

REFERENCES

- 1] "Provably Robust Score based Diffusion Posterior Sampling for Plug-and-Play image reconstruction"
- Xinyu Xu, Yujie Chi
- 2] "Diffusion Posterior Sampling is computationally intractable"
- Shyam Gupta, Ajil Jalal, Aditya Pamalekar, Eric Price, Zhiyang Xun
- 3] "Provable benefit of Annealed Langevin Monte Carlo for non-log-concave sampling"
- Wei Gu, Molei Tao, Yongxin Chen.
- 4] "Structured log-concave sampling with a restricted Gaussian oracle"
- Yin Tat Lee, Ruqi Shen, Kevin Tian
- 5] "Towards a theory of Non-Log-concave Sampling: First order stationarity guarantees for Langevin Monte Carlo"
- Krishnakumar Balasubramanian, Sinho Chewi, Murat A. Eldem, Adil Salim, Matthew Zhang
- 6] "The sample size required in Importance Sampling"
- Sourav Chatterjee, Persi Diaconis
- 7] "Efficient Approximate Posterior Sampling via Annealed Langevin Monte Carlo"
- Advait Parmekar, Lita Rout, Karthikeyan Shanmugam, Sanjay Shakkottai

RECAP

INVERSE PROBLEMS

Recall the inverse problem set up of Lecture 16 (Posterior Sampling overview).

We wish to recover some unknown x given access to noisy measurements

$$y = d(x) + \eta$$

observed measurement measurement operator unknown noise

MLE : $x^* = \underset{x}{\operatorname{argmin}} \|y - d(x)\|^2$

doesn't incorporate a "prior" over x

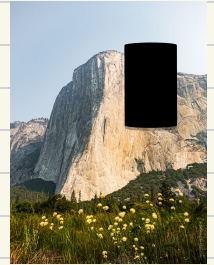
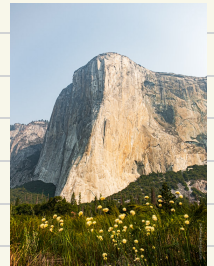
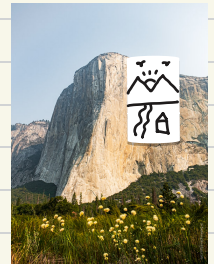
lets instead sample q_x such that

$$q_x = \underset{q}{\operatorname{argmin}} \mathbb{E}_{x \sim q} \|y - d(x)\|^2 + KL(q \| p)$$

p is our PRIOR. we assume we have trained scores

$S_\theta(x; t) \approx \nabla \ln p_t(x)$ where p_t is the noised prior

$$p_t = (p(e^t x) * \mathcal{N}_{1-e^{-t}}) e^{-td}$$



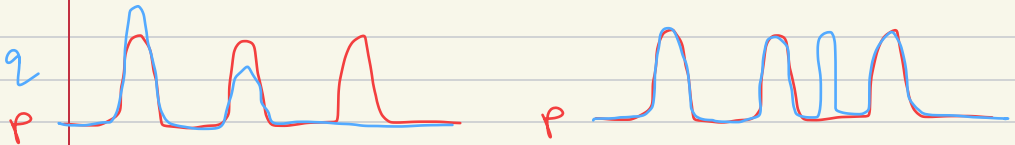
$$\underset{q}{\operatorname{argmin}} \mathbb{E}_{x \sim q} \|y - A(x)\|^2 + KL(q \| p)$$

① $\mathbb{E}_{x \sim q} \|y - A(x)\|^2$ small ensures consistency with y

② $KL(q \| p)$ small ensures "semantic" correctness to p

small $KL(q \| p)$

large $KL(q \| p)$



Generally we can use a loss $L_y(x)$ instead of $\|y - A(x)\|^2$

$L_y(x)$ is some known, smooth convex loss function

$$q_x = \underset{q}{\operatorname{argmin}} \mathbb{E}_{x \sim q} L_y(x) + KL(q \| p)$$

$$\Rightarrow q_x(x) = \frac{p(x) e^{-L_y(x)}}{Z} \quad \leftarrow \text{normalizing partition function}$$

$$Z = \int p(x) e^{-L_y(x)} dx$$

POSTERIOR SAMPLING

We can interpret this as sampling from a posterior

Bayes rule:

$$q_x \rightarrow \text{posterior} = \frac{\overset{e^{-L_y(x)}}{\downarrow} \text{likelihood} \times \overset{p(x)}{\nwarrow} \text{prior}}{\text{evidence} \leftarrow \int e^{-L_y(x)} p(x) dx}$$

COMPOSITE OPTIMIZATION/SAMPLING

Sampling from p is similar to maximizing $\ln p(x)$

max $\ln p$	sample from p
$dx = \nabla \ln p(x) dt$ (gradient flow)	$dx = \nabla \ln p(x) dt + \sqrt{2} db_t$ (Langevin Dynamics)

In our new setting, we want to sample from

is this related to $q_x \propto p e^{-L_y} = e^{\ln p - L_y}$

$\max -L_y + \ln p$? We can sample from e^{-L_y} using LD
We can sample from $e^{\ln p}$ using DDPM
we will see in a couple pages can we sample from the composite?

→ Different ways of looking at the same problem

"inverse problem $\equiv$ posterior sampling $\equiv$ composite sampling"

ALGORITHMS

Sanjay previously discussed a couple of algorithms,

- ① "Split Gibbs"
- ② "Monte-carlo importance sampling"

we will look in more detail about why they fail.

Split - Gibbs

Reference ①

Can we borrow techniques from optimizing

$$\max -L_y(x) + \ln p(x) \quad ?$$

If p has an efficient prox operator

$$\text{prox}_\varepsilon f(x) = \underset{z}{\operatorname{argmin}} f(z) + \|x - z\|^2 / 2\varepsilon$$

and $L_y(x)$ is smooth convex, then

"proximal gradient method" is an efficient algorithm.

Repeat

$$\begin{aligned} x'_t &= x_{t-1} - \nabla f(x_{t-1}) \cdot \eta \\ x_t &= \text{prox}_\varepsilon(x'_t) \end{aligned}$$

Sampling version of $\text{prox}_\varepsilon f$ is the "proximal sampler"

$$\tilde{f}_\varepsilon(x) \propto e^{\ln f(z) - \|x - z\|^2 / 2\varepsilon}$$

Interestingly, for p_t we actually almost have a proximal sampler.

$$\begin{aligned} \tilde{p}_t(x) &\propto e^{\ln p(z) + \|x - z\|^2 / 2\varepsilon} \\ &\propto \Pr(x_0 = x \mid x_t = z) \end{aligned}$$

where $x_t = x_0 + n \cdot \varepsilon$ (comes from heat process not on, but very related)

Motivates the following modification.

Repeat

$$\begin{array}{|l} x'_t = x_{t-1} - \eta \nabla L_y(x) + \sqrt{2\eta} \cdot n \quad (\text{replace grad with LMC step}) \\ x_t \sim \tilde{p}_\eta(x'_t) \quad (\text{replace prox operator with a sample from proximal sampler}) \end{array}$$

Havent really seen this. not sure why not.

Split Gibbs sampling modifies the first step

Repeat

$$\begin{array}{|l} x'_t \sim e^{-L_y(x_{t-1}) + \|x - x_{t-1}\|^2 / \eta} \\ x_t \sim \tilde{p}_\eta(x'_t) \end{array} \quad \left. \vphantom{\begin{array}{|l} x'_t \sim e^{-L_y(x_{t-1}) + \|x - x_{t-1}\|^2 / \eta} \\ x_t \sim \tilde{p}_\eta(x'_t) \end{array}} \right\} \begin{array}{l} \text{alternating} \\ \text{prox samplers} \end{array}$$

Note that $\tilde{p}_\eta(x'_t)$ is the conditional

$$p(x \mid x + \eta \cdot n = x'_t)$$

Not

$$p(x \mid e^{-\eta} x + \sqrt{1 - e^{-2\eta}} n = x'_t)$$

but we will ignore this for now.

The prox operator acts like a "projection".

Suppose

$$L_y(x) = \begin{cases} 0 & \text{if } x \in S \\ \infty & \text{if } x \notin S \end{cases}$$

convex, compact
set S

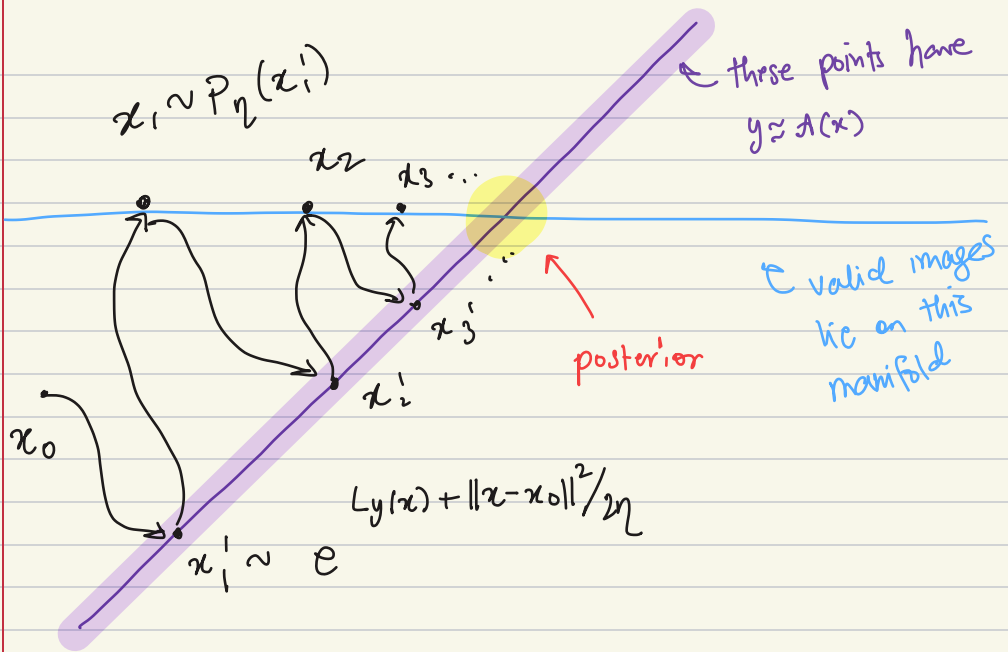
This is not smooth or convex; just for illustration.

then

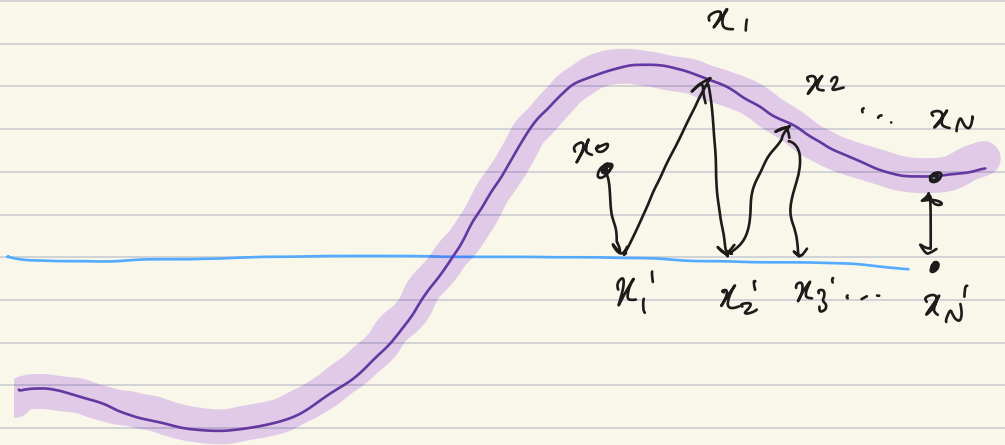
$$\exp(L_y(x) + \|x - z\|^2 / 2\eta)$$

is supported entirely on S , and as $\eta \rightarrow 0$,
converges to the projection of z onto S .

So split cribbs can be seen as alternately
"projecting" onto the set that aligns our iterate with
the measurements and the set that is consistent with
the prior.



Looks like it works well sometimes.



but could get stuck in local minima, rescued by the randomness of the sampling, but generally no "actual" convergence guarantee.

Reference ①

Run algorithm

Initialize $x_0 \sim \mathcal{N}$,

Repeat

$$\begin{cases} x'_i \sim \exp(-L_y(x) + \|x - x_{i-1}\|^2/2\eta) \\ x_i \sim \exp(\ln p(x) + \|x - x'_i\|^2/2\eta) \end{cases}$$

Let q_η denote the distribution

$$q_\eta(x, x') \propto \exp(\ln p(x) - L_y(x') + \|x - x'\|^2/\eta)$$

then

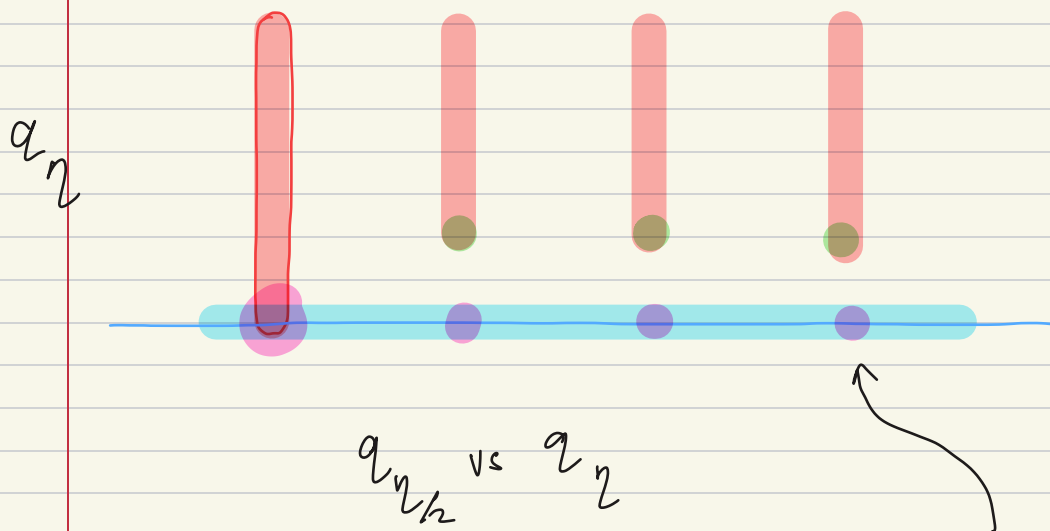
$$\text{Law}(x_i, x'_i) \rightarrow q_\eta$$

Notes

① $\lim_{\eta \rightarrow 0} q_\eta = p e^{-L} / Z$, our desired

posterior. The parameter η controls how "decoupled" the prior and reward model problems are allowed to be.

② Consider the following instance



Allowing larger η allows q_{η} to put density here

$$\exp(\ln p(x) - L_y(x') + \|x - x'\|^2 / \eta)$$

③ There is no bound on how long it takes for algorithm to converge to q_η for even a fixed η . For posterior sampling, we would need to also slowly send $\eta \rightarrow 0$. And as we saw, generally q_η could be far from $q_{\eta-\varepsilon}$ for small ε .

Monte carlo importance weighted Posterior Score estimation

$$\nabla \ln p_t(x_t | y) = \frac{E_{x \sim p_t(\cdot | x_t)} \left[p(y | x) \frac{x - x_t}{\sigma_t^2} \right]}{E_{x \sim p_t(\cdot | x_t)} [p(y | x)]}$$

We can compute the expected values using sample averages, but this could take a LOT of samples.

What we are really trying to do is compute the expected value of the posterior

$$E_{x \sim p_t(\cdot | x_t, y)} \frac{x - x_t}{\sigma_t^2}$$

using samples from $p_t(\cdot | x_t)$, which we can simulate using DDPM.

From Reference (6), the number of samples needed to estimate E_{ν} if using samples from μ goes as $\exp(KL(\nu || \mu))$. In our setting this is

$$\exp(KL(p(\cdot | x_t, y) || p(\cdot | x_t)))$$

Unclear how to get a handle on this. Lets look at t large, so

$$P_t(\cdot | x_t, Y) \approx P(\cdot | Y) \text{ and } P_t(\cdot | x_t) \approx P(\cdot)$$

then this the KL distance between the posterior

we want to sample from and the prior that

we can sample from efficiently

$$\exp(KL(p e^{-Ly}/Z \parallel P))$$

$$= \exp\left(\int \ln \frac{p e^{-Ly}}{Z \cdot P} d \frac{p e^{-Ly}}{Z}\right)$$

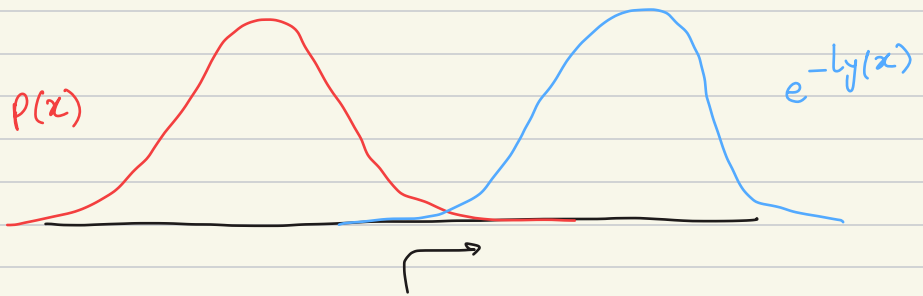
$$= \exp\left(\int \ln e^{-Ly} d \frac{p e^{-Ly}}{Z} - \int \ln Z d \frac{p e^{-Ly}}{Z}\right)$$

$$= \exp\left(\int \ln e^{-Ly} d \frac{p e^{-Ly}}{Z} - \ln Z\right)$$

$$= \exp\left(\frac{E}{p e^{-Ly}/Z} [-Ly]\right) / Z$$

the problem is that Z could be quite small.

$$Z = \int p(x) e^{-Ly(x)} dx$$



the points that matter the most for the expectation are very unlikely under the prior $p(z)$.

Fundamentally, still stuck on the Rejection Sampling issue that points with low $Ly(z)$ could occur very infrequently under the prior but still matter a lot for the posterior.

Satisfying posterior sampling guarantee remains elusive...

Hardness

There is a prior p which we can sample from in polynomial time, and quadratic L

$$L = \|y - Ax\|^2$$

for some A , such that the distribution

$$q \propto p e^{-L}$$

cannot be sampled from in

polynomial time under a standard cryptographic hardness assumption.

We will only discuss the hard instance and some intuition. See reference (2) for details.

Instance

Take a one way function $f^* : \mathbb{R}^d \rightarrow \mathbb{R}^d$, such that for all $x \in \mathbb{R}^d$, $f(x)$ can be computed in polynomial time, but for all $y \in \mathbb{R}^d$, we cannot find x such that $y = f(x)$ in polynomial time

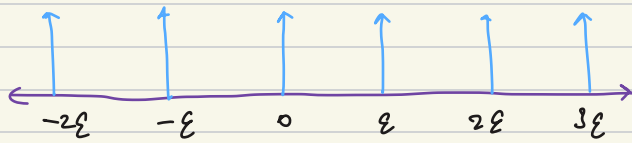
* hypothesized to exist. This is the "cryptographic hardness assumption"

We will construct the prior $p.$ over $\mathbb{R}^{2d}$ as a uniform mixture over product distributions g_s for $s \in \{0,1\}^d$

For any $s \in \{0,1\}^d$, define g_s as below.

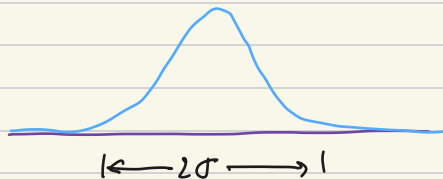
$$\text{Let } \text{Comb}_\varepsilon(x) = \sum_{k=-\infty}^{\infty} \delta(x - k\varepsilon)$$

denote the Dirac comb distribution



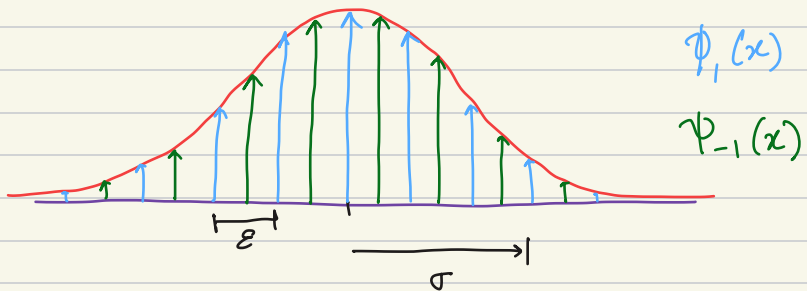
and let $\gamma_\sigma(x)$ denote the

σ^2 variance gaussian



Let $\psi_b(x)$ denote the discretized gaussian

$$\psi_b(x) \propto \gamma_1(x) \text{comb}_\varepsilon\left(x - \frac{\varepsilon}{2} \frac{1-b}{2}\right)$$



① The marginal on the first d coordinates is

$$x_i \sim \gamma_1(x - R s_i) \text{ for } i \leq d$$

② Let f be a DWF

$$x_i \sim \psi_{f(s)_{i-d}} \text{ for } i > d$$

$$\text{So } p = \sum_{s \in \{0,1\}^d} g_s \cdot \frac{1}{2^d}$$

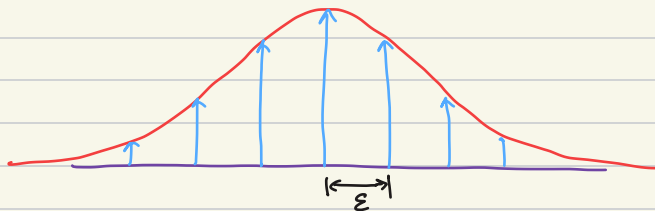
Lets break this down.

For every $(s, f(s))$, $s \in \{0,1\}^d$, there is a component g_s .

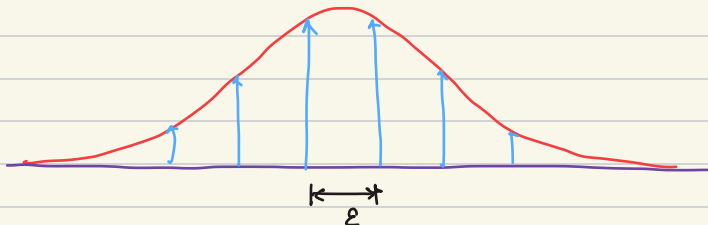
g_s is a distribution over $\mathbb{R}^{2d}$, the marginal over the first d -coordinates is a gaussian centered at Rs .

Meanwhile, the last d -coordinates encode the value of $f(s)$. Recall that f is a OWF.

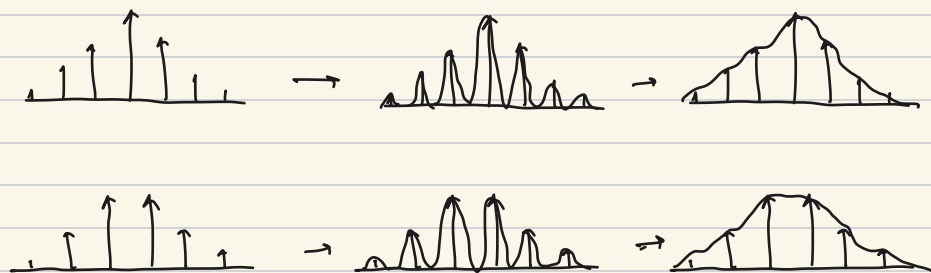
if $f(s)_{i-d} = 1$, then the marginal of the i^{th} coordinate of g_s is a comb-gaussian



if $f(s)_{i-d} = -1$, then the marginal of the i^{th} coordinate of g_s is a shifted comb-gaussian.



Note that $(g_s)_t$, the t -noised distribution quickly erases the distinction between the different shifts.



so for large enough t , $(g_s)_t$ is

$$\left((g_s)_t \right)_i(x) = \left(\gamma_i(x - R \cdot s_i) \right)_t \quad i \leq d$$

$\nwarrow$ marginal on i^{th} coordinate
 $\uparrow$ gaussian centered at $R \cdot s_i$, noised to time t .

$$\left((g_s)_t \right)_i(x) = \left(\gamma_i(x) \right)_t \quad i > d$$

$\uparrow$ "comb" aspect is washed out

and the mixture $p = \sum_{s \in \{0,1\}^d} g_s \cdot \frac{1}{2^d}$ is

$$p_i(x) = \text{Unif}(\mathcal{H}^d) \quad i \leq d$$

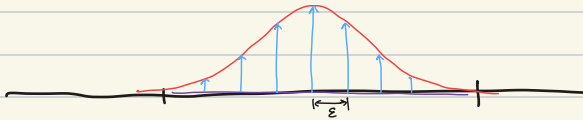
Illustration of g_s . Suppose $S = \{-1, 1\}$, $f(s) = \{1, -1\}$



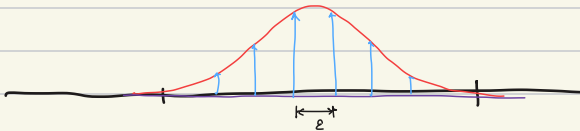
$$s_0 = -1$$



$$s_1 = 1$$



$$f(s)_0 = 1$$



$$f(s)_1 = -1$$

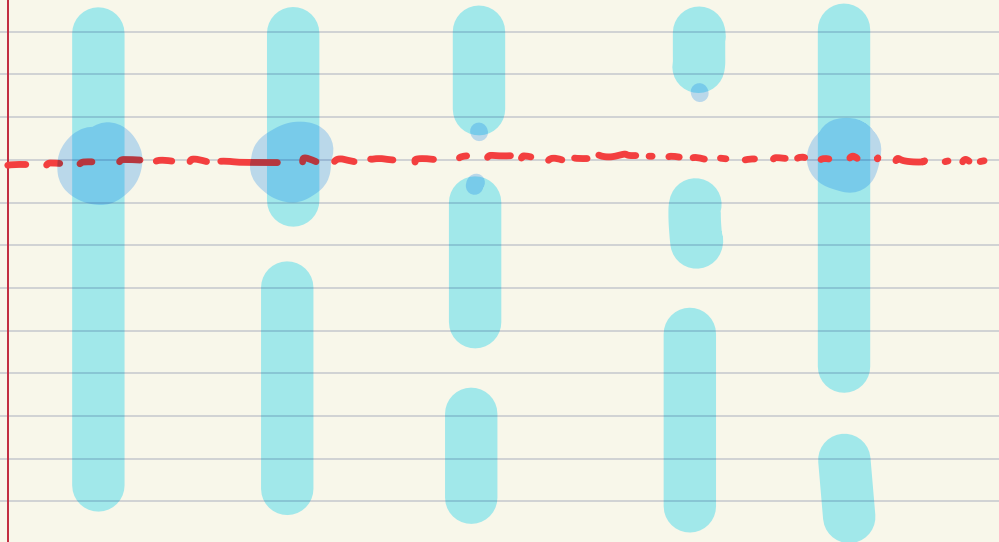
Measurement $y = Ax + \eta$ where $A = \begin{pmatrix} 0 & 0 \\ 0 & I_d \end{pmatrix}$,

so only the latter d coordinates (the ones encoding $f(s)$) are observed.

Idea is that the coordinates we observe are consistent with some specific $f(s)$ for some s . Because f is a onf, we cannot determine s in polynomial time from just $f(s)$, but if we had a posterior sampler, it would sample $(s, f(s))$ given $f(s)$ and invert the onf.

Difficulty is showing that the p we just defined admits a poly time sampler.

Note that $L_y(x) \sim \|y - Ax\|^2$ is actually a simple quadratic with A linear so the difficulty of posterior sampling isn't immediately removed by considering a simple L_y .



distribution consists of several vertical bars with gaps
the measurement is denoted by the horizontal red line
but when you note the prior, you can't tell which of
the modes will be consistent with the measurement.
By the time you can tell, you have already
committed to a mode.