Approximation of Subsurface Flow via Physics Informed Neural Netwokrs

Denis Basharin[a], Anatolii Frolov[a], Daniil Maksimov[a], Kirill Katsuba[a]

[a] *Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Bolshoy Boulevard, Moscow, 121205, , Russia*

Abstract

This project proposes a novel approach to training diffusion models on datasets that suffer from inherently missing data. While standard diffusion models require clean, complete samples for training, this project introduces a method to learn the true underlying data distribution directly from corrupted (masked) samples. We validate this approach on synthetic image data (MNIST) with the ultimate goal of applicability to tabular data.

Keywords: Machine-Learning, Deep Learning, Diffusion models, Bayesian ML, MDLM

1. Introduction

Recent advancements in Generative Artificial Intelligence have been dominated by Denoising Diffusion Probabilistic Models (DDPMs), which have achieved state-of-the-art results in image synthesis and data generation. However, the standard training protocols for these models rely on a critical assumption: access to a large dataset of complete, clean samples from the underlying distribution. In real-world scenarios such as medical records or tabular datasets this assumption rarely holds. Instead, practitioners often face ambient or corrupted data, where a significant portion of the training instances are missed.

Our project "Ambient Masked Diffusion via Consistency Property" addresses the challenge of training diffusion models directly on datasets with inherently missing values. Unlike traditional two-stage approaches that require separate imputation before training, we propose an end-to-end framework that learns to recover the true data distribution from corrupted samples alone.

Our methodology integrates Masked Diffusion Language Models [1] with Ambient Learning principles [2]. By enforcing a Consistency Property which ensures stability between diffusion time steps and utilizing an Expectation-Maximization algorithm to iteratively refine missing data, we demonstrate that it is possible to generate high-quality samples without ever seeing a full ground-truth image during initialization. We validate this approach on the MNIST dataset, showing significant improvements over baseline methods.

2. Related Work

Our work sits at the intersection of discrete generative modeling, inverse problem solving or ambient diffusion, and data imputation. In this section, we review the foundational methodologies that motivate our proposed MDLM with consistency regularization.

2.1. Masked Diffusion Language Models (MDLM)

Standard diffusion models typically operate in continuous space using Gaussian noise. However, for categorical data or binarized images, discrete diffusion processes offer a more natural formulation. **Masked Diffusion Language Models** [1] simplify the complex transition matrices of earlier discrete diffusion works by utilizing a simple absorbing state.

In MDLM, the forward process $q(x_t|x_0)$ independently corrupts tokens in the input sequence x_0 by replacing them with a special `[MASK]` token with probability γ_t. The transition probability for a single token x at time t is given by:

$$q(x_t|x_0) = \begin{cases} 1-\gamma_t & \text{if } x_t = x_0 \\ \gamma_t & \text{if } x_t = \texttt{[MASK]} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The generative process reverses this by training a network $p_\theta(x_0|x_t)$ to predict the original unmasked tokens. Loss of the model have following structure:

$$\mathcal{L}_{\text{MDLM}} = \mathbb{E}_{t\sim U(0,1), x_t\sim q(x_t|x_0)} \left[-\sum_{i\in\mathcal{M}_t} \log p_\theta(x_0^{(i)}|x_t) \right] \tag{2}$$

where $\mathcal{M}_t$ represents the set of indices masked at time t.

2.2. Consistent Diffusion Meets Tweedie

The challenge of learning generative models from corrupted observations—often or **Ambient Diffusion** is addressed in **Consistent Diffusion Meets Tweedie** [2]. This work focuses on the scenario where training data consists only of noisy measurements $\mathbf{y} = \mathbf{x}_0 + \mathbf{n}$.

Authores propose a training objective that leverages Tweedie's Formula, which relates the posterior mean of the clean signal to the score of the data distribution. For a Gaussian noise level σ, the conditional expectation is:

$$\mathbb{E}[\mathbf{x}_0|\mathbf{z}_t] = \mathbf{z}_t + \sigma_t^2 \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t) \tag{3}$$

They introduce a **consistency loss** that enforces the denoising network $f_\theta(\mathbf{z}_t)$ to be consistent with the available noisy observations $\mathbf{y}$. By deriving a pseudo-target $\hat{\mathbf{x}}(\mathbf{y}, t)$ based on the conditional posterior $p(\mathbf{x}_0|\mathbf{y})$, they consider following los function:

$$\mathcal{L}_{\text{consist}} = \mathbb{E}_{t,\mathbf{z}_t} \left\| f_\theta(\mathbf{z}_t) - \hat{\mathbf{x}}(\mathbf{y}, t) \right\|^2 \tag{4}$$

This allows the model to hallucinate the missing high-frequency details lost in $\mathbf{y}$ while remaining consistent with the low-frequency information present. Our project adapts this consistency principle—specifically the idea that the expectation of the clean image must remain stable across time steps—from the continuous Gaussian domain to the discrete masked domain.

2.3. DiffPuter: Diffusion-based Imputation

Generative models have increasingly been deployed for missing value imputation. **DiffPuter** [3] introduces a framework that frames imputation as an Expectation-Maximization problem. Given observed data $\mathbf{x}_{obs}$ and missing data $\mathbf{x}_{mis}$, DiffPuter aims to maximize the log-likelihood of the observed components $\log p_\theta(\mathbf{x}_{obs})$.

The method iteratively refines the missing values (E-step) and updates the model parameters (M-step). The objective can be decomposed via the ELBO as:

$$\mathcal{L}_{\text{EM}} = \mathbb{E}_{q(\mathbf{x}_{mis}|\mathbf{x}_{obs})} \left[\log p_\theta(\mathbf{x}_{obs}, \mathbf{x}_{mis})\right] \tag{5}$$

In practice, DiffPuter implements this by injecting noise into the imputed values and denoising them conditioned on the fixed $\mathbf{x}_{obs}$, effectively learning the conditional distribution $p(\mathbf{x}_{mis}|\mathbf{x}_{obs})$.

3. Methodology

We propose a framework for training generative models on heavily corrupted data by using Discrete Diffusion, Consistency Regularization, and Expectation-Maximization.

3.1. Discrete Diffusion with Absorbing States

Unlike standard Gaussian diffusion, we model the data $\mathbf{x}$ as a sequence of discrete tokens from a set $V = \{0, 1\}$. The corruption process is modeled as a transition to a special absorbing token $[\texttt{MASK}]$. Let the extended set be $V^+ = V \cup \{[\texttt{MASK}]\}$.

The forward diffusion process $\{\mathbf{z}_t\}_{t=0}^T$ is a Markov chain that gradually masks the clean input $\mathbf{x}_0$. The transition probability for a single token at time t is defined by the transition matrix $\mathbf{Q}_t$:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \text{Cat}(\mathbf{z}_t; \mathbf{z}_{t-1}\mathbf{Q}_t) \tag{6}$$

where $\mathbf{Q}_t$ is designed such that tokens have a probability β_t of transitioning to $[\texttt{MASK}]$ and remaining there. A key property of this formulation is the closed-form marginal at any timestep t:

$$q(\mathbf{z}_t|\mathbf{x}_0) = \text{Cat}(\mathbf{z}_t; \mathbf{x}_0\bar{\mathbf{Q}}_t) \tag{7}$$

where $\bar{\mathbf{Q}}_t = \mathbf{Q}_1 \dots \mathbf{Q}_t$. Effectively, at time t, a pixel is observed with probability α_t and masked with probability $1 - \alpha_t$.

The reverse process $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$ is parameterized by a neural network $f_\theta(\mathbf{z}_t, t)$ that predicts the categorical distribution of the clean image, denoted as $\hat{p}_\theta(\mathbf{x}_0|\mathbf{z}_t)$. The reverse transition is then computed analytically using Bayes' rule:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) \propto q(\mathbf{z}_t|\mathbf{z}_{t-1})\hat{p}_\theta(\mathbf{x}_0|\mathbf{z}_t) \tag{8}$$

3.2. The Ambient Learning Objective

In our setting, we do not have access to $\mathbf{x}_0$. Instead, we observe a corrupted subset $\mathbf{x}_{obs}$. Let $m \in \{0, 1\}^D$ be a binary mask indicating observed indices. The training objective approximates ELBO by decomposing the loss into two components based on observability.

3.2.1. 1. The Easy Loss (Supervised)

For pixels that are observed, we have the ground truth. We apply standard diffusion training logic here. The model takes a further corrupted version $\mathbf{z}_t$ and tries to recover the values known in $\mathbf{x}_{obs}$.

$$\mathcal{L}_{\text{easy}} = -\sum_{i:m_i=1} \log p_\theta(x_0^{(i)} = x_{obs}^{(i)} \mid \mathbf{z}_t) \tag{9}$$

3.2.2. Consistency

For missing pixels, we lack ground truth. To supervise the model, we enforce the **Consistency Property**. Theoretical results in continuous diffusion (Tweedie's formula from [2]) suggest that the conditional expectation of the clean data, $\mathbb{E}[\mathbf{x}_0|\mathbf{z}_t]$, forms a martingale. We translate this to the discrete domain.

We require that the model's prediction of $\mathbf{x}_0$ at a noisy step $t+1$ should be consistent with the expected prediction at a cleaner step t. Since $\mathbf{z}_{t+1}$ contains less information than $\mathbf{z}_t$, the prediction at t (what we called teacher) is generally more accurate than at $t+1$ (what we called student).

$$\mathcal{L}_{\text{hard}} = \mathbb{E}_{\mathbf{z}_t \sim q(\cdot|\mathbf{z}_{t+1})} \left[D_{KL} \left(\text{sg}[\hat{p}_\theta(\mathbf{x}_0|\mathbf{z}_t)] \parallel \hat{p}_\theta(\mathbf{x}_0|\mathbf{z}_{t+1}) \right) \right] \tag{10}$$

where sg$[\cdot]$ denotes the stop-gradient operator. This prevents the teacher from degrading to match the student, ensuring information flows from lower noise levels to higher noise levels.

3.3. EM algorithm

While consistency regularization stabilizes training, it assumes the model can eventually infer the correct distribution from noise. To bootstrap this process when 90% of data is missing, we use EM:

- **E-Step (Imputation):** Given the current model parameters $\theta^{(k)}$, we estimate the missing values $\mathbf{x}_{miss}$ by sampling from the generative model conditioned on the observed parts:

  $$\hat{\mathbf{x}}_{miss}^{(k)} \sim p_{\theta^{(k)}}(\mathbf{x}_{miss}|\mathbf{x}_{obs}) \tag{11}$$

  This creates a completed dataset $\mathcal{D}^{(k)}$.

- **M-Step (Maximization):** We update $\theta^{(k+1)}$ by minimizing the combined loss on the completed dataset $\mathcal{D}^{(k)}$:

  $$\theta^{(k+1)} \leftarrow \arg\min_\theta \left(\mathcal{L}_{\text{easy}}(\mathcal{D}^{(k)}) + \lambda \mathcal{L}_{\text{hard}}(\mathcal{D}^{(k)}) \right) \tag{12}$$

In the result the imputed data improves the model, and the improved model yields better imputations.

4. Results

We evaluated our method on the Binarized MNIST dataset under a severe corruption setting where 90% of pixels were masked during training. We compared three configurations:

1. **Baseline:** Standard MDLM training on visible pixels only.
2. **EM Only:** Iterative imputation without consistency regularization.
3. **EM + Consistency:** Our full proposed method.

4.1. Qualitative Analysis

As shown in our experimental visualizations, the **Baseline** model fails completely. Due to the sparsity of the signal (only 10% visible), the model cannot learn global digit structure and produces largely incoherent noise.

The **EM Only** approach demonstrates the ability to latch onto digit modalities. After several iterations of self-imputation, the model begins to generate recognizable shapes. However, the digits often lack sharpness and exhibit artifacts, suggesting the model converges to a local optimum biased by early incorrect imputations.

The **EM + Consistency** approach yields the highest fidelity results. The consistency constraint forces the generative process to be robust across time steps, effectively smoothing out the noise introduced during the imputation phase. The resulting digits exhibit clean strokes and clear class separability, closely resembling the ground truth distribution despite the model never observing a complete image during initialization.

4.2. Challenges and Stability

We observed that training discrete consistency models is highly sensitive. If the model's prediction at time t is low-confidence, forcing the model at $t + 1$ to match it simply propagates uncertainty. We addressed this by forcing pixels—discretizing the teacher's soft predictions before using them as targets—which stabilized the learning objective.

5. Conclusion

In this work, we presented a framework for **Ambient Masked Diffusion**, enabling the training of generative language models on datasets with severe missingness. By integrating the theoretical consistency property into the discrete MDLM architecture and wrapping the training in an EM loop, we successfully recovered the MNIST data distribution from 90% masked samples.

Our results highlight that while imputation provides a pathway to learn from missing data, it is the addition of **Consistency Regularization** that ensures the learned distribution is coherent and sharp. This method holds significant promise for real-world applications such as tabular data imputation or imaging reconstruction, where fully observed ground truth data is often unavailable. Future work will focus on stabilizing the teacher-student dynamics further and extending this method to continuous-valued tabular datasets.

Hope, we can continue to work on this project to make strong paper.

References

[1] S. S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, T. Marwala, T. Kolda, and C. Raffel, “Simple and effective masked diffusion language models,” *arXiv preprint arXiv:2406.07524*, 2024.

[2] B. Kawar, S. Zada, H. Ben-Hamu, and M. Elad, “Consistent diffusion meets tweedie: Training exact ambient diffusion models with noisy data,” *arXiv preprint arXiv:2404.10177*, 2024.

[3] Y. Xu *et al.*, “Diffputer: A diffusion-based imputation method for missing values,” *arXiv preprint arXiv:2405.20690*, 2024.