ON THE USE OF STOCHASTIC LOCALIZATION FOR SAMPLING.

Goal: The message delivered in this lecture will be twofold. For the first part of this tutorial, we will assume the existence of any oracle of desire when it comes to calculating (possibly) hard observable such as transition kernels and conditional expectations. This is done to illustrate the noising technique from stochastic localization which is used to derive sampling algorithms of varying hardness. Hopefully, we will also make connections between simulated annealing and diffusion processes.

The second part of this tutorial will be to show how to provide rigorous sampling error bounds when one must have a sufficiently fine discretization scheme and a high quality approximation of the oracle of desire. An application to $\mathbb{Z}_2$-synchronization will be discussed.

So Let's get started.

Q: What is the overall goal of sampling?

A: Given some distribution μ, perhaps specified from some random data $\mathbf{Y}$, our goal is sample an atom $\mathbf{x}^* \sim \mu$ in some computationally efficient way. Usually, computational efficiency comes from relates to one's ability to compute value such as $\mu(\mathbf{x}_1)/\mu(\mathbf{x}_2)$ or $\nabla \log \mu(\mathbf{x})$ for points $\mathbf{x}_1, \mathbf{x}_2 \in \text{supp}(\mu)$.

1. General Stochastic Localization Sampling [Mon23]

In contrast to the observable from MCMC and diffusion processes, we will take a different approach to sampling that is based on estimating transition kernels. We aim to create a general scheme to be able to sample an arbitrary distribution, μ, supported in $\mathbb{R}^n$.

Mimicking the notion of "noising" and "denoising" from stochastic localization, and the previous talks given over the past few weeks, the following plan seems realistic:

First we will write a noising algorithm in "forward" time,

(1) Draw an Element $\mathbf{x} \sim \mu$ (Somewhat confusing since one would need a sampling algorithm already, but we address this later)
(2) Given interval $I = [0, T]$, "noise" the sample $\mathbf{x}$ continuously (or in discrete intervals of continuous time) in such a way that after a reasonably long runtime t, the law of the noised element is some tractable distribution, say ν. Stated formally, consider an observation (i.e. noising) sequence $(\mathbf{Y}_t)_{t\in I}$ where for each $k \in \mathbb{N}$ and $t_1 < \cdots < t_k \in I$, we have that $\mathbf{x} \longrightarrow \mathbf{Y}_{t_1} \longrightarrow \cdots \longrightarrow \mathbf{Y}_{t_k}$ forms a Markov Chain.

Remark 1.1. Note, as we will see later, the noising process is allowed to expand beyond the support of μ, contrasting MCMC methods.

We can now implement our stocastic localization algorithm in the backwards direction ("denoising"),

(1) Sample from $\mathbf{Y}_T \sim \nu$, the tractable fully noised distribution. Often times this is simply an i.i.d random vector.
(2) Iteratively sample from the reverse posterior distribution, i.e.
$$\mathbb{P}_{t-\epsilon,t}(\cdot) = \mathbb{P}(\mathbf{Y}_{t-\epsilon} \in \cdot | \mathbf{Y}_t),$$
where this distribution is derived by the noising procedure given above.

Under this definition, one has essentially abstracted away alot of the difficulty of implementing this algorithm (i.e. the need to discretize reverse time and inaccuracies in estimating the transition kernel $\mathbb{P}_{t,t'}$ for specific

Date: November 2024.
Manuscript Written and Typeset By Max Lovig.

values of $t, t' \in I$). Regardless, this idea should hopefully give us a valid sampling algorithm under the final assumption that, given the continuous path of all noised elements $(\mathbf{Y}_t)_{t\in I}$, we have

$$\mathbb{P}(\mathbf{x} \in \mathcal{A} | (\mathbf{Y}_t)_{t\in I}) \in \{0, 1\}.$$

So why does this even work?

From previous talks, if we have a sequence of random measures $\mu_0, \dots, \mu_T$ (with T possibly being infinite and this sequence of measures varying continuously) where $\mu_0 = \mu$ (the target measure), $\mu_T = \mathbb{1}_{x^* \in \mathcal{A}}$ and $\mathbb{E}[\mu_t | \mu_s] = \mu_s$ (for all $s < t$, i.e. a martingale) for $s \leq t$, then we have that $\mathbb{E}[\mu_T] = \mu_0 = \mu$, as μ_T is a Dirac measure then the final step of our stochastic localization process is a sample from μ, which is trivial.

To construct this sequence of measure we consider a Markov chain,

$$\mathbf{x}^* \longrightarrow \mathbf{Y}_1 \longrightarrow \mathbf{Y}_2 \longrightarrow \cdots \longrightarrow \mathbf{Y}_T,$$

where $\mathbf{x}^* \sim \mu$. We can then define the Doob martingale of $\mu_t = \mathbb{P}(x^* \in \cdot | \mathbf{Y}_t, \dots, \mathbf{Y}_T) = \mathbb{P}(\mathbf{x}^* \in \cdot | \mathbf{Y}_t)$. As the noising process is localized once the whole sequence of $\mathbf{Y}$'s are revealed, we have that $\mu_T = \mathbb{1}_{x^* \in \mathcal{A}}$ and if the noising process forces $\mathbf{Y}_T$ to be independent of x^*, then $\mu_T = \mathbb{P}(x^* \in \cdot | \mathbf{Y}_T) = \mathbb{P}(\mathbf{x}^* \in \cdot) = \mu$. We further have, by definition, that $\mu_0 = \mu$ as it is a forced localization.

So, in a sense, this setup gets the localization process backwards, so we follow the denoting process, as the forward process is Markovian, this is also a Markov chain, thus we need to just estimate $\mathbb{P}(x^* \in \cdot | \mathbf{Y}_t, \dots, \mathbf{Y}_T) = \mathbb{P}(\mathbf{Y}_{t-1} \in \cdot | \mathbf{Y}_t)$ and simply daisy chain these transition kernels together.

Now lets get a taste of how to derive SL algorithms using the noising-denoising process. We will present three different model / noising algorithms, to give some basic commentary on how our noising process constructs our SL algorithm. Our three examples, defined by how they noise $\mathbf{x}^*$, are:

- The Eraser Noiser
- A "Simulated Annealing-lite" Binary Switch Noiser
- The Isotropic Gaussian Noiser

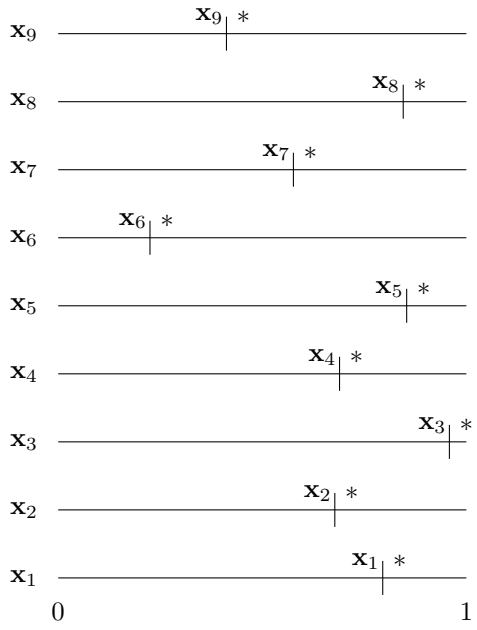

FIGURE 1. The Eraser Noiser

1.1. **The Eraser Noiser.** The erasure noiser, noises a sample $\mathbf{x} \sim \mu$, with $\mathbf{x} \in \mathbb{R}^n$ in the following manner:

First we generate n random variables $\{t_1, \dots, t_n\} \overset{\text{i.i.d}}{\sim} \text{Unif}([0, 1])$ as displayed by the vertical lines in Figure 1. As time t moves forward, when $t > t_i$ we erase the i-th element of $\mathbf{x}$.

The related denoising process is very straightforward and is equivalent to a sequential sampling procedure,

(1) First, one random generates an order for which the elements of $\mathbf{x}$ we erased in, for simplicity lets choose the most natural ordering, numerical.
(2) We then sample $\mathbf{x}_j$ from the conditional law $\mu(\mathbf{x}_j \mid \mathbf{x}_i \text{ for all } i \in [j-1])$.

If we consider a distribution supported on the hyper-cube, which is $\{-1, +1\}$ valued, then we can see that the above sample procedure in step two can be reduced to calculating the conditional mean,

$$\mathbb{E}[\mathbf{x}_j \mid \mathbf{x}_i \text{ for all } i \in [j-1]]. \tag{1.1}$$

Equation (1.1) represents on an overall trend we will see in stochastic localization algorithms, a reduction from sampling to estimation of a conditional expectation. This immediately draws comparisons with Diffusion which reduces sampling to estimation of the score function. When we see

the isotropic Gaussian noiser, an application of Tweedie's formula will codify the relationship between these two techniques.

Remark 1.2. Another interesting fact, although our noising process is continuous in time, there are simple discretization of the denoising process which will never need to be further refine, essentially removing the concern of time discretization.

1.2. **The Binary Switch Noiser.** Now we will consider the binary switch noiser, this presentation differs from the localization algorithm in [Mon23] in that we fix the flipping Rademacher variable to have probability $1/2$ instead of probability $(1+t)/2$, this is just to ellicit a simpler reverse process.

The running informal description of the noising process is as follows:

(For this example we will noise backwards in time to maintain the notation of [Mon23])

(1) Draw $\mathbf{x} \sim \mu$ supported on $\{-1,1\}^n$.
(2) At $t = 1$ we set $\mathbf{Z}_1 = \mathbf{1}$, let $\mathbf{Y}_1 = \mathbf{x} \odot \mathbf{Z}_1$.
(3) Focusing on one coordinate, for each time interval $[t-\delta, t)$, we flip a coin with weight $\delta/t + o(\delta)$, if the coin is heads we set $\mathbf{Z}_{t,i} = \text{Bern}(1/2)$, now let $\mathbf{Y}_t = \mathbf{x} \odot \mathbf{Z}_t$.

If we preform this over each δ small intervals then our vector $\mathbf{Y}_0$ is an element-wise Rademacher $1/2$ random variable. Meaning, our forward process for sampling from μ is as follows:

For each δ step, uniformly sample an order of the variables, to update and then flip the value of each coordinate with probability $\mathbb{P}(\mathbf{Y}_{t+\delta} = \mathbf{y}_{\text{flip}} | \mathbf{Y}_t = \mathbf{y}) = p_i(\mathbf{y}, t)\delta + o(\delta)$, where

$$p_i(\mathbf{y}, t) = \frac{(1-t)}{2t} - \mathbf{y}_i \frac{\mathbb{E}[\mathbf{x}_i | \mathbf{Y}_t = \mathbf{y}] + 1}{2}.$$

This again reduces the problem of sampling to estimating the conditional expectation $\mathbb{E}[\mathbf{x}_i | \mathbf{Y}_t = \mathbf{y}]$.

Remark 1.3. If we think about the reverse process here when t is very small, p_i is dominated by the first term, which just says that one should flip irrespective of the alignment of y with the conditional mean. Once t approaches 1, then the second term dominates, and we see that the alignment of y with the conditional mean determines the likelihood of the flip. This effectively raises the temperature of the walk as time moves forward, akin to simulated annealing, but in a more abstract way.

[Mon23] istead considered the variable $\mathbf{Z}_{t,i}$ to instead be a Rademacher variable with probability $(2-t)/2$ of resulting in $\{1\}$, this is a more natural choice, as it leads to the following equivalency in the noising process. Let $(\mathbf{X}_s)_{s>0}$, with $\mathbf{X}_0 = \mathbf{x} \sim \mu$, be a continuous random walk on the $\{-1,1\}^n$ hypercube, in each interval $[s, s+\delta)$ we flip coordinate i with probability $1/2$, we then have the equivalence of $\mathbf{Y}_t = \mathbf{X}_{\log(1/t)}$ for $t \in (0,1]$.

1.3. **The Isotropic Gaussian Noiser.** The final example is the process most associated with stochastic localization, the isotropic Gaussian noiser.

For simplicity, we will first describe the denoising process without reference to the noising process. Consider an $x^* \sim \mu$, we consider the Gaussian process,

$$\mathbf{Y}_t = t\mathbf{x}^* + \mathbf{W}_t \tag{1.2}$$

where $\mathbf{W}_t$ is Brownian motion. It can be easily shown that this process is Markovian due to Markov properties of Brownian motion, moreover as $t \longrightarrow \infty$ we have that the measure must localize to a Dirac at x^* (just divide by t!). Of course this process is intractable to run because it relies on the unknown sample $\mathbf{x}^*$, but there is an alternate form of the stochastic equation (1.2) that is feasible to run if we have access to a specific conditional expectation oracle.

We can equivalently write the change in $\mathbf{Y}_t$ as follows:

$$d\mathbf{Y}_t = \mathbb{E}[\mathbf{x} \mid t\mathbf{x} + \sqrt{t}\mathbf{G} = \mathbf{y}]dt + d\mathbf{W}_t. \tag{1.3}$$

As with previous techniques the relation between stochastic localization and diffusion is close-knit, for this case our denoiser can be directly related to score estimation through Tweedie's formula.

Theorem 1.4 (Tweedie's Formula)**.** *Suppose that $x^* \sim \mu$, and we have $\mathbf{z} \sim \mathcal{N}(\mathbf{x}^*, \sigma^2)$, then $\mathbb{E}[\mathbf{x}^* \mid \mathbf{z}] = \mathbf{z} + \sigma^2 \nabla_z \log f(z)$ where $f = \mathcal{N}(0, \sigma^2) * \mu$.*

As $\mathbf{Y}_t \sim \mathcal{N}(x^*, t)$, we have that,

$$\mathbb{E}[\mathbf{x}^* \mid \mathbf{Y}_t] = \mathbf{Y}_t + t\nabla_y \log f(y)$$

where $f = \mathcal{N}(0, t) * \mu$. So we could equivalently run the stochastic process,

$$d\mathbf{Y}_t = (\mathbf{Y}_t + t\nabla_y \log f(y))dt + d\mathbf{W}_t.$$

So, what is the noising process related to this sampling algorithm, it is simply the reverse time SDE of (1.2) initialized at a sample $x^* \sim \mu$.

2. Rigourus Stochastic Localization Bounds And The Spiked Wigner Model[MW24]

Now we move onto the second part of this lecture, providing rigorous sampling bound under discretization and oracle approximation.

Now we can move onto a rigorous analysis of sampling from a general distribution μ using the Isotropic Gaussian SL algorithm.

Per the previous section, our goal is to run SDE (1.3) to sample from a measure μ, perhaps now under some set of observations $\mathcal{D}$. Unfortunately we now lack access to the oracle

$$\mathbb{E}[\mathbf{x} \mid \mathcal{D}, t\boldsymbol{\theta} + \sqrt{t}\mathbf{G} = \mathbf{y}], \tag{2.1}$$

and have to deal with discretization errors of approximating the continuous denoising process.

The below Theorem, from [MW24], provides a rigorous bound on the error between μ and sample drawn from the SL algorithm in terms of Wasserstein distance. This said algorithm takes the following from:

Given an oracle approximation $\hat{m} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, which takes in the time in the denoising process and an estimate for $\mathbf{y}$, returning an estimate for the conditional expectation (2.1). Under a discretization with step size Δ, we then run the following,

(1) Set $\hat{\mathbf{y}} = 0$.
(2) Draw a random variable $w \sim \mathcal{N}(0, \mathrm{Id}_n)$
(3) Update $\hat{\mathbf{y}} = \hat{\mathbf{y}} + \Delta\hat{m}(\hat{\mathbf{y}}, \Delta(\text{numSteps})) + \sqrt{\Delta}w$.
(4) Repeat steps 2 and 3 until some desired stopping time.

In order to bound the error between μ and the law of samples from the above algorithm, we require the following conditions to hold.

For transparency, we paste the full set of assumptions below from [MW24], in this work they consider a anisotropic extension to the Isotropic Gaussian SL algorithm where the run the SL algorithm using,

$$\mathbf{y}(t) = tH\boldsymbol{\theta} + \mathbf{W}_t$$

where $H \in \mathbb{R}^{m \times n}$ and$\mathbf{w}_t$ is Brownian motion in $\mathbb{R}^m$. This still has a SDE reduction, with $\mathbf{y}_0 = 0$, to running:

$$d\mathbf{y}(t) = m(\mathbf{y}(t), t) + d\mathbf{W}_t,$$

where, with $\mathbf{G} \sim \mathcal{N}(0, \mathrm{Id})$, $m(\mathbf{y}(t), t) = \mathbb{E}[H\boldsymbol{\theta} | t\mathbf{H}\boldsymbol{\theta} + \sqrt{t}\mathbf{G} = \mathbf{y}(t)]$.

There is then a final step of transferring the estimation of $\mathbf{H}\boldsymbol{\theta}$ to just θ which requires another approximator m_θ, but our application is conducted with $\mathbf{H} = \mathrm{Id}$ and thus $m = m_\theta$.

Assumptions [MW24]

(probabilities below are with respect to the distribution of the stochastic localization process $\{y(t)\}$, at fixed D):

(A1) (Posterior mean consistency) With probability at least $1-\eta$, it holds that

$$\max_{\ell\in\{0,\dots,L-1\}} \frac{1}{\sqrt{N}}\|m(y(\ell\Delta),\ell\Delta)-\hat{m}(y(\ell\Delta),\ell\Delta)\|_2 \le \epsilon_1.$$

Further, with the same probability, $\|m_\theta(y(T),T)-\hat{m}_\theta(y(T),T)\|_2 \le \epsilon_1\sqrt{n}$, where we recall that $T=L\Delta$.

(A2) (Path regularity) With probability at least $1-\eta$, it holds that

$$\max_{\ell\in\{0,\dots,L-1\}} \sup_{t\in[\ell\Delta,(\ell+1)\Delta]} \frac{1}{\sqrt{N}}\|m(y(t),t)-m(y(\ell\Delta),\ell\Delta)\|_2 \le C_1\sqrt{\Delta}+\epsilon_2.$$

(A3) (Lipschitz continuity) There exists a sequence $\{r_\ell\}_{0\le\ell\le L}\subset\mathbb{R}_+$, such that letting $B(\ell):=\{y\in\mathbb{R}^N : \|y-y(\ell\Delta)\|_2\le r_\ell\sqrt{N}\}$, then the following holds with probability at least $1-\eta$:

$$\max_{\ell\in\{0,\dots,L-1\}} \sup_{\substack{y_1\neq y_2\\ y_1,y_2\in B(\ell)}} \left[\frac{1}{\sqrt{N}}\|\hat{m}(y_1,\ell\Delta)-\hat{m}(y_2,\ell\Delta)\|_2 - \frac{C_2}{\sqrt{N}}\|y_1-y_2\|_2\right] \le \epsilon_3.$$

Further, we assume that $r_\ell > (C_1\sqrt{\Delta}+\epsilon_1+\epsilon_2+\epsilon_3)e^{C_2\ell\Delta/C_2}$ for all $\ell\in\{0,\dots,L\}$. We also assume that with the same probability, $\|m_\theta(y_1,T)-m_\theta(y_2,T)\|_2/\sqrt{n} \le C_2\|y_1-y_2\|_2/\sqrt{N}+\epsilon_3$ for all $y_1,y_2\in B(L)$.

The dependence on constants C_1, C_2 will be tracked in the statement below.

Theorem 1 [MW24]

Theorem 2.1. *Assume that $\|m_\theta(y,T)\|_2 \le R\sqrt{n}$ for all $y\in\mathbb{R}^N$ (this can always be achieved by projection onto the ball $B^n(0,R\sqrt{n})$) and that conditions (A1), (A2) and (A3) hold. Letting $\mu_D^{alg} = Law(\theta^{alg})$ be the distribution of the samples generated by Algorithm 1, then we have:*

$$W_{2,n}(\mu_D,\mu_D^{alg}) \le 2(C_1\sqrt{\Delta}+\epsilon_1+\epsilon_2+\epsilon_3)\cdot e^{C_2T} + \frac{1}{n}\mu_D\left(\|\theta\|_2^2\cdot 1_{\|\theta\|_2\ge R\sqrt{n}}\right)^{1/2} + 10R\eta$$
$$+W_{2,n}\left(\mu_D, Law(m_\theta(y(T),T))\right).$$

If in addition H has full column rank, and $\int(\|\theta\|_2^2/n)^{c_0}\mu_D(d\theta)\le R^{2c_0}$ for some $c_0>1$, then

$$W_{2,n}(\mu_D,\mu_D^{alg}) \le 2(C_1\sqrt{\Delta}+\epsilon_1+\epsilon_2+\epsilon_3)\cdot e^{C_2T} + C(c_0)R\eta^{(c_0-1)/c_0} + \frac{1}{\sqrt{T}}Tr\left((H^\top H)^{-1}\right)^{1/2},$$

where $C(c_0)$ is a positive constant that depends only on c_0.

2.1. **Applying Theorem 2.1 To The Spiked Wigner Model.** First, a quick recap on what this model represents. We have a prior π_Θ which generates the iid vector $\boldsymbol{\theta}\overset{\text{iid}}{\sim}\pi_\Theta$, we are then given the following observation,

$$bX = \frac{\beta}{n}\boldsymbol{\theta}\boldsymbol{\theta}^\top + \mathbf{W} \tag{2.2}$$

where $\mathbf{W}$ is a GOE(n) random matrix.

This is a generalization of the $\mathbb{Z}_2$-synchronization model to a general prior distribution besides Rademacher. A simple calculation gives that the posterior of $\boldsymbol{\theta}$ is given by

$$\mu_{\mathbf{X}}(\boldsymbol{\theta}) \propto \exp\left(\frac{\beta}{2}\langle\boldsymbol{\theta},\mathbf{X}\boldsymbol{\theta}\rangle - \frac{\beta^2}{4n}\|\boldsymbol{\theta}\|_2^4\right)\pi_\Theta(\boldsymbol{\theta})^{\otimes n}$$

This model has been incredibly well studied since its introduction for the asymmetric case by Johnstone in 2001. Spiked Wigner is meant to be a toy model used to study the ability for algorithms that do PCA-like computation.

Although the basis of this model is much older. Specifically if we consider the non-planted version of this model, we can consider the Hamiltonian $H(\boldsymbol{\sigma}) = \boldsymbol{\sigma}^\top \mathbf{W}\boldsymbol{\sigma}$, of the infamous Sherrington-Kirkpatrick model if we restrict $\boldsymbol{\sigma} \in \{+1, -1\}^n$.

We can apply Theorem 2.1 to this model if we can find an oracle m, which can get arbitrary accuracy on the $\sqrt{n}$ order, arbitrarily close to a Lipschitz function, is computationally efficient, and we can consider a fine enough discretization, then we get this statement

$$W_2(\mu, \mu^{\mathrm{SL}}) \leq \epsilon\sqrt{n}$$

for an arbitrary $\epsilon > 0$.

Favorably, the choice of AMP as the approximating oracle, run for a sufficiently long but $O(1)$ run time, will satisfy these conditions under further conditions on the prior π. We won't go into AMP too much, but, it is important to mention that, the isotropic Gaussian SL oracle and approximate is intimately connected. To get a similar stochastic localization argument for any measure one must consider the available computationally efficient algorithms for a given model and then see which SL algorithms can be applicable. For example, AMP has been shown to be suboptimal in the tensor version of (2.2) so a different algorithm must be used and perhaps there is a more amenable choice of SL oracle to approximate.

Remark 2.2. A very subtle point is when this theorem is actually meaningful, this problem is very delicate as the posterior does not concentrate on the true solution in the same manner as a planted clique model would. Meaning that there is a fundamental gap in the Wasserstein distance between the posterior and an estimator of the true signal. This leads to issues for sampling when considering sublinear sparse signals which have all-or-nothing phenomena.

Further Reading

- [BBdBM24] for an analysis on the stages of diffusion where speciation, and memorization occur. Could be relevant to understanding the time dynamics of stochastic localization, although the analysis is non-rigorous.
- [AMS24] for an analysis of the noiseless case where the model is SK. This paper also gives a notion of when SL will fail, related with a phenomenon known as the overlap gap phenomena.
- [MS24] for a nice introduction to the spiked Wigner model and gives analysis using infinite replica symmetric breaking, which is vital to analyze [AMS24]. The rigorization of ∞-RSB was proven in [Tal11].
- [FVRS21] for an introduction to an approximate message passing, which is a vital tool to confirm the assumptions of the general theorem presented today.

References

[AMS24] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization, 2024.

[BBdBM24] Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models, 2024.

[FVRS21] Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A unifying tutorial on approximate message passing, 2021.

[Mon23] Andrea Montanari. Sampling, diffusions, and stochastic localization, 2023.

[MS24] Andrea Montanari and Subhabrata Sen. A friendly tutorial on mean-field spin glass techniques for non-physicists, 2024.

[MW24] Andrea Montanari and Yuchen Wu. Posterior sampling in high dimension via diffusion processes, 2024.

[Tal11] Michel Talagrand. *The Parisi Formula*, pages 349–474. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

Proof of Theorem 2.1

We couple $\{w_\ell\}_{\ell \le L}$ and $\{B(t)\}_{0 \le t \le T}$ by letting $w_\ell = B(\ell\Delta) - B((\ell-1)\Delta)$. We also define

$$A_\ell = \|y_\ell - y(\ell\Delta)\|_2/\sqrt{N} \text{ for all } \ell \in \{0\} \cup [L], \text{ and write } t_\ell := \ell\Delta.$$

Let Ω be the intersection of the events listed in points (A1), (A2), and (A3). By taking a union bound, we obtain that $\mathbb{P}(\Omega) \ge 1 - 5\eta$. We will prove by induction that, on Ω, the following holds for all $\ell \le L$:

$$\hat{y}_\ell \in B(\ell), \quad \text{and} \quad A_\ell \le \frac{C_1\sqrt{\Delta} + \epsilon_1 + \epsilon_2 + \epsilon_3}{C_2}\left(e^{C_2\ell\Delta} - 1\right). \tag{52}$$

By definition, we see that $A_0 = 0$ and $\hat{y}_0 = y(0) = 0 \in B(0)$. Next, assume that the induction hypothesis holds up to step $\ell - 1$. On the event Ω:

$$\begin{aligned}
A_\ell - A_{\ell-1} &\le \frac{1}{\sqrt{N}} \int_{t_{\ell-1}}^{t_\ell} \|\hat{m}(\hat{y}_{\ell-1}, t_{\ell-1}) - m(y(t), t)\|_2 \, dt \\
&\le \frac{\Delta}{\sqrt{N}} \|\hat{m}(y(t_{\ell-1}), t_{\ell-1}) - m(y(t_{\ell-1}), t_{\ell-1})\|_2 \\
&+ \sup_{t \in [t_{\ell-1}, t_\ell]} \frac{\Delta}{\sqrt{N}} \|m(y(t), t) - m(y(t_{\ell-1}), t_{\ell-1})\|_2 \\
&+ \frac{\Delta}{\sqrt{N}} \|\hat{m}(y(t_{\ell-1}), t_{\ell-1}) - \hat{m}(\hat{y}_{\ell-1}, t_{\ell-1})\|_2 \\
&\le \Delta \cdot \left(\epsilon_1 + C_1\sqrt{\Delta} + \epsilon_2 + C_2 A_{\ell-1} + \epsilon_3\right).
\end{aligned}$$

Substituting in the induction hypothesis, we obtain $A_\ell \le \frac{C_1\sqrt{\Delta} + \epsilon_1 + \epsilon_2 + \epsilon_3}{C_2}\left(e^{C_2\ell\Delta} - 1\right)$ as desired. The claim $\hat{y}_\ell \in B(\ell)$ follows from the stated condition on r_ℓ. This completes the induction proof for Ω.

Applying the bound given in Eq. (52) with $\ell = L$ and using the moment assumptions (A1) and (A3), we have:

$$\begin{aligned}
\frac{1}{\sqrt{n}} \|m_\theta(y(T), T) - \hat{m}_\theta(y_L, L\Delta)\|_2 &\le \frac{1}{\sqrt{n}} \|m_\theta(y(T), T) - m_\theta(y_L, L\Delta)\|_2 \\
&+ \frac{1}{\sqrt{n}} \|\hat{m}_\theta(y(T), T) - \hat{m}_\theta(y_L, L\Delta)\|_2 \\
&\le \epsilon_1 + C_2 A_L + \epsilon_3 \\
&\le \epsilon_1 + \epsilon_3 + (C_1\sqrt{\Delta} + \epsilon_1 + \epsilon_2 + \epsilon_3) \cdot e^{C_2 T} =: \Delta_0. \qquad (53)
\end{aligned}$$

The above upper bound further implies that (denoting by $\mathbb{P}_R$ the projection onto $B^n(0, R\sqrt{n})$):

$$\begin{aligned}
W_{2,n}(\mu_D, \mu_D^{\text{alg}}) &\le W_{2,n}\left(\text{Law}(m_\theta(y(T), T)), \mu_D^{\text{alg}}\right) + W_{2,n}\left(\text{Law}(m_\theta(y(T), T)), \mu_D\right) \qquad (54) \\
&\le \frac{1}{\sqrt{n}} \mathbb{E}\|m_\theta(\hat{y}_L, L\Delta) - m_\theta(y(T), T)\|_2^{1/2} + W_{2,n}\left(\mu_D, \text{Law}(m_\theta(y(T), T))\right) \\
&\le \frac{1}{\sqrt{n}} \left(\mathbb{E}\|\mathbb{P}_R(\hat{m}_\theta(\hat{y}_L, L\Delta)) - \mathbb{P}_R(m_\theta(y(T), T))\|_2^2\right)^{1/2} + 10R\eta + W_{2,n}\left(\mu_D, \text{Law}(m_\theta(y(T), T))\right). \qquad (55)
\end{aligned}$$

This implies Eq. (13). Eq. (14) follows by using the moment assumption to bound the expectation on the right-hand side and optimizing over R, and finally applying Lemma B.1.