# Improved probabilistic regression using diffusion models

Carlo Kneissl*[1,2], Christopher Bülte*[1,2], Philipp Scholl*[1,2], and Gitta Kutyniok[1,2,3,4]

[1]Ludwig-Maximilians-Universität München
[2]Munich Center for Machine Learning (MCML)
[3]University of Tromsø
[4]DLR-German Aerospace Center
{kneissl, buelte, scholl, kutyniok}@math.lmu.de

## Abstract

Probabilistic regression models the entire predictive distribution of a response variable, offering richer insights than classical point estimates and directly allowing for uncertainty quantification. While diffusion-based generative models have shown remarkable success in generating high-dimensional data, their usage in regression tasks often lacks uncertainty-related evaluation. We propose a novel diffusion-based framework for probabilistic regression where we model the full distribution of the diffusion noise, enabling adaptation to diverse tasks and enhanced uncertainty quantification.

## 1 Introduction

Supervised regression aims at predicting a response variable $\boldsymbol{y} \in \mathcal{Y}$ from covariates $\boldsymbol{c} \in \mathcal{C}$. Classical approaches estimate the conditional mean $\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{c}]$, whereas probabilistic regression models the full predictive distribution $p_{\mathcal{Y}}(\boldsymbol{y} \mid \boldsymbol{c})$ [1, 2], typically via a Gaussian [3, 4]. Recent work has emphasized more flexible, non-parametric alternatives [2, 5].

Diffusion-based generative models have emerged as state-of-the-art approaches for high-dimensional data generation, achieving remarkable results in tasks such as photorealistic image [6] and video synthesis [7].

Recently, diffusion models have been applied to various regression tasks such as depth estimation [8], autoregressive flow prediction [9, 10], and weather forecasting [11, 12], often achieving state-of-the-art performance. Despite their inherently probabilistic nature, evaluations rarely emphasize uncertainty-related metrics or calibration. Recent efforts aim to extract uncertainty estimates from diffusion models [13–15], but typically rely on training multiple networks, incurring substantial computational overhead. Furthermore, the intimate relation between uncertainty quantification and the noise modeling within the diffusion process remains underexplored.

**Contributions:** In this work, we address these limitations by adapting the diffusion process to yield calibrated probabilistic predictions. Building on recent advances from generative modeling [16], we introduce a novel loss for diffusion-based regression models that enables learning a parameterized noise distribution, and offers task-specific trade-offs between expressivity and computational efficiency.

*Equal contribution

## 2 Background

### 2.1 Probabilistic regression

Let $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$ denote the response variables of interest and $\boldsymbol{c} \in \mathcal{C} \subseteq \mathbb{R}^{d_c}$ the corresponding conditioning variables. Given training data $\mathcal{D} = (\boldsymbol{c}_i, \boldsymbol{y}_i)_{i=1}^N$, the goal is to recover the predictive distribution $p_{\mathcal{Y}}(\boldsymbol{y} \mid \boldsymbol{c}, \mathcal{D})$.

### 2.2 Diffusion models

Diffusion probabilistic models (DPMs) aim to learn a target distribution $p_0$ on $\mathbb{R}^d$ from samples by estimating the reverse dynamics of a diffusion process. We follow the non-Markovian formulation of denoising diffusion implicit models (DDIM) [17].

Let $T \in \mathbb{N}$, $\boldsymbol{x}_0 \sim p_0$, and $\beta_{1:T} \in [0,1]^T$ denote a noise schedule. Define $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$. The *forward process* produces intermediate states

$$\boldsymbol{X}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{X}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad (1)$$

such that each $x_t$ is a noisy version of $x_0$. For sufficiently small $\bar{\alpha}_T$, $x_T$ is close to a standard normal, providing a tractable prior.

To generate samples, one must approximate the reverse process $p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$, which is intractable in general. DPMs approximate it with a latent-variable model

$$p_\theta(\boldsymbol{x}_{0:T}) := p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t), \qquad (2)$$

which is a Markov chain that samples from $\boldsymbol{x}_T$ to $\boldsymbol{x}_0$, that is referred to as the *generative process*. For sufficiently small $\beta_t$, the reverse transition $p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ is well approximated by a Gaussian,

thus allowing to set $p_\theta(\boldsymbol{x}_T) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and specifying the latent variable model as a neural network, $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t))$.

Training proceeds by minimizing the MSE loss of the noise samples

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t,x_0,\epsilon_t}\left[\|\epsilon_t - \epsilon_\theta(\boldsymbol{x}_t, t)\|_2^2\right]. \quad (3)$$

The target distribution is set as $p_0 = p_{\mathcal{Y}}(\cdot \mid \boldsymbol{c})$, yielding an approximate predictive distribution $p_\theta(\cdot \mid \boldsymbol{c}) \approx p_{\mathcal{Y}}(\cdot \mid \boldsymbol{c})$.

While diffusion models have shown great success in conditional and unconditional generative modeling, note that due to the objective in Equation (3), the latent variable model only learns to approximate $\epsilon_\theta(\boldsymbol{x}_t, t) \approx \mathbb{E}[\epsilon_t \mid \boldsymbol{x}_t]$ and does not capture information about the full distribution $p(\epsilon_t \mid \boldsymbol{x}_t)$.

### 2.3 Scoring rules

Let $\mathcal{P}$ be a convex set of probability measures on $\mathcal{Y}$ and $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$. A scoring rule $S(\mathbb{P}, \boldsymbol{y})$ [18] is a function that measures the discrepancy between a predictive distribution $\mathbb{P}$ and an observation $\boldsymbol{y} \in \mathcal{Y}$. The expected score is defined as $S(\mathbb{P}, \mathbb{Q}) := \mathbb{E}_{Y\sim\mathbb{Q}}[S(\mathbb{P}, Y)]$. A scoring rule is called *proper* if $S(\mathbb{Q}, \mathbb{Q}) \leq S(\mathbb{P}, \mathbb{Q})$ for all $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$, and *strictly proper* if equality holds if and only if $\mathbb{Q} = \mathbb{P}$. Therefore, strictly proper scoring rules ensure that the true distribution is the unique optimum and have been successfully applied in training neural networks for various tasks [19–21].

## 3 Our Methodology

### 3.1 Learning $p_\theta^\epsilon(\cdot \mid \boldsymbol{x}_t)$

Most diffusion frameworks, following DDPM [22], fix the variance of the noise distribution in each denoising step. This design was originally motivated by two observations: (i) learning the variance often destabilizes training, and (ii) variance modeling showed little benefit for image generation benchmarks, for example, with respect to the FID [23].

However, subsequent work [24] demonstrated that learning the variance improves likelihood estimates, indicating that recovering only the mean is insufficient for faithfully approximating the conditional distribution. Furthermore, the Gaussian approximation of $p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ is only valid when the number of timesteps $T$ is large. Yet, large $T$ is computationally costly, and recent results suggest that using as few as 20–50 steps can yield superior performance in regression tasks [10, 12], particularly with improved noise schedulers and solvers [25, 26].

These observations motivate us to go beyond estimating the first two moments and instead learn the full distribution of $\epsilon_t$. Specifically, we reinterpret $\epsilon_\theta$: rather than treating it as a point estimate of $\epsilon_t$, we view it as a random variable. This perspective naturally suggests replacing the mean-squared error loss with a criterion that compares probability distributions rather than point predictions. To this end, we adopt the framework of strictly proper scoring rules [20, 27].

### 3.2 Parametrization of $p_\theta^\epsilon(\cdot \mid \boldsymbol{x}_t)$

Concurrent work by Bortoli et al. [16] reached a similar conclusion. They proposed modeling the full posterior distribution $p(\boldsymbol{x}_0 \mid \boldsymbol{x}_t)$ by a neural network and generating samples via noise concatenation to the inputs.

This, however, comes at significant computational cost: multiple samples increase the runtime. Empirically, training is reported to be 1.3× to 7× slower than standard diffusion models [16].

As an alternative to nonparametric sampling-based approaches, we propose to model $p_\theta^\epsilon(\cdot \mid \boldsymbol{x}_t)$ directly through a parametrized distribution. This aligns naturally with scoring rule minimization, since closed-form expressions for the training objective are available, and thereby reducing the training time significantly as we do not need to generate samples. Depending on the choice of parametrization, one obtains a trade-off between computational efficiency and flexibility, which may vary across tasks.

Specifically, we consider for $p_\theta^\epsilon(\epsilon_t \mid \boldsymbol{x}_t)$ the general Gaussian mixture form

$$\sum_{k=1}^{K} \pi_{\theta,k} \mathcal{N}(\epsilon_t; \boldsymbol{\mu}_{\theta,k}^\epsilon(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_{\theta,k}^\epsilon(\boldsymbol{x}_t, t)). \quad (4)$$

From this general specification, we highlight three concrete parametrizations that offer a trade-off between the expressivity of the distribution and the simplicity of the model training:

**Univariate Gaussian:** For $K = 1$ and $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) = \text{diag}(\sigma_\theta^2(\boldsymbol{x}_t, t))$, $\sigma_\theta^2(\boldsymbol{x}_t, t) \in \mathbb{R}_{>0}^{d_y}$, we obtain a simple baseline.

**Univariate Gaussian mixture.** Setting $\boldsymbol{\Sigma}_{\theta,k}(\boldsymbol{x}_t, t) = \text{diag}(\sigma_{\theta,k}^2(\boldsymbol{x}_t, t))$ with $\sigma_{\theta,k}^2(\boldsymbol{x}_t, t) \in \mathbb{R}_{>0}^{d_y}$ and $K > 1$ we get a Gaussian mixture that approximates arbitrary densities under mild assumptions [1, 28].

**Multivariate Gaussian:** Choosing $K = 1$ with full covariance $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ allows modeling correlations between the marginals of $\epsilon_t$.

## 4 Conclusion

We propose to learn the full diffusion noise distribution using a Gaussian mixture model, aiming to improve probabilistic regression performance and decrease training time.

# References


[1] C. Bishop. *Mixture density networks*. English. WorkingPaper. Aston University, 1994.

[2] X. Shen and N. Meinshausen. "Engression: extrapolation through the lens of distributional regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 87.3 (Nov. 2024), pp. 653–677. ISSN: 1369-7412. DOI: 10.1093/jrsssb/qkae108. URL: https://doi.org/10.1093/jrsssb/qkae108.

[3] D. Nix and A. Weigend. "Estimating the mean and variance of the target probability distribution". In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. Vol. 1. 1994, 55–60 vol.1. DOI: 10.1109/ICNN.1994.374138.

[4] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

[5] D. M. Kelen, Á. Jung, P. Kersch, and A. A. Benczur. "Distribution-Free Data Uncertainty for Neural Network Regression". In: *The Thirteenth International Conference on Learning Representations*. Oct. 2024. (Visited on 06/04/2025).

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[7] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. "Video Diffusion Models". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 8633–8646. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

[8] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler. "Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 9492–9502. DOI: 10.1109/CVPR52733.2024.00907.

[9] M. A. Finzi, A. Boral, A. G. Wilson, F. Sha, and L. Zepeda-Nunez. "User-defined Event Sampling and Uncertainty Quantification in Diffusion Models for Physical Dynamical Systems". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 10136–10152. URL: https://proceedings.mlr.press/v202/finzi23a.html.

[10] G. Kohl, L.-W. Chen, and N. Thuerey. *Benchmarking Autoregressive Conditional Diffusion Models for Turbulent Flow Simulation*. en. arXiv:2309.01745 [cs]. Dec. 2024. DOI: 10.48550/arXiv.2309.01745. URL: http://arxiv.org/abs/2309.01745 (visited on 06/12/2025).

[11] G. Couairon, R. Singh, A. Charantonis, C. Lessig, and C. Monteleoni. *ArchesWeather & ArchesWeatherGen: a deterministic and generative model for efficient ML weather forecasting*. 2024. arXiv: 2412.12971 [cs.LG]. URL: https://arxiv.org/abs/2412.12971.

[12] I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, and M. Willson. "Probabilistic Weather Forecasting with Machine Learning". In: *Nature* 637.8044 (Jan. 2025), pp. 84–90. ISSN: 1476-4687. DOI: 10.1038/s41586-024-08252-9. (Visited on 08/21/2025).

[13] L. Berry, A. Brando, and D. Meger. "Shedding Light on Large Generative Networks: Estimating Epistemic Uncertainty in Diffusion Models". In: *The 40th Conference on Uncertainty in Artificial Intelligence*. 2024. URL: https://openreview.net/forum?id=512IkGDqA8.

[14] M. A. Chan, M. J. Molina, and C. Metzler. "Estimating Epistemic and Aleatoric Uncertainty with a Single Model". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024. URL: https://openreview.net/forum?id=WPxa6OcIdg.

[15] D. Shu and A. B. Farimani. *Zero-Shot Uncertainty Quantification using Diffusion Probabilistic Models*. 2024. arXiv: 2408.04718 [cs.LG]. URL: https://arxiv.org/abs/2408.04718.

[16] V. D. Bortoli, A. Galashov, J. S. Guntupalli, G. Zhou, K. P. Murphy, A. Gretton, and A. Doucet. "Distributional Diffusion Models with Scoring Rules". In: *Forty-second International Conference on Machine Learning*. 2025. URL: https://openreview.net/forum?id=N82967FcVK.

[17] J. Song, C. Meng, and S. Ermon. "Denoising Diffusion Implicit Models". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=St1giarCHLP.

[18] T. Gneiting and A. E. Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378. ISSN: 0162-1459. DOI: 10.1198/016214506000001437.

[19] J. Chen, T. Janke, F. Steinke, and S. Lerch. "Generative machine learning methods for multivariate ensemble postprocessing". In: *The Annals of Applied Statistics* 18.1 (2024), pp. 159–183. DOI: 10.1214/23-AOAS1784. URL: https://doi.org/10.1214/23-AOAS1784.

[20] L. Pacchiardi, R. A. Adewoyin, P. Dueben, and R. Dutta. "Probabilistic Forecasting with Generative Networks via Scoring Rule Minimization". In: *Journal of Machine Learning Research* 25.45 (2024), pp. 1–64. URL: http://jmlr.org/papers/v25/23-0038.html.

[21] C. Bülte, P. Scholl, and G. Kutyniok. "Probabilistic neural operators for functional uncertainty quantification". In: *Transactions on Machine Learning Research* (2025). ISSN: 2835-8856. URL: https://openreview.net/forum?id=gangoPXSRw.

[22] J. Ho, A. Jain, and P. Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

[24] A. Q. Nichol and P. Dhariwal. "Improved Denoising Diffusion Probabilistic Models". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 8162–8171. URL: https://proceedings.mlr.press/v139/nichol21a.html.

[25] H. Chung, B. Sim, and J. C. Ye. "Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 12403–12412. DOI: 10.1109/CVPR52688.2022.01209.

[26] T. Karras, M. Aittala, T. Aila, and S. Laine. "Elucidating the Design Space of Diffusion-Based Generative Models". In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho. 2022. URL: https://openreview.net/forum?id=k7FuTOWMOc7.

[27] A. P. Dawid, M. Musio, and L. Ventura. "Minimum Scoring Rule Inference". In: *Scandinavian Journal of Statistics* 43.1 (2016), pp. 123–138. DOI: https://doi.org/10.1111/sjos.12168. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjos.12168. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12168.

[28] K. Plataniotis and D. Hatzinakos. "Gaussian mixtures and their applications to signal processing". In: Jan. 2000.