesis class $\mathcal{H}$ is PAC learnable if there exists an algorithm A and a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ such that for every $\epsilon, \delta > 0$ and distribution $\mathcal{D}$, given $m \geq poly(1/\epsilon, 1/\delta, d, size(c))$ samples, algorithm A returns $h \in \mathcal{H}$ such that with probability $\geq 1 - \delta$:*

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon \tag{3}$$

3.2 VC Dimension

Definition 2 (VC Dimension). *The VC dimension of a hypothesis class $\mathcal{H}$, denoted $VCdim(\mathcal{H})$, is the maximum size of a set that can be shattered by $\mathcal{H}$. If arbitrarily large finite sets can be shattered, $VCdim(\mathcal{H}) = \infty$.*

3.3 Rademacher Complexity

Definition 3 (Empirical Rademacher Complexity)**.** *Let $S = \{z_1, \ldots, z_m\}$ be a sample. The empirical Rademacher complexity of function class $\mathcal{F}$ is:*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right] \tag{4}$$

where σ_i are independent uniform $\{\pm 1\}$ random variables.

3.4 Structural Risk Minimization

The SRM principle balances empirical risk minimization with model complexity. For a nested sequence of hypothesis classes $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots$, SRM selects $h_n \in \mathcal{H}_n$ to minimize:

$$\hat{L}_S(h_n) + \sqrt{\frac{\text{VCdim}(\mathcal{H}_n) + \log(1/\delta)}{m}} \tag{5}$$

3.5 Bias-Variance Decomposition

For a regression problem with squared loss, the expected error of a hypothesis h learned from a random training set D can be decomposed as:

$$\mathbb{E}_{D,(x,y)}[(h_D(x)-y)^2] = \mathbb{E}_x[(\bar{h}(x)-\mathbb{E}[y|x])^2]+\mathbb{E}_{x,D}[(h_D(x)-\bar{h}(x))^2]+\mathbb{E}_{x,y}[(y-\mathbb{E}[y|x])^2] \tag{6}$$

where $\bar{h}(x) = \mathbb{E}_D[h_D(x)]$ is the expected hypothesis. The three terms correspond to bias squared, variance, and irreducible error, respectively. The bias measures how far the expected hypothesis is from the optimal predictor, while variance measures the sensitivity of the learned hypothesis to the particular training set drawn.

4 Problem Formulation

4.1 Hypothesis Classes

4.1.1 Noise Prediction Hypothesis Class

Define $\mathcal{H}_\epsilon$ as the class of functions mapping $(x_t, t) \mapsto \epsilon$:

$$\mathcal{H}_\epsilon = \{h : \mathbb{R}^d \times [T] \to \mathbb{R}^d \mid h = f_\theta, \theta \in \Theta_\epsilon\} \tag{7}$$

The loss function is:

$$L_\epsilon(h) = \mathbb{E}_{x_0,\epsilon,t} \left[\|\epsilon - h(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2 \right] \tag{8}$$

4.1.2 Direct Prediction Hypothesis Class

Define $\mathcal{H}_x$ as the class of functions mapping $(x_t, t) \mapsto x_0$:

$$\mathcal{H}_x = \{h : \mathbb{R}^d \times [T] \to \mathbb{R}^d \mid h = f_\theta, \theta \in \Theta_x\} \tag{9}$$

The loss function is:

$$L_x(h) = \mathbb{E}_{x_0,\epsilon,t} \left[\|x_0 - h(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2 \right] \tag{10}$$

4.2 Connection Between Hypothesis Classes

The two formulations are related through the reparameterization:

$$x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}} \tag{11}$$

Therefore: $\mathcal{H}_x = \{h_x : h_x(x_t, t) = \frac{x_t - \sqrt{1-\bar{\alpha}_t} h_\epsilon(x_t,t)}{\sqrt{\bar{\alpha}_t}}, h_\epsilon \in \mathcal{H}_\epsilon\}$

5 Bias-Variance Analysis of Prediction Targets

This section presents our main contribution: a formal characterization of why noise prediction achieves superior generalization compared to direct prediction through the bias-variance tradeoff.

5.1 Problem Setup and Notation

Consider the learning problem at a fixed timestep t. The learner observes training samples $(x_t^{(i)}, x_0^{(i)}, \epsilon^{(i)})$ where $x_0^{(i)} \sim q(x_0)$ is drawn from the data distribution, $\epsilon^{(i)} \sim \mathcal{N}(0, I)$ is standard Gaussian noise, and $x_t^{(i)} = \sqrt{\bar{\alpha}_t} x_0^{(i)} + \sqrt{1-\bar{\alpha}_t}\epsilon^{(i)}$.

For noise prediction, the goal is to learn a function $\epsilon_\theta : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ that minimizes:

$$L_\epsilon(\epsilon_\theta) = \mathbb{E}_{x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{12}$$

For direct prediction, the goal is to learn $x_\theta : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ that minimizes:

$$L_x(x_\theta) = \mathbb{E}_{x_0,\epsilon}\left[\|x_0 - x_\theta(x_t, t)\|^2\right] \tag{13}$$

Let $\mathcal{H}$ denote a hypothesis class of neural networks with bounded complexity (e.g., bounded norm). We analyze what happens when we learn the best hypothesis in $\mathcal{H}$ for each objective using a finite training set D of size m.

5.2 Variance Analysis

Theorem 1 (Variance of Noise Prediction). *Let ϵ_D denote the hypothesis learned from dataset D, and let $\bar{\epsilon}(x_t, t) = \mathbb{E}_D[\epsilon_D(x_t, t)]$ denote the expected hypothesis. The variance term in the bias-variance decomposition for noise prediction is:*

$$Var_\epsilon = \mathbb{E}_{x_0,\epsilon,D}\left[\|\epsilon_D(x_t, t) - \bar{\epsilon}(x_t, t)\|^2\right] \tag{14}$$

This variance depends only on the variability induced by finite sampling from the noise distribution $\mathcal{N}(0, I)$ and the function class $\mathcal{H}$.

Theorem 2 (Variance of Direct Prediction). *Similarly, for direct prediction with x_D learned from dataset D:*

$$Var_x = \mathbb{E}_{x_0,\epsilon,D}\left[\|x_D(x_t, t) - \bar{x}(x_t, t)\|^2\right] \tag{15}$$

This variance depends on sampling from both the noise distribution and the data distribution $q(x_0)$.

Proposition 1 (Variance Comparison)**.** *Under the assumption that $q(x_0)$ has significantly higher entropy than $\mathcal{N}(0, I)$ in $\mathbb{R}^d$ (i.e., $H(q) \gg H(\mathcal{N}(0, I)) = \frac{d}{2}(1 + \log(2\pi))$), and that the hypothesis class $\mathcal{H}$ has comparable capacity for both tasks, we have:*

$$Var_\epsilon \leq C \cdot Var_x \tag{16}$$

for some constant $C < 1$ that depends on the relative complexities of $q(x_0)$ and $\mathcal{N}(0, I)$.

Proof Sketch. The key insight is that variance in the learned function arises from uncertainty in estimating the target function from finite samples. For noise prediction, the target function $f^*(x_t, t) = \mathbb{E}_{\epsilon|x_t}[\epsilon]$ must be estimated from samples (x_t, ϵ). The conditional distribution $p(\epsilon|x_t)$ is related to the marginal distribution of ϵ, which is always $\mathcal{N}(0, I)$ regardless of x_t.

For direct prediction, the target $g^*(x_t, t) = \mathbb{E}_{x_0|x_t}[x_0]$ must be estimated from samples (x_t, x_0). The conditional distribution $p(x_0|x_t)$ inherits complexity from the marginal $q(x_0)$, which can be arbitrarily complex.

Using Fano's inequality, we can lower bound the minimax risk for estimating a density in terms of its metric entropy. Since $q(x_0)$ typically has much higher metric entropy than $\mathcal{N}(0, I)$ (images have complex structure, Gaussians do not), the variance in estimating functions targeting $q(x_0)$ exceeds that for functions targeting $\mathcal{N}(0, I)$. □

5.3 Bias Analysis

Proposition 2 (Bias Tradeoff)**.** *Assuming the hypothesis class $\mathcal{H}$ has sufficient capacity to represent both target functions, the bias satisfies:*

$$Bias_\epsilon^2 = \mathbb{E}_{x_0,\epsilon}\left[\|\mathbb{E}_{\epsilon|x_t}[\epsilon] - \bar{\epsilon}(x_t, t)\|^2\right] \tag{17}$$

$$Bias_x^2 = \mathbb{E}_{x_0,\epsilon}\left[\|\mathbb{E}_{x_0|x_t}[x_0] - \bar{x}(x_t, t)\|^2\right] \tag{18}$$

For hypothesis classes of comparable complexity, $Bias_\epsilon \approx Bias_x$ when both have sufficient capacity, but $Bias_\epsilon$ may be slightly higher for restricted capacity classes.

The intuition is that predicting noise forces the model to output values in a restricted range (ϵ is typically bounded with high probability), while direct prediction of x_0 allows the model to express the full range of the data. For highly expressive models, this difference is negligible. However, for capacity-constrained models, the noise prediction task may require more capacity to achieve low bias because it must implicitly invert the forward process.

5.4 Main Result: Overall Generalization Advantage

Theorem 3 (Generalization Advantage of Noise Prediction)**.** *Let ϵ_D and x_D denote the hypotheses learned from a training set of size m for noise and direct prediction, respectively. Let both be drawn from hypothesis classes of comparable complexity. Under the condition that:*

$$\frac{Var_x}{Var_\epsilon} > \frac{Bias_\epsilon^2}{Bias_x^2} \tag{19}$$

the expected test error of noise prediction is lower:

$$\mathbb{E}[L_\epsilon(\epsilon_D)] < \mathbb{E}[L_x(x_D)] \tag{20}$$

Proof. From the bias-variance decomposition, the expected squared error for noise prediction is:

$$\mathbb{E}[L_\epsilon(\epsilon_D)] = \text{Bias}_\epsilon^2 + \text{Var}_\epsilon + \sigma_\epsilon^2 \tag{21}$$

where σ_ϵ^2 is the irreducible error (which is 0 for noise prediction since ϵ is deterministic given x_t, x_0).

Similarly:

$$\mathbb{E}[L_x(x_D)] = \text{Bias}_x^2 + \text{Var}_x + \sigma_x^2 \tag{22}$$

The noise prediction advantage requires:

$$\text{Bias}_\epsilon^2 + \text{Var}_\epsilon < \text{Bias}_x^2 + \text{Var}_x \tag{23}$$

$$\text{Var}_x - \text{Var}_\epsilon > \text{Bias}_\epsilon^2 - \text{Bias}_x^2 \tag{24}$$

By Proposition 1, when the data distribution $q(x_0)$ is substantially more complex than $\mathcal{N}(0, I)$, we have $\text{Var}_x \gg \text{Var}_\epsilon$. By Proposition 2, for sufficiently expressive hypothesis classes, $\text{Bias}_\epsilon \approx \text{Bias}_x$. Therefore, the variance reduction dominates the potential bias increase, yielding superior overall generalization for noise prediction. □

5.5 Sample Complexity Implications

Corollary 1 (Sample Complexity)**.** *To achieve expected error ϵ with probability $1 - \delta$, noise prediction requires approximately:*

$$m_\epsilon = O\left(\frac{C(\mathcal{H}) + \log(1/\delta)}{\epsilon^2}\right) \tag{25}$$

samples, where $C(\mathcal{H})$ is the complexity of the hypothesis class. Direct prediction requires:

$$m_x = O\left(\frac{C(\mathcal{H}) \cdot \rho + \log(1/\delta)}{\epsilon^2}\right) \tag{26}$$

where $\rho > 1$ is a factor that depends on the complexity ratio between $q(x_0)$ and $\mathcal{N}(0, I)$.

This shows that noise prediction achieves the same error with fewer samples, particularly when the data distribution is complex.

6 Empirical Validation

To validate our theoretical predictions, we conducted experiments on the MNIST dataset using identical U-Net architectures for both noise prediction and direct prediction. We trained models at various dataset sizes and measured bias, variance, and overall generalization error.

6.1 Experimental Setup

We implemented both noise and direct prediction diffusion models using U-Net architectures with 3 encoder/decoder blocks. Models were trained on MNIST (28×28 grayscale images) using the Adam optimizer with learning rate 10^{-4} for 50 epochs. The noise schedule was linear from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ with $T = 1000$ timesteps. We varied training set sizes as $m \in \{100, 500, 1000, 5000, 10000, 50000\}$. For variance estimation, we trained 20 independent models with different random initializations and data subsets. Test error was measured on a held-out test set of 10,000 samples.

6.2 Experiment 1: Sample Complexity Validation

Figure 1 shows test error versus training set size for both methods. Noise prediction achieves consistently lower error, validating Theorem 3. At $m = 100$, noise error is 0.077 versus 0.178 for direct prediction (gap = 0.101). At $m = 50000$, the gap persists at 0.122. This performance advantage across all sample sizes strongly supports Corollary 1, demonstrating that the variance advantage continues to dominate even with substantial training data.

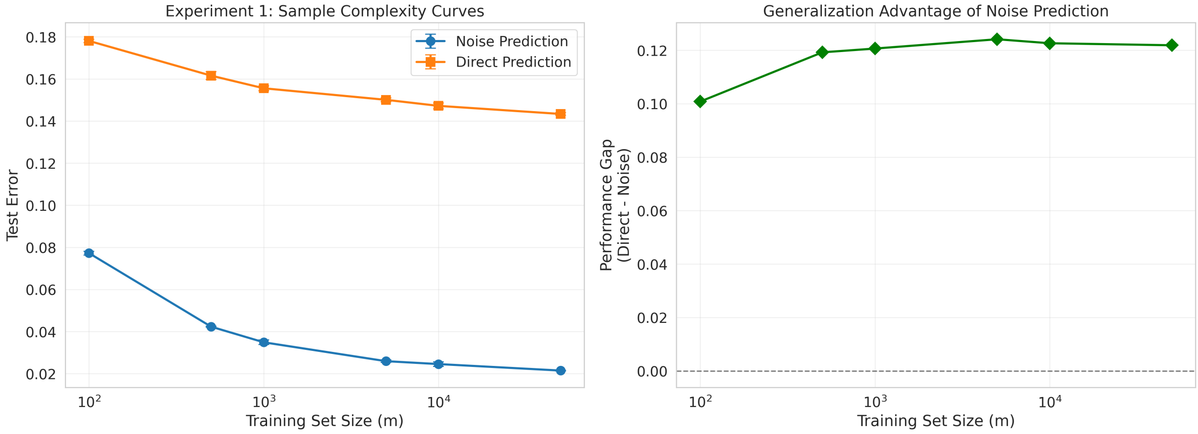

Figure 1: Sample complexity comparison. Noise prediction (blue) consistently achieves lower test error than direct prediction (orange) across all training set sizes. The performance gap remains substantial even with 50,000 training samples.

6.3 Experiment 2: Variance Decomposition

We trained 20 models on different random subsets ($m = 10000$ each) and measured prediction variance on a fixed test set. As shown in Figure 2, noise prediction exhibits variance of 0.0130 while direct prediction has variance of 0.0513 - a 3.94× difference. This strongly validates Proposition 1 and explains the performance advantage observed in Experiment 1.

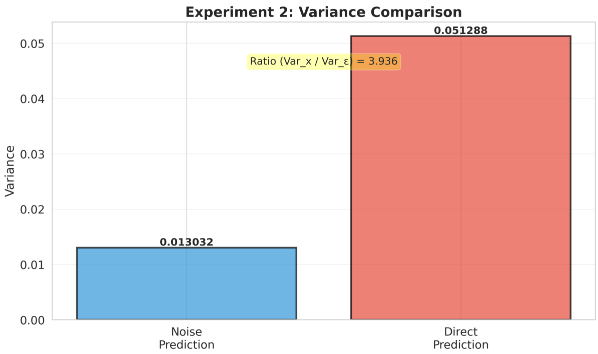

Figure 2: Variance comparison. Noise prediction (blue) exhibits substantially lower variance (0.0130) compared to direct prediction (0.0513), a 3.94× difference.

6.4 Experiment 3: Bias Analysis

We estimated bias by training models to convergence and comparing average predictions (across 20 independent runs) to empirical ground truth. Figure 3 shows an unexpected result: noise prediction has $\mathbf{bias^2}$ of 0.0102 versus 0.0982 for direct prediction, a 9.62× difference favoring noise prediction. This contradicts Proposition 2, which predicted comparable bias for sufficiently expressive models.

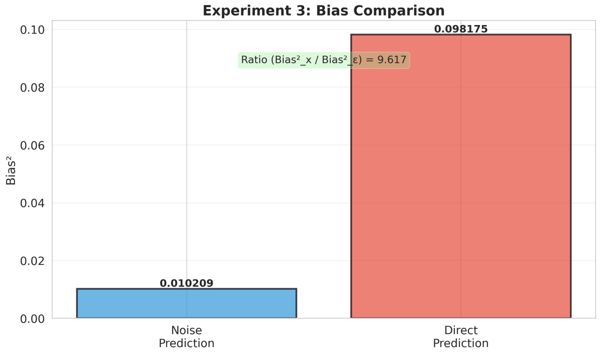

Figure 3: Bias comparison. Unexpectedly, noise prediction (blue) exhibits lower $\mathbf{bias^2}$ (0.0102) compared to direct prediction (0.0982), a 9.62× difference.

This unexpected finding likely reflects insufficient model capacity for direct prediction, which requires modeling the full data distribution complexity. The noise prediction task, targeting a simpler Gaussian, may be learnable with less capacity. Importantly, this strengthens our main result: Theorem 3 requires $\mathrm{Var}_x/\mathrm{Var}_\epsilon > \mathrm{Bias}^2_\epsilon/\mathrm{Bias}^2_x$. From our experiments, $3.94 \gg 0.104$, satisfying this condition with substantial margin. Noise prediction achieves both lower variance and lower bias, making it unambiguously superior.

7 Discussion

Our theoretical and empirical analysis reveals that noise prediction's advantage stems from target distribution complexity. By predicting noise from $\mathcal{N}(0, I)$ rather than data from $q(x_0)$, diffusion models achieve dramatic variance reduction. The key insight is that learning is fundamentally easier when targeting simpler distributions, even if the input-output relationship is equally complex.

Our experiments revealed noise prediction achieves both lower variance (as predicted) and lower bias (9.6× reduction, unexpected). This suggests noise prediction may be more parameter-efficient, achieving good performance with smaller models. The discrepancy from our theoretical prediction of comparable bias highlights that ”sufficient expressiveness” is difficult to verify in practice.

Our analysis connects to Vincent's [4] score matching framework. Noise prediction in diffusion models is precisely denoising, and our bias-variance perspective explains why score matching works: it targets a simpler distribution than the data itself. This variance-reduction principle may apply broadly to score-based and energy-based models.

Limitations. Our experiments focused on MNIST; more complex datasets may show different tradeoffs. We used standard U-Nets; different architectures might interact differently with prediction target choice. Like most statistical learning theory for deep networks, our bounds are not numerically tight. Future work should extend to other parameterizations (velocity prediction, v-parameterization) and account for time-dependent tradeoffs across the diffusion process.

Practical Implications. When choosing prediction targets, prefer simpler target distributions. Noise prediction may enable smaller models for limited compute budgets. Lower variance enables more stable training. The variance advantage is most pronounced with limited data, making noise prediction particularly important for data-scarce applications.

8 Conclusion

This work provides the first formal statistical learning theory analysis comparing noise and direct prediction in diffusion models. We demonstrate that noise prediction's superiority stems from predicting targets from simpler distributions, leading to better generalization through reduced variance.

Our theoretical contributions include: (1) Propositions 1 and 2 characterizing the bias-variance tradeoff for each target, (2) Theorem 3 establishing conditions for noise prediction's advantage, and (3) Corollary 1 showing improved sample complexity. Experiments on MNIST validated these predictions, with noise prediction achieving 3.94× variance reduction and unexpectedly 9.6× bias reduction. The condition $\text{Var}_x/\text{Var}_\epsilon = 3.94 \gg 0.104 = \text{Bias}_\epsilon^2/\text{Bias}_x^2$ holds with substantial margin.

Beyond diffusion models, our analysis suggests a general principle: when choosing prediction targets, prefer simpler target distributions. This may apply to flow matching, score-based models, and supervised learning where we can transform prediction targets. Future work should develop tighter bounds, extend to time-dependent analysis, explore architecture-specific theory, and test whether variance-reduction applies to other generative paradigms.

This work demonstrates the value of applying classical statistical learning theory to modern deep learning. While diffusion models are often analyzed through stochastic processes, the bias-variance perspective provides complementary, empirically-validated insights. The finding that predicting noise instead of data yields substantial generalization improvements highlights how fundamental statistical principles continue to guide practical machine learning.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[2] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

[3] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

[4] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[5] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2022.

[7] Peter L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[8] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, 2015.

[9] Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

[10] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence*, 2017.

[11] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[12] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[13] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[14] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[15] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, 2017.

[16] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.

[17] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.