# Stochastic Interpolants and Ising Samples

Luca Raffo
EPFL, Institute of Mathematics - luca.raffo@epfl.ch
Capital Fund Management - luca.raffo@cfm.com

June 2025

## Contents



## 1 Introduction

*Stochastic interpolants*, recently introduced in [1], provide a general and flexible framework for generative modeling. They naturally extend the score-based diffusion models of [4], which, however, rely on the assumption that the underlying distributions are continuous.

In many applications of interest, the data is inherently discrete, with the Ising model being a prominent example. In such cases, one can either treat the samples as continuous and apply the standard diffusion framework, or instead build upon the ideas of [3], who introduced *discrete diffusion models*, specifically designed to handle probability measures supported on discrete spaces.

In this work we show how we can use both to learn how to sample from the Ising model. Later on, we will see how to induce sparsity.

# 2 Stochastic Interpolants

Suppose there exists an unknown absolutely continuous measure $\mu_1$ with density $\rho_1$. Our goal is to find a way to sample from $\mu_1$.

The strategy is as follows: we fix a simple absolutely continuous measure $\mu_0$, with density $\rho_0$ (for instance a Gaussian), and we learn a vector field $b_t$ such that

$$\begin{cases} X_{t=0} \sim \mu_0 \\ \dot{X}_t = b_t(X_t), \end{cases} \qquad \Longrightarrow \qquad X_1 \sim \mu_1.$$

Let us formalize this idea.

## 2.1 Flows of measures

The starting points is the *Stochastic Interpolant*

$$I_t := \alpha_t X_0 + \beta_t X_1,$$

where we require that $\alpha_0 = \beta_1 = 1$, $\alpha_1 = \beta_0 = 0$ (think for instance as $\alpha_t = 1 - t, \beta_t = t$).

Both $X_0$ and $X_1$ are stochastic objects (assumed here to be also independent), and accordingly $I_t$ will be a (doubly-)stochastic object. In particular, it can be shown that (almost always) it is distributed according to an absolutely continuous measure which we call $\mu_t$, with density $\rho_t$.

Let us define the vector field $b_t$ as

$$b_t(x) := \mathbb{E}[\dot{I}_t \mid I_t = x].$$

The main result is due to the following theorem.

**Theorem 2.1.** Let $X_t$ be the solution to the system

$$\begin{cases} X_{t=0} \sim \mu_0 \\ \dot{X}_t = b_t(X_t), \end{cases} \tag{1}$$

then $X_t$ is distributed according to $\mu_t$ (same as $I_t$).

The proof can be found in the appendix. In particular, this means that if $X_{t=0} \sim \mu_0$, then $X_{t=1} \sim \mu_1$.

## 2.2 Learning the vector field

Suppose we are only given some samples from $\mu_1$. We can generate as many samples as we want from $\mu_0$. How do we find an approximate $b_t$? We use a smart characterization of conditional expectation, the proof can be found in the appendix.

**Theorem 2.2.** The vector field $b_t(x) = \mathbb{E}[\dot{I}_t \mid I_t = x]$ can be locally characterized as

$$b_t = \arg\min_{\hat{b}_t} \mathbb{E}[|\hat{b}_t(I_t)|^2 - 2\dot{I}_t \cdot \hat{b}_t(I_t)]$$

and also globally characterized as

$$b = \arg\min_{\hat{b}} \int_0^1 \mathbb{E}[|\hat{b}_t(I_t)|^2 - 2\dot{I}_t \cdot \hat{b}_t(I_t)]\, dt.$$

In practice, we will learn a deep neural network $\hat{b}_t^\theta$, that aims at minimizing the loss function

$$\mathcal{L}(\theta) := \mathbb{E}[|\hat{b}_t(I_t)|^2 - 2\dot{I}_t \cdot \hat{b}_t(I_t)], \tag{2}$$

where the expectation is over $X_0 \sim \mu_0, X_1 \sim \mu_1$ and $t \sim uniform[0, 1]$.

## 2.3 Adding some noise

Looking at (1), we can write down the corresponding continuity equation (that the density $\rho_t$ has to satisfy),

$$\partial_t \rho_t + \nabla \cdot (b_t \rho_t) = 0. \tag{3}$$

We define $s_t(x) := \nabla \log \rho_t(x)$, and as the identity $\nabla \cdot (s_t \rho_t) = \Delta \rho_t$ holds, (3) is equivalent to

$$\partial_t \rho_t + \nabla \cdot (b_t \rho_t) + \epsilon_t \nabla \cdot (s_t \rho_t) = \epsilon_t \Delta \rho_t, \tag{4}$$

so that if we call $X_t^{fp}$ the solution of

$$\begin{cases} X_{t=0} \sim \mu_0 \\ dX_t^\epsilon = b_t(X_t^\epsilon)dt + \epsilon_t s_t(X_t^\epsilon) + \sqrt{2\epsilon_t} dW_t, \end{cases} \tag{5}$$

then $X_t^{fp}$ will have the same distribution as $X_t$. To learn the score $s_t$, we can use another functional to minimize with a neural network.

The functional to minimize, introduced in [2], is

$$\mathcal{L}(\phi) = \mathbb{E}[\|s_t^\phi(I_t)\|^2 + 2\nabla \cdot s_t^\phi(I_t)], \tag{6}$$

where the expectation is again intended to be upon $X_0 \sim \mu_0, X_1 \sim \mu_1$ and $t \sim uniform[0,1]$.

In particular, if $\mu_0$ is Gaussian, then learning $b_t$ is enough as $s_t$ can be written in close form as a function of the former.

**Theorem 2.3.** If $\mu_0$ is a Gaussian measure, then $s_t(x) = -\frac{1}{\alpha_t}\mathbb{E}[X_0 \mid I_t = x]$.

This is shown in the appendix, and the only idea behind is the Stein's lemma.

Once we have this, we can derive our claim.

$$\begin{cases} b_t(x) = \dot{\alpha}_t \mathbb{E}[X_0 \mid I_t = x] + \dot{\beta}_t \mathbb{E}[X_1 \mid I_t = x] \\ s_t(x) = -\frac{1}{\alpha_t}\mathbb{E}[X_0 \mid I_t = x] \\ x = \alpha_t \mathbb{E}[X_0 \mid I_t = x] + \beta_t \mathbb{E}[X_1 \mid I_t = x], \end{cases}$$

is solved by

$$\alpha_t s_t(x) = \frac{\beta_t b_t(x) - \dot{\beta}_t}{\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t}.$$

There are furthermore some theoretical guarantees for the choice of $\epsilon_t$ which can be found in [1].

## 2.4 Link with score-based diffusion models

In score-based diffusion models we start with the following proxy. We have some samples from $\mu_1$ which we evolve towards gaussian noise (with a OU process). In the process, we learn a score function, which is sufficient to reverse the process, i.e. to move gaussian samples towards $\mu_1$ samples.

Formally, we define $\tilde{X}_\tau$ as the solution of

$$\begin{cases} \tilde{X}_{\tau=0} = X_1 \sim \mu_1 \\ d\tilde{X}_\tau = -\tilde{X}_\tau d\tau + \sqrt{2} dW_\tau. \end{cases} \tag{7}$$

This is integrated as

$$\begin{aligned} \tilde{X}_\tau &= e^{-\tau} X_1 + \sqrt{2} \int_0^\tau e^{-\tau+\tau'} \, dW_{\tau'} \\ &\overset{d}{=} e^{-\tau} X_1 + \sqrt{1 - e^{-2\tau}} X_0 \overset{\tau \to +\infty}{\longrightarrow} X_0, \end{aligned}$$

where $X_0 \sim \mu_0$ is a sample from a Gaussian.

Let us call $\tilde{\rho}_\tau := dist(\tilde{X}_\tau)$. Then the attached Fokker-Planck equation reads

$$\partial_\tau \tilde{\rho}_\tau - \nabla \cdot (x \tilde{\rho}_\tau) = \Delta \tilde{\rho}_\tau. \tag{8}$$

Now let $T \gg 1$ and let $\rho_\tau := \tilde{\rho}_{T-\tau}$, so that $\tilde{\rho}_T \approx \rho_0$. Then,

$$\begin{aligned} \partial_\tau \rho_\tau &= -\partial_\tau \tilde{\rho}_{T-\tau} \\ &= -\nabla \cdot (x \tilde{\rho}_{T-\tau}) - \Delta \tilde{\rho}_{T-\tau} \\ &= -\nabla \cdot (x \rho_\tau) - 2\nabla \cdot (\tilde{s}_{T-\tau} \rho_\tau) + \Delta \rho_\tau, \end{aligned}$$

Since $X_\tau := \tilde{X}_{T-\tau}$ (and therefore $dist(X_\tau) = dist(\tilde{X}_{T-\tau})$), the time reversed SDE allows us to move from $\tilde{\rho}_T \approx \rho_0$ towards $\tilde{\rho}_0 = \rho_1$, given that we are able

to learn the score function just by minimizing the functional (6) (instead of using $I_t$ of course we use $\tilde{X}_\tau$):

$$dX_\tau = X_\tau d\tau + 2\nabla \log \tilde{\rho}_{T-\tau}(X_\tau)d\tau + \sqrt{2}dW_\tau.$$

The link with interpolants is easy:

$$\begin{aligned}\tilde{X}_\tau &= e^{-\tau}X_1 + \sqrt{2}\int_0^\tau e^{-\tau+\tau'}\, dW_{\tau'} \\ &\overset{d}{=} e^{-\tau}X_1 + \sqrt{1-e^{-2\tau}}X_0\end{aligned}$$

and

$$e^{-\tau} = t \Longleftrightarrow \tau = -\log t,$$

so that finally

$$\tilde{X}_{-\log t} \overset{d}{=} \overset{\beta_t}{t}X_1 + (\overset{\alpha_t}{1-t})X_0 = I_t.$$

# 3 Discrete Diffusion Models

# 4 Sampling of Ising Models

# 5 NESS Sampling

# References

[1] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2023.

[2] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[3] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024.

[4] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

# A Appendix

Here we collect additional derivations, proofs, and technical details that complement the main text.