AN OPINIONATED INTRODUCTION TO STOCHASTIC LOCALIZATION

RUSHIL MALLARAPU

ABSTRACT. Stochastic localization is a technique at the intersection of high-dimensional probability theory and stochastic process theory which aims to understand complicated high-dimensional distributions by evolving them in a controlled way. This notion, which dates back to the 1970s, is both general, and in certain cases of interest, very amenable to direct analysis and simulation. In this paper, we will introduce this notion, familiarize ourselves with the basics of this analysis, and conclude by presenting a application of stochastic localization to constructing low-entropy decompositions of probability measures, inspired by information theory.

These is my final paper for 18.676, a Spring 2024 course on stochastic calculus taught by Nike Sun. All mistakes are my own; please reach out to me if you spot anything!

CONTENTS

1. INTRODUCTION: WHAT IS STOCHASTIC LOCALIZATION

Suppose we are given a probability measure μ on a high-dimensional space, like $\mathbf{R}^n$. Oftentimes, in statistical physics or machine learning, these are complicated or otherwise opaque distributions we want to understand better. Stochastic localization, broadly construed, is a controlled way of deforming such measures by localizing them at a point.

Definition 1.1. A *stochastic localization* of μ is a probability measure-valued stochastic process $(\mu_t)_{t\geq 0}$ with $\mu_0 = \mu$ such that

1. $\mu_t \Rightarrow \delta_{\mathbf{x}^*}$ for a (random) $\mathbf{x}^* \in \mathbf{R}^n$, i.e. the process localizes;
2. $(\mu_t)_{t\geq 0}$ is a martingale.

The latter means that $\mu_t(A)$ is a martingale for all measurable $A \subset \mathbf{R}^n$, or equivalently, that $\mathbf{E}_{\mathbf{x}\sim\mu_t}[f(x)]$ is a martingale for all bounded continuous $f \in C_b(\mathbf{R}^n)$.

To motivate this definition, consider the following computation: let (μ_t) a stochastic localization of μ. Weak convergence says for any bounded continuous $f \in C_b(\mathbf{R}^n)$,

$$\int f(\mathbf{x})\,\mu_t(\mathrm{d}\mathbf{x}) \to \int f(\mathbf{x})\,\delta_{\mathbf{x}^*}(\mathrm{d}\mathbf{x}) = f(\mathbf{x}^*),$$

Date: May 14, 2024.

so by bounded convergence, we have

$$\mathbf{E}\left[\int f(\mathbf{x})\,\mu_t(\mathrm{d}\mathbf{x})\right] \to \mathbf{E}[f(\mathbf{x}^*)].$$

However, the martingale condition implies the expectation on the left is constant for all t, and in particular is equal to $\int f(\mathbf{x})\mu(\mathrm{d}\mathbf{x})$, as μ is a deterministic measure. Thus,

$$\int f(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}) = \mathbf{E}[f(\mathbf{x}^*)] \text{ for all } f \in C_b(\mathbf{R}^n).$$

That is, $\mathbf{x}^*$ is distributed as a random sample from μ.

Example 1.2 (Coordinate-by-coordinate localization)**.** Suppose μ is supported on the vertices of the hypercube $\{-1,+1\}^n$. Pick $X \sim \mu$ and, independently, a uniformly random permutation $(k_1, \dots, k_n)$ of $1, \dots, n$. Then, the process

$$\mu_j(\cdot) \coloneqq \mu(X \in \cdot \mid X_{k_1}, \dots, X_{k_j}),$$

i.e. the conditional distribution of X given $X_{k_1}, \dots, X_{k_j}$, is a stochastic localization of μ. It is a martingale by the conditional version of Adam's law, i.e. the tower law for non-Harvard trained readers, and it localizes as for $j \geq n$, we have conditioned on all the information about where X is. Visually, we can imagine at each integer time step bisecting the hypercube along a randomly chosen axis, pushing all the mass to one side or another – eventually, the mass will be localized on a single point. This localization scheme, while simple, has applications towards analysis of Gibbs sampling [CE22].

Example 1.3 (Isotropic Gaussian localization)**.** We return to a general μ; let $\mathbf{x} \sim \mu$ and B_t be an independent n-dimensional Brownian motion started from 0. Define

$$\mathbf{y}_t \coloneqq t\mathbf{x} + B_t$$

and let

$$\mu_t(\cdot) \coloneqq \mu(\mathbf{x} \in \cdot \mid \mathbf{y}_t).$$

This is also a stochastic localization. It is a martingale for the same reason as above, and as t, which we can interpret as a signal-to-noise ratio, increases, the relative contribution of the Gaussian noise decreases, so $\mathbf{y}_t \approx t\mathbf{x}$ [Mon23, §1.3].

Why might one care about stochastic localization? There are two (overlapping) motivations I find worth considering:

1. (High-dimensional geometry) Within the context of convex geometry, there is a notion of "localization," which reduces high-dimensional inequalities to 1-dimensional ones to prove statements about isoperimetric or concentration phenomena. Very roughly, the idea is to pick a line, or "needle" running through a given convex body and project along it, and many approaches towards milestone conjectures in this area, such as the Kannan-Lovász-Simonovitz conjecture, rely on being able to show "most" needles have good isoperimetry [LV18, §3.3]. One use of stochastic localization is to globalize this process, localizing a given measure along a randomly chosen needle, and establishing isoperimetric inequalities by showing that for small times, the isoperimetry along this process improves without shrinking the variance too much [Eld13].

2. (Sampling algorithms) As we saw, stochastic localization lets us push all the mass of a complicated measure onto a random sample from that measure. Thus, if we can efficiently simulate a stochastic localization scheme, we get a sampling algorithm for an otherwise intractible sampling problem. As we'll see, the success of such methods depend on the specific choice of localization scheme, but this idea has already been used to great effect in [EMS24].

We'll see soon that both of these applications have information-theoretic motivations; one might interpret certain stochastic localization schemes as a way of decomposing a measure into a simple mixture of simpler measures.

The remainder of this paper has two goals. First, we will discuss the abstract theory of stochastic localizations in more detail, unpacking the historical connection between localizations and "non-linear filters." We will also expand Example 1.3 to a natural and broadly applicable family of localization schemes, characterizable by an explicit SDE. Second, we will return to the question of low-entropy decompositions, and, following [AM21], prove a theorem of Eldan on the behavior of a stochastic localization-inspired decomposition [Eld19].

2. Observation Processes and the Localization SDE

To start, notice how in Examples 1.2 and 1.3, our stochastic localizations were constructed by sampling $X \sim \mu$ and conditioning on a sequence of random variables that grew more "informative" about X as $t \to \infty$. This idea of localizing a measure by conditioning on more data has natural information-theoretic interpretations, which is formalized by the notion of observation processes [Mon23, §3].

In this section, we introduce observation processes, and then discuss the linear-tilt localization SDE, which provides both a theoretical guarantee that certain classes of localizations exist and are tractable as well as an alternative (and often more practical) method of simulating them. This is meant to be an overview of what (in the author's opinion) are the key fundamentals behind stochastic localization in the literature, and readers familiar with the idea should instead skip ahead to section Section 3.

2.1. Observation Processes.

Definition 2.1. Given $\mathbf{x} \sim \mu$, a random process $(\mathbf{y}_t)_{t\geq 0}$ is an *observation process* for $\mathbf{x}$ if

1. for each sequence $t_1 < \cdots < t_k$, $\mathbf{x}, \mathbf{y}_{t_k}, \dots, \mathbf{y}_{t_1}$ forms a Markov chain, i.e. the process gets more informative over time;
2. given a measurable $A \subset \mathbf{R}^n$,
$$\mu_\infty(A) \coloneqq \mu(\mathbf{x} \in A \mid \mathbf{y}_t,\ t \geq 0) \in \{0, 1\},$$
i.e. the process gives complete information about $\mathbf{x}$.

The *associated stochastic localization scheme* is given by
$$\mu_t(\cdot) \coloneqq \mu(\mathbf{x} \in \cdot \mid \mathbf{y}_t).$$

One way of interpreting the first condition in Definition 2.1 is to realize that asking
$$\mathbf{P}(\mathbf{y}_{t_i} \in \cdot \mid \mathbf{y}_{t_{i+1}}, \dots, \mathbf{y}_k, \mathbf{x}) = \mathbf{P}(\mathbf{y}_{t_i} \in \cdot \mid \mathbf{y}_{t_{i+1}})$$
means requiring y_{t_i} to contain no additional information about $\mathbf{x}$ other than what was known from $\mathbf{y}_{t_{i+1}}$. In particular,
$$\mu(\mathbf{x} \in \cdot \mid \mathbf{y}_{t_k}, \dots, \mathbf{y}_{t_1}) = \mu(\mathbf{x} \in \cdot \mid \mathbf{y}_{t_k}) = \mu_{t_k}(\cdot),$$

by Bayes' rule. This notion of "ordering by physical degradation" has existed for some time in the context of information theory, and some of the earliest work on what we now call stochastic localization came from considering observation processes for non-linear filtering problems [Ber73; FKK72]. Of course, the second condition ensures that the localization scheme thus constructed fulfils condition 1. of Definition 1.1.

Remark 2.2. In [CE22], the authors consider stochastic localization schemes given by sampling $\mathbf{x} \sim \mu$ and conditioning on a filtration $\mathcal{F}_t$ which is *precise*, in that $\bigcup \mathcal{F}_t = \sigma(\mathbf{x})$. These are referred to as "Doob localization schemes," given their similarity to the more familiar notion of Doob martingales (in fact, such a process is a martingale for precisely the same reason a Doob martingale is a martingale). In fact, under mild assumptions, it is always possible to write a stochastic localization scheme as a Doob localization; even better, it is possible to write it as the localization scheme associated to an honest observation process, although said process might not be valued on the same space as $\mathbf{x}$. See [Mon23, Remark 3.2] for more details.

Now we will generalize Example 1.3 to a broad class of localizations driven by Gaussian noise with increasing signal-to-noise ratio. For the remainder of the section, let $\mathbf{Q}$ be a positive (semi)definite $n \times n$ matrix $\mathbf{Q}$. Any such matrix induces a positive (semi)definite bilinear form and (semi)norm given by

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{Q}} := \langle \mathbf{x}, \mathbf{Q}\mathbf{y} \rangle, \qquad \text{and } \|\mathbf{x}\|_{\mathbf{Q}}^2 := \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{Q}}.$$

Definition 2.3. The *anisotropic Gaussian localization* of μ is the stochastic localization scheme associated to the observation process

$$\mathbf{y}_t = t\mathbf{x} + \mathbf{Q}^{1/2} B_t,$$

where $\mathbf{x} \sim \mu$ and B_t is an independent standard Brownian motion. That is,

$$\mu_t(\cdot) := \mu(\mathbf{x} \in \cdot \mid \mathbf{y}_t).$$

Anisotropic, in this context, means having non-identity covariance.

As the conditional density of $\mathbf{y}_t$ given $\mathbf{x}$ is proportional to

$$\exp\Big(-\frac{1}{2t}(\mathbf{y}_t - t\mathbf{x})^\top \mathbf{Q}^{-1}(\mathbf{y}_t - t\mathbf{x})\Big) = \exp\Big(-\frac{1}{2t}\|\mathbf{y}_t - t\mathbf{x}\|_{\mathbf{Q}^{-1}}^2\Big),$$

and

$$\|\mathbf{y}_t - t\mathbf{x}\|_{\mathbf{Q}^{-1}}^2 = \|\mathbf{y}_t\|_{\mathbf{Q}^{-1}}^2 - 2t\langle \mathbf{y}_t, \mathbf{x} \rangle_{\mathbf{Q}^{-1}} + t^2 \|\mathbf{x}\|_{\mathbf{Q}^{-1}}^2,$$

a quick application of Bayes' rule shows that (knowing $\mu_t \ll \mu$), we have

$$\mu_t(\mathrm{d}\mathbf{x}) = \frac{1}{Z_t} \exp\Big(\langle \mathbf{y}_t, \mathbf{x} \rangle_{\mathbf{Q}^{-1}} - \frac{t}{2}\|\mathbf{x}\|_{\mathbf{Q}^{-1}}^2\Big)\mu(\mathrm{d}\mathbf{x}), \tag{1}$$

where we absorb all terms independent of $\mathbf{x}$ into the normalizing constant Z_t. This computation is why some authors, e.g. [CE22], call this scheme the "linear-tilt localization"; here we evolve our observation process in a way that increasingly reveals more about the originally sampled $\mathbf{x}$, and exponentially tilt our measure in this random direction.

2.2. **The Likelihood SDE.** The stochastic localization process in Definition 2.3 is both general and highly applicable, as we'll see in the next section. However, there is a problem: recall that one of our motivations was in constructing methods of sampling from μ but how might we sample $\mathbf{y}_t$? It seems like we would need to already sample $\mathbf{x} \sim \mu$, which is problematic.

Luckily, we can circumvent this issue if we can characterize μ_t in a way that doesn't directly reference $\mathbf{y}_t$. Define the *likelihood ratio process*

$$L_t(\mathbf{x}) \coloneqq \frac{\mu_t(\mathrm{d}\mathbf{x})}{\mu(\mathrm{d}\mathbf{x})} = \frac{1}{Z_t} \exp\Big(\langle \mathbf{y}_t, \mathbf{x}\rangle_{\mathbf{Q}^{-1}} - \frac{t}{2}\|\mathbf{x}\|^2_{\mathbf{Q}^{-1}}\Big).$$

Moreover, let $\mathbf{a}_t = \int \mathbf{x}\, \mu_t(\mathrm{d}\mathbf{x})$ denote the conditional expectation $\mathbf{E}[\mathbf{x} \mid \mathbf{y}_t]$.

Theorem 2.4. *There exists a Brownian motion $(W_t)_{t\geq 0}$ adapted to the filtration generated by $\mathbf{y}_t$ such that for all $\mathbf{x} \in \mathbf{R}^n$ and $t \geq 0$, we have*

$$\mathrm{d}L_t(\mathbf{x}) = L_t\langle \mathbf{x} - \mathbf{a}_t, \mathbf{Q}^{-1/2}\, \mathrm{d}W_t\rangle,$$

and $L_0(\mathbf{x}) = 1$.

This is great because it gives us a self-contained way to generate or analyze a linear-tilt localization, as in Definition 2.3, without worrying about conditional distributions or having to sample from μ. We still need to compute conditional expectations, but this is markedly easier in many applications.

To prove Theorem 2.4, we will need the following classical lemma on Itô diffusions with random drift:

Lemma 2.5. *Let $\mathbf{y}_t$ be an Itô process with $\mathbf{y}_0 = 0$ and differential $\mathrm{d}\mathbf{y}_t = \mathbf{x}\,\mathrm{d}t + \mathbf{Q}^{1/2}\,\mathrm{d}B_t$, and let $\mathcal{F}_t = \{\mathbf{y}_s : 0 \leq s \leq t\}$. Assume that μ has finite second moment, and let $\alpha_t \coloneqq \mathbf{E}[\mathbf{x} \mid \mathcal{F}_t]$. Then*

$$W_t \coloneqq \mathbf{Q}^{-1/2}\left(\mathbf{y}_t - \int_0^t \alpha_s\, \mathrm{d}s\right)$$

is a $\mathcal{F}_t$-adapted Brownian motion.

Proof. We have

$$W_t = B_t + \mathbf{Q}^{-1/2}\int_0^t (\mathbf{x} - \alpha_s)\, \mathrm{d}s,$$

i.e. $\mathrm{d}W_t = \mathrm{d}B_t + \mathbf{Q}^{-1/2}(\mathbf{x} - \alpha_t)\,\mathrm{d}t$, so $\langle W\rangle_t = t\mathbf{I}_n$. Thus, by Itô's lemma, we have for $0 \leq s \leq t$,

$$\begin{aligned} \mathrm{d}e^{iz(W_t - W_s)} &= ize^{iz(W_t-W_s)}\,\mathrm{d}W_t - \frac{z^2}{2}e^{iz(W_t-W_s)}\,\mathrm{d}t \\ \implies e^{iz(W_t-W_s)} &= 1 + iz\int_s^t e^{iz(W_u-W_s)}\,\mathrm{d}B_u \\ &\quad + iz\int_s^t e^{iz(W_u-W_s)}\mathbf{Q}^{-1/2}(\mathbf{x} - \alpha_u)\,\mathrm{d}u - \frac{z^2}{2}\int_s^t e^{iz(W_u-W_s)}\,\mathrm{d}u \end{aligned} \tag{2}$$

Our moment bounds imply $\mathbf{E}\big[\int_s^t e^{iz(W_u-W_s)}\,\mathrm{d}B_u \mid \mathcal{F}_s\big] = 0$ (as it is a martingale) and

$$\mathbf{E}\left[\int_s^t e^{iz(W_u-W_s)}\mathbf{Q}^{-1/2}(\mathbf{x} - \alpha_u)\,\mathrm{d}u \mid \mathcal{F}_s\right] = \mathbf{E}\left[\int_s^t e^{iz(W_u-W_s)}\mathbf{Q}^{-1/2}\,\mathbf{E}[\mathbf{x} - \alpha_u \mid \mathcal{F}_u]\,\mathrm{d}u \mid \mathcal{F}_s\right] = 0$$

by the tower law. Thus, taking conditional expectations of Eq. (2) gives

$$\mathbf{E}\big[e^{iz(W_t-W_s)} \mid \mathcal{F}_s\big] = 1 - \frac{z^2}{2}\int_s^t \mathbf{E}\big[e^{iz(W_u-W_s)} \mid \mathcal{F}_s\big]\,\mathrm{d}u.$$

Solving this differential equation gives

$$\mathbf{E}\big[e^{iz(W_t-W_s)} \mid \mathcal{F}_s\big] = e^{-\frac{z^2}{2}(t-s)},$$

so (as in the proof of the Lévy characterization) W_t is a Brownian motion, as claimed (adapted from [LS77, Theorem 7.12]). //

One way to interpret this lemma is to realize that, given such a $\mathbf{y}_t$, we already have the right quadratic covariation for W_t. Thus, by the Lévy characterization, we only need to check that this process is still a martingale. However, the only FV term that could cause problems is given by integrating a term which has conditional expectation 0, essentially by construction!

With this in hand, we can prove Theorem 2.4.

Proof of Theorem 2.4. We know

$$\mathrm{d}\log L_t(\mathbf{x}) = \langle \mathrm{d}\mathbf{y}_t, \mathbf{x}\rangle_{\mathbf{Q}^{-1}} - \frac{1}{2}\|\mathbf{x}\|_{\mathbf{Q}^{-1}}^2\,\mathrm{d}t - \mathrm{d}\log Z_t, \tag{3}$$

so we can start by analyzing the last term. Writing $h_t(\mathbf{x}) = \langle \mathbf{y}_t, \mathbf{x}\rangle_{\mathbf{Q}^{-1}} - \frac{t}{2}\|\mathbf{x}\|_{\mathbf{Q}^{-1}}^2$ for the exponential tilting factor, we begin by using Itô's lemma to compute

$$\begin{aligned}
\mathrm{d}Z_t &= \mathrm{d}\left(\int_{\mathbf{R}^n} e^{h_t(\mathbf{x})}\,\mu(\mathrm{d}\mathbf{x})\right)\\
&= \int_{\mathbf{R}^n}\left(\langle \mathrm{d}\mathbf{y}_t, \mathbf{x}\rangle_{\mathbf{Q}^{-1}} - \frac{1}{2}\|\mathbf{x}\|_{\mathbf{Q}^{-1}}^2\,\mathrm{d}t\right)e^{h_t(\mathbf{x})}\,\mu(\mathrm{d}\mathbf{x}) + \frac{1}{2}\left(\int_{\mathbf{R}^n}\|\mathbf{x}\|_{\mathbf{Q}^{-1}}^2 e^{h_t(\mathbf{x})}\,\mu(\mathrm{d}\mathbf{x})\right)\mathrm{d}t,\\
&= \left\langle \mathbf{Q}^{-1}\,\mathrm{d}\mathbf{y}_t, \int_{\mathbf{R}^n}\mathbf{x}e^{h_t(\mathbf{x})}\,\mu(\mathrm{d}\mathbf{x})\right\rangle = Z_t\langle \mathbf{Q}^{-1}\,\mathrm{d}\mathbf{y}_t, \mathbf{a}_t\rangle.
\end{aligned}$$

Note that the quadratic covariation of $\mathbf{y}_t$ is $\mathbf{Q}t$, as $\mathrm{d}\mathbf{y}_t = \mathbf{x}\,\mathrm{d}t + \mathbf{Q}^{1/2}\,\mathrm{d}B_t$. In particular, the quadratic variation $\langle Z\rangle_t$ satisfies

$$\mathrm{d}\langle Z\rangle_t = Z_t^2\|\mathbf{a}_t\|_{\mathbf{Q}^{-1}}^2\,\mathrm{d}t.$$

Thus, another application of Itô's lemma shows

$$\mathrm{d}\log Z_t = \frac{\mathrm{d}Z_t}{Z_t} - \frac{1}{2}\frac{\mathrm{d}\langle Z\rangle_t}{Z_t^2} = \langle \mathbf{a}_t, \mathrm{d}\mathbf{y}_t\rangle_{\mathbf{Q}^{-1}} - \frac{1}{2}\|\mathbf{a}_t\|_{\mathbf{Q}^{-1}}^2\,\mathrm{d}t,$$

and reinserting into Eq. (3) gives

$$\begin{aligned}
\mathrm{d}\log L_t(\mathbf{x}) &= \langle \mathrm{d}\mathbf{y}_t, \mathbf{x}\rangle_{\mathbf{Q}^{-1}} - \frac{1}{2}\|\mathbf{x}\|_{\mathbf{Q}^{-1}}^2\,\mathrm{d}t - \langle \mathbf{a}_t, \mathrm{d}\mathbf{y}_t\rangle_{\mathbf{Q}^{-1}} + \frac{1}{2}\|\mathbf{a}_t\|_{\mathbf{Q}^{-1}}^2\,\mathrm{d}t\\
&= \langle \mathbf{x}-\mathbf{a}_t, \mathrm{d}\mathbf{y}_t\rangle_{\mathbf{Q}^{-1}} - \frac{1}{2}\big(\|\mathbf{x}\|_{\mathbf{Q}^{-1}} - \|\mathbf{a}_t\|_{\mathbf{Q}^{-1}}\big)\,\mathrm{d}t\\
&= \langle \mathbf{x}-\mathbf{a}_t, \mathrm{d}\mathbf{y}_t - \mathbf{a}_t\,\mathrm{d}t\rangle_{\mathbf{Q}^{-1}} - \frac{1}{2}\|\mathbf{x}-\mathbf{a}_t\|_{\mathbf{Q}^{-1}}^2\,\mathrm{d}t.
\end{aligned}$$

Here is where Lemma 2.5 comes in: we observe that as $\mathbf{y}_t$ is a sufficient statistic for $\mathbf{x}$ under μ_t, $\mathbf{a}_t$ as defined here is equal to the process α_t in the lemma; thus we can take the Brownian motion W_t constructed therein, observing it has

$$\mathrm{d}W_t = \mathbf{Q}^{-1/2}(\mathrm{d}\mathbf{y}_t - \mathbf{a}_t\,\mathrm{d}t).$$

In particular, we find that

$$\mathrm{d}\log L_t(\mathbf{x}) = \langle \mathbf{x} - \mathbf{a}_t, \mathbf{Q}^{-1/2}\,\mathrm{d}W_t\rangle - \frac{1}{2}\|\mathbf{x} - \mathbf{a}_t\|^2_{\mathbf{Q}^{-1}}\,\mathrm{d}t.$$

Applying Itô's lemma one final time, we observe $\mathrm{d}\langle \log L(\mathbf{x})\rangle_t = \|\mathbf{x} - \mathbf{a}_t\|^2_{\mathbf{Q}^{-1}}$, and so

$$\mathrm{d}L_t(\mathbf{x}) = L_t\langle \mathbf{x} - \mathbf{a}_t, \mathbf{Q}^{-1/2}\,\mathrm{d}W_t\rangle,$$

thus completing the proof.[1] //

Remark 2.6. The localization SDE in Theorem 2.4 is highly useful, but it is still an infinite-dimensional family of SDEs, so directly simulating it (without working on a subspace of much smaller size, e.g. the unit hypercube as in Example 1.2) might still be inefficient. However, another benefit of Lemma 2.5 and the linear-tilt form of Definition 2.3 is that we can instead simulate $\mathbf{y}_t$ via

$$\mathrm{d}\mathbf{y}_t = \mathbf{a}_t\,\mathrm{d}t + \mathbf{Q}^{1/2}\,\mathrm{d}W_t,$$

with $\mathbf{y}_0 = 0$, and then simulate μ_t by directly computing the tilts of μ, as in Eq. (1). This idea was leveraged to great effect in [EMS24], where the authors use it to efficiently sample from the Sherrington-Kirkpatric Gibbs measure in the high-temperature regime. Alternatively, one can approach the linear-tilt localization from the perspective of defining it as the measure associated to a solution of the localization SDE. This lends itself well to deriving quantitative bounds on the evolution of $\mathrm{Cov}(\mu_t)$, which is an approach more common when applying these ideas to high-dimensional geometry.

3. Low-Entropy Decompositions of Measures

The original motivation for [AM21], and indeed a good fraction of the stochastic localization literature, is in decomposing a high-dimensional μ into a mixture of simpler measures:

$$\mu = \mathbf{E}_\theta\,\mu_\theta = \int_\Theta \mu_\theta\,\rho(\mathrm{d}\theta), \tag{4}$$

where Θ is a parameter space with a probability measure ρ. A bad approach would be to let $\rho = \mu$, $\Theta = \mathbf{R}^n$, and $\mu_\theta = \delta_\theta$: then Eq. (4) would be satisfied, but we would be shifting the problem to ρ, without balancing out the complexity of μ among the μ_θ's. What would be better is for the μ_θ's to be well-controlled (e.g. close to product measures), with ρ having low entropy.

Below, we present a construction of such a decomposition, originally due to [Eld19]. However, it was reinterpreted and reproved in information-theoretic terms by [AM21], who managed to greatly simplify what were otherwise direct computational proofs. To motivate this reinterpretation, we first need to recall some concepts from information theory.

[1] On a personal note, I initially struggled with the applications of Itô's lemma presented here. I encourage the curious reader to work carefully through the computation of $\mathrm{d}Z_t$, as a wonderful exercise on using the multidimensional version of Itô's lemma; it may also help to assume $\mathbf{Q} = \mathbf{I}$

3.1. **A whirlwind tour of information theory.** We recall some notions from information theory which will be key in establishing the low-entropy nature of our putative decomposition.

Definition 3.1. The *differential entropy* of a r.v. X with density f (w.r.t. Lebesgue measure) is

$$h(X) \coloneqq \mathbf{E}_{X\sim f}[-\log f(X)] = -\int f(\mathbf{x}) \log f(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Example 3.2 (Gaussian differential entropy)**.** Let $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, for Σ a symmetric positive-definite $n \times n$ covariance matrix. What is $h(\mathbf{x})$? It is easy to see that the differential entropy is translation invariant, so we can assume $\mu = 0$. Recall the density of $\mathbf{x}$ is

$$\phi(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\Big(-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1}\mathbf{x}\Big).$$

Thus,

$$\log \phi(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x} - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma.$$

Note that the last two terms do not depend on $\mathbf{x}$, so it remains to compute

$$\begin{aligned}\mathbf{E}_\phi\, \mathbf{x}^\top\Sigma^{-1}\mathbf{x} &= \mathbf{E}_\phi \operatorname{tr}(\mathbf{x}^\top\Sigma^{-1}\mathbf{x}) = \mathbf{E}_\phi \operatorname{tr}(\Sigma^{-1}\mathbf{x}\mathbf{x}^\top)\\ &= \operatorname{tr}\big(\Sigma^{-1}\,\mathbf{E}_\phi(\mathbf{x}\mathbf{x}^\top)\big) = \operatorname{tr}(\Sigma^{-1}\Sigma) = \operatorname{tr}(\mathbf{I}_n) = n.\end{aligned}$$

Using that $\log\det\Sigma = \operatorname{tr}(\log\Sigma)$ (both are the sum of the log-eigenvalues of Σ), we conclude

$$h(\mathbf{x}) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\operatorname{tr}(\log\Sigma).$$

Intuitively, entropy corresponds to how much "information" is contained in a r.v., or alternatively, how much the probability density is spread out. We can also ask what the relative entropy between two measures is:

Definition 3.3. Suppose $\mu \ll \nu$. The *Kullback-Leibler divergence* is

$$\mathrm{D}(\mu \parallel \nu) \coloneqq \mathbf{E}_\mu \log \frac{\mu(\mathrm{d}\mathbf{x})}{\nu(\mathrm{d}\mathbf{x})}.$$

The most important quality of the KL divergence is its nonnegativity; indeed, by Jensen's inequality applied to the convex function $x\log x$,

$$\mathrm{D}(\mu \parallel \nu) = \mathbf{E}_\nu\left[\frac{\mu(\mathrm{d}\mathbf{x})}{\nu(\mathrm{d}\mathbf{x})}\log\frac{\mu(\mathrm{d}\mathbf{x})}{\nu(\mathrm{d}\mathbf{x})}\right] \geq \left(\mathbf{E}_\nu \frac{\mu(\mathrm{d}\mathbf{x})}{\nu(\mathrm{d}\mathbf{x})}\right)\log\left(\mathbf{E}_\nu \frac{\mu(\mathrm{d}\mathbf{x})}{\nu(\mathrm{d}\mathbf{x})}\right) = 0,$$

as

$$\mathbf{E}_\nu \frac{\mu(\mathrm{d}\mathbf{x})}{\nu(\mathrm{d}\mathbf{x})} = \int \frac{\mu(\mathrm{d}\mathbf{x})}{\nu(\mathrm{d}\mathbf{x})}\nu(\mathrm{d}\mathbf{x}) = \int \mu(\mathrm{d}\mathbf{x}) = 1.$$

One consequence of this nonnegativity is the amazing fact that the multivariate Gaussian distribution has maximal entropy among all distributions with the same covariance.

Proposition 3.4 (Gaussian has maximal entropy)**.** *Suppose f is any probability density on $\mathbf{R}^n$ with (positive-definite) covariance Σ, and let ϕ be the $\mathcal{N}(0,\Sigma)$ density. Then $h(f) \leq h(g)$.*

Proof. Without loss of generality, we can assume f has mean 0. We can compute the KL divergence between f and ϕ; this is well-defined as $\phi > 0$, so $f\,\mathrm{d}\mathbf{x} \ll \phi\,\mathrm{d}\mathbf{x}$:

$$\mathrm{D}(f \parallel g) = \int f(\mathbf{x})\log\left(\frac{f(\mathbf{x})}{\phi(\mathbf{x})}\right)\mathrm{d}\mathbf{x} = -h(f) - \int f(\mathbf{x})\log\phi(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

To control the latter term, we can mimic the computation of Example 3.2:

$$\begin{aligned} -\int f(\mathbf{x}) \log \phi(\mathbf{x}) \, d\mathbf{x} &= \frac{1}{2} \int f(\mathbf{x}) \mathbf{x}^\top \Sigma \mathbf{x} + \frac{1}{2} \int \log((2\pi)^n \det \Sigma) f(\mathbf{x}) \, d\mathbf{x} \\ &= \frac{1}{2} \mathbf{E}_f[\mathbf{x}^\top \Sigma \mathbf{x}] + \frac{1}{2} \log((2\pi)^n \det \Sigma) = h(\phi), \end{aligned}$$

where we compute $\mathbf{E}_f[\mathbf{x}^\top \Sigma \mathbf{x}] = n$ exactly as before. By nonnegativity of KL divergence, we have $h(\phi) - h(f) \geq 0$, thus completing the proof (adapted from Wikipedia, vis-a-vis [CT06, Theorem 8.6.5]). //

Finally, we want a quantitative way of understanding how far away two r.v.s X, Y are from being independent, i.e. how much information they contain about each other.

Definition 3.5. The *mutual information* of X and Y is

$$I(X;Y) \coloneqq \mathrm{D}(\mu_{X,Y} \parallel \mu_X \times \mu_Y),$$

where μ_X, μ_Y are the marginals of X, Y and $\mu_{X,Y}$ is their joint distribution.

In some cases, it might be that we don't know the joint distribution explicitly, but we do know the marginal and conditional distributions satisfy $\mu_X, \mu_{X|Y}(\cdot \mid \mathbf{y}) \ll \nu_X$, for some reference ν_X.

Proposition 3.6. *In the above situation, we have*

$$I(X;Y) = \mathbf{E}_{\mathbf{y}} \, \mathrm{D}(\mu_{X|Y}(\cdot \mid \mathbf{y}) \parallel \nu_X) - \mathrm{D}(\mu_X \parallel \nu_X).$$

Proof. We can compute

$$\begin{aligned} \mathbf{E}_{\mathbf{y}} \, \mathrm{D}(\mu_{X|Y}(\cdot \mid y) \parallel \nu_X) &= \mathbf{E}_{\mathbf{y}} \int \log \frac{\mu_{X|Y}(d\mathbf{x} \mid \mathbf{y})}{\nu_X(d\mathbf{x})} \mu_{X|Y}(d\mathbf{x} \mid \mathbf{y}) \\ &= \int \log\left(\frac{\mu_{X|Y}(d\mathbf{x} \mid \mathbf{y})}{\nu_X(d\mathbf{x})} \cdot \frac{\mu_Y(d\mathbf{y})}{\mu_Y(d\mathbf{y})} \right) \mu_{X|Y}(d\mathbf{x} \mid \mathbf{y}) \mu_Y(d\mathbf{y}) \\ &= \int \log\left(\frac{\mu_{X,Y}(d\mathbf{x}, d\mathbf{y})}{\nu_X(d\mathbf{x}) \mu_Y(d\mathbf{y})} \right) \mu_{X,Y}(d\mathbf{x}, d\mathbf{y}). \end{aligned}$$

Similarly, we find that

$$\begin{aligned} \mathrm{D}(\mu_X \parallel \nu_X) &= \int \log \frac{\mu_X(d\mathbf{x})}{\nu_X(d\mathbf{x})} \mu_X(d\mathbf{x}) \\ &= \int \left(\int \log \frac{\mu_X(d\mathbf{x})}{\nu_X(d\mathbf{x})} \mu_{Y|X}(d\mathbf{y} \mid \mathbf{x}) \right) \mu_X(d\mathbf{x}) \\ &= \int \log \frac{\mu_X(d\mathbf{x}) \mu_Y(d\mathbf{y})}{\nu_X(d\mathbf{x}) \mu_Y(d\mathbf{y})} \mu_{X,Y}(d\mathbf{x}, d\mathbf{y}). \end{aligned}$$

Subtracting these two terms, using properties of log to cancel the denominators, gives $I(X;Y) = \mathrm{D}(\mu_{X,Y} \parallel \mu_X \times \mu_Y)$, as desired. //

We can also relate mutual information to differential entropy: if $\mathbf{x}$ and $\mathbf{y}$ have densities w.r.t. Lebesgue measure, then it is easy to see that

$$I(\mathbf{x};\mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y}) - h(\mathbf{x},\mathbf{y}),$$

where the latter term is the mutual entropy of the joint distribution. Moreover, the *conditional entropy* of X, Y, with joint density $f(\mathbf{x}, \mathbf{y})$, is defined as

$$h(Y|X) := -\int f(\mathbf{x}, \mathbf{y}) \log f(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} \, d\mathbf{x}.$$

It satisfies $h(Y|X) = h(X, Y) - h(X)$, and has the property that it is the average of the differential entropy of the conditional distribution $Y|X = \mathbf{x}$:

$$h(Y|X) = \int \left(\int -f(\mathbf{y}|\mathbf{x}) \log f(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} \right) f(\mathbf{x}) \, d\mathbf{x} = \mathbf{E}_{\mathbf{x}} \, h(Y|X = \mathbf{x}).$$

Thus, another interpretation of Proposition 3.6 becomes

$$I(\mathbf{x}; \mathbf{y}) = h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x}). \tag{5}$$

3.2. **The low-entropy decomposition.** We finally come to our desired low-entropy decomposition of μ. Fix a positive semidefinite $n \times n$ matrix $\mathbf{Q}$, let $\mathbf{x} \sim \mu$, and consider the output of our anisotropic Gaussian localization in Definition 2.3 at a uniformly random time $\tau \in \mathrm{Unif}([1, 2])$:

$$\mathbf{y} := \mathbf{y}_\tau = \tau \mathbf{x} + \mathbf{Q}^{1/2} B_\tau.$$

Here $\mathbf{x}, \tau, B_\tau$ are all independent (note that we could have simply sampled $\mathbf{z} \sim \mathcal{N}(0, I_n)$ and considered $\sqrt{\tau}\mathbf{z}$ instead, but this way the notation remains consistent). Let $\theta = (\mathbf{y}, \tau)$ and $\mu_\theta(\cdot) = \mu(\cdot \mid \theta)$, so that $\mathbf{E}_\theta \, \mu_\theta = \mu$, as in Eq. (4) (by the tower law). One way to motivate this decomposition is by considering noisy observations of $\mathbf{x}$ through a Gaussian channel; given said noisy observations, how much variance is there left in $\mathbf{x}$, and how much does the our observation depend on $\mathbf{x}$ or the noise itself?

Theorem 3.7. *With the setup above, we have*

$$\mathbf{E}_\theta \operatorname{Cov}(\mu_\theta) \preceq \mathbf{Q}; \tag{6}$$

$$0 \le I(\theta; \mathbf{x}) \le \frac{1}{2} \log \det\left(\mathbf{I}_n + 2\mathbf{Q}^{-1} \operatorname{Cov}(\mu)\right); \tag{7}$$

$$\mathbf{E}_\theta\left[\operatorname{Cov}(\mu_\theta)\mathbf{Q}^{-1} \operatorname{Cov}(\mu_\theta)\right] \preceq \operatorname{Cov}(\mu). \tag{8}$$

Intuitively, Eqs. (6) and (8) control the covariance of the component measures μ_θ of the decomposition, while Eq. (7) controls the overall entropy of the mixture. That is, the lower $I(\theta; \mathbf{x})$ is, the more "independent" $\mathbf{x}$ and the decomposition parameters are, avoiding the maximum entropy situation of the pathological decomposition discussed in the intro, where $\theta = \mathbf{x}$ and $I(\mathbf{x}; \mathbf{x}) = \infty$. The idea is now that we can choose $\mathbf{Q}$ to trade off the complexity in μ between the complexity of the component measures and the entropy of the decomposition.

The theorem above is originally due to [Eld19], but a far more informative (no pun intended) proof was developed by [AM21], which we recount in part here. We start with Eq. (6).

Proof of Eq. (6). Imagine we are in the estimation-theoretic setting, trying to estimate the value of a parameter $\mathbf{x}$ from noisy observations $\theta = (\mathbf{y}, \tau)$.[2] In particular, let us compare the estimators

$$\hat{\mathbf{x}}_{\mathrm{MLE}} = \frac{\mathbf{y}}{\tau} \quad \text{and} \quad \hat{\mathbf{x}}_{\mathrm{Bayes}} = \mathbf{E}[\mathbf{x} \mid \theta] = \int \mathbf{x} \mu_\theta(d\mathbf{x}).$$

[2]Any statisticians – myself included – are likely bothered by this notation, but bear with us.

By the multivariate (conditional) bias-variance tradeoff, we know that $\hat{\mathbf{x}}_{\text{Bayes}}$ minimizes the mean-square error $\mathbf{E}[(\mathbf{x}-\hat{\mathbf{x}})^{\otimes 2}]$ among all estimators $\hat{\mathbf{x}}$, in the positive-semidefinite ordering [Jac20]. In particular, we have

$$\mathbf{E}_\theta[\operatorname{Cov}(\mu_\theta)] = \mathbf{E}_\theta\Big[\mathbf{E}_{x\sim\mu_\theta}\Big(\mathbf{x}-\hat{\mathbf{x}}_{\text{Bayes}}^{\otimes 2}\Big)\Big] \preceq \mathbf{E}\Big[\Big(\mathbf{x}-\hat{\mathbf{x}}_{\text{MLE}}^{\otimes 2}\Big)\Big] = \operatorname{Cov}(\mathbf{Q}^{1/2}B_\tau) = \mathbf{E}[\tau^{-1}]\mathbf{Q} \preceq \mathbf{Q},$$

as desired. //

Next, we can utilize the toolkit developed in the previous section to quickly dispatch Eq. (7).

Proof of Eq. (7). We want to bound $I(\mathbf{x};\theta)$ from above and below. The lower bound $I(\mathbf{x};\theta) \geq 0$ was established following Definition 3.3. For the upper bound, consider the joint distribution of $\mathbf{x}, \mathbf{y}$ given $\tau = t$. We recall by Eq. (5) that

$$I(\mathbf{x};\mathbf{y}) = h(\mathbf{y}) - h(\mathbf{y} \mid \mathbf{x}).$$

By Example 3.2 and the fact that $\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(t\mathbf{x}, \mathbf{Q}t)$, we know

$$h(\mathbf{y} \mid \mathbf{x}) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\operatorname{tr}\log \mathbf{Q}t$$

Moreover, as the Gaussian distribution has maximal entropy, Proposition 3.4 gives

$$h(\mathbf{y}) \leq \frac{n}{2}\log(2\pi e) + \frac{1}{2}\operatorname{tr}\log \operatorname{Cov}(\mu_{\mathbf{y}}),$$

so overall, as $\operatorname{Cov}(\mu_{\mathbf{y}}) = t^2 \operatorname{Cov}(\mu) + \mathbf{Q}t$, we get that

$$I(\mathbf{x};\mathbf{y}) \leq \frac{1}{2}\operatorname{tr}\log(\mathbf{Q}t + t^2 \operatorname{Cov}(\mu)) - \frac{1}{2}\operatorname{tr}\log(\mathbf{Q}t) = \frac{1}{2}\log\det(\mathbf{I}_n + t\mathbf{Q}^{-1}\operatorname{Cov}(\mu)).$$

Finally, as $t \leq 2$, we conclude. //

We will omit the proof of Eq. (8), on the grounds that, unlike the previous two proofs, it has no illustrative flavor and is instead computational, being a direct application of Gaussian integration by parts and analysis of the minimum mean square error as in the proof of Eq. (6).

4. Conclusions

Overall, stochastic localization is a technique in the analysis of complicated, high-dimensional probability measures, with longstanding motivations, myriad applications, and well-developed theoretical underpinnings. While the key ideas go back to the 70s, the past decade has seen a resurgence in using stochastic localizations for everything from improved asymptotic isoperimetric coefficients to predicted but previously unrealizable fast sampling algorithms to improved mixing time bounds on Markov chains. As we've seen, while the notion is very general, in specific circumstances we can perform precise analysis using nothing more than basic stochastic process theory. Moreover, as a benefit of the underlying simplicity of these concepts, it is possible to effectively motivate their application, as we did in the case of low-entropy decompositions of measures, and thus construct efficient proofs of the associated quantitative estimates. It is the hope of the author that this and similar techniques will go on to be used to even more dramatic effect in high-dimensional probability theory in the future.

ACKNOWLEDGEMENTS

Many thanks to Prof. Sun for being a wonderful teacher and for making stochastic calculus seem intuitive and familiar. I would also like to thank Prof. Mark Sellke and Prof. Subhabrata Sen, for introducing me to stochastic localization, and my pset partners, Serena An, Enrique Rivera Ferraiuoli, Daniel Ogbe, and Xialu Zheng, for their help throughout the semester!

REFERENCES

[FKK72] Masatoshi Fujisaki, G. Kallianpur, and Hiroshi Kunita. "Stochastic differential equations for the non linear filtering problem". In: *Osaka Journal of Mathematics* 9.1 (1972), pp. 19–40.

[Ber73] P. Bergmans. "Random coding theorem for broadcast channels with degraded components". en. In: *IEEE Transactions on Information Theory* 19.2 (Mar. 1973), pp. 197–207. ISSN: 0018-9448. DOI: 10.1109/TIT.1973.1054980. URL: http://ieeexplore.ieee.org/document/1054980/.

[LS77] R. S. Liptser and A. N. Shiryayev. *Statistics of Random Processes I*. New York, NY: Springer, 1977. ISBN: 9781475716672. DOI: 10.1007/978-1-4757-1665-8. URL: http://link.springer.com/10.1007/978-1-4757-1665-8.

[CT06] T. M. Cover and Joy A. Thomas. *Elements of information theory*. 2nd ed. Hoboken, N.J: Wiley-Interscience, 2006. ISBN: 9780471241959.

[Eld13] Ronen Eldan. "Thin shell implies spectral gap up to polylog via a stochastic localization scheme". In: *Geometric and Functional Analysis* 23.2 (Apr. 2013). arXiv:1203.0893 [math], pp. 532–569. ISSN: 1016-443X, 1420-8970. DOI: 10.1007/s00039-013-0214-y. URL: http://arxiv.org/abs/1203.0893.

[LV18] Yin Tat Lee and Santosh S. Vempala. "The Kannan-Lovász-Simonovits Conjecture". In: arXiv:1807.03465 (July 2018). arXiv:1807.03465 [cs, math]. URL: http://arxiv.org/abs/1807.03465.

[Eld19] Ronen Eldan. "Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation". In: arXiv:1811.11530 (May 2019). arXiv:1811.11530 [math-ph]. URL: http://arxiv.org/abs/1811.11530.

[Jac20] Johannes Jacob Meyer. *The multivariate bias-variance decomposition*. July 2020. URL: https://johannesjakobmeyer.com/blog/005-multivariate-bias-variance-decomposition/.

[AM21] Ahmed El Alaoui and Andrea Montanari. "An Information-Theoretic View of Stochastic Localization". In: arXiv:2109.00709 (Sept. 2021). arXiv:2109.00709 [cs, math]. URL: http://arxiv.org/abs/2109.00709.

[CE22] Yuansi Chen and Ronen Eldan. "Localization Schemes: A Framework for Proving Mixing Bounds for Markov Chains". In: arXiv:2203.04163 (June 2022). arXiv:2203.04163 [math-ph, stat]. URL: http://arxiv.org/abs/2203.04163.

[Mon23] Andrea Montanari. "Sampling, Diffusions, and Stochastic Localization". In: arXiv:2305.10690 (May 2023). arXiv:2305.10690 [cs]. URL: http://arxiv.org/abs/2305.10690.

[EMS24] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. "Sampling from the Sherrington-Kirkpatrick Gibbs measure via algorithmic stochastic localization". In: arXiv:2203.05093 (Feb. 2024). arXiv:2203.05093 [cond-mat]. URL: http://arxiv.org/abs/2203.05093.

Email address: rushil_mallarapu@college.harvard.edu