

REFERENCES

- ① "Provable Benefit of Annealed Langevin Monte Carlo for Non-Log-concave Sampling"
 - Wei Cui, Molei Tao, Yongxin Chen.
- ② "Towards a theory of non-log-concave sampling: First order stationarity Guarantees for Langevin Monte Carlo"
 - Krishnakumar Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, Matthew Zhang
- ③ "Efficient Approximate Posterior Sampling with Annealed Langevin Monte Carlo"
 - Advait Panigrahy, Litu Rout, Kartikeyan Suresh, Sanjay Shakkottai

ANNEALED LANGEVIN DYNAMICS

Reference 1

Note: Here "annealed" is used in a different, more general sense than when one says "annealed scores".

For now, we mean it in a way that encompasses things like "simulated annealing".

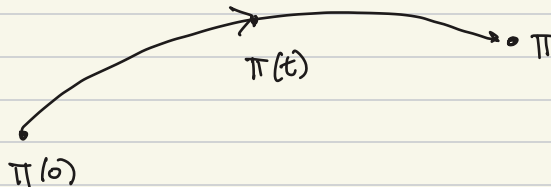
Recall Langevin Dynamics.

$$\left[\begin{array}{l} dX = \nabla \ln \pi(x) dt + \sqrt{2} dB_t \\ \text{initialized at } X(0) \sim \pi(0) \end{array} \right.$$

Let $\text{Law}(X(t)) = \pi(t)$. Then we have

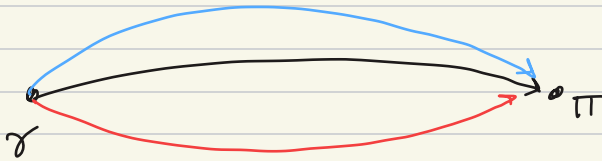
$$KL(\pi(t) \parallel \pi) \leq e^{-ct} KL(\pi(0) \parallel \pi)$$

where c depends on LSI .



This is an example of a method that transports one distribution to another. Generally all sampling related problems can be thought of in this way.

You can almost always^{*} view a sampling algorithm from the perspective of the path that the law of the iterates takes.



$$\pi^t = e^{\ln \pi \cdot t - x^2(1-t)} \quad \dots \text{"temperature annealing"}$$

$$\pi^t = \pi_t \quad \dots \text{"OU process"}$$

The different paths have different uses.

"Reverse OU process" $\rightarrow$ iterates move less, but also difficult to compute

"temperature annealing" $\rightarrow$ easy to compute but takes a long time.

We have seen that any AC path $t \mapsto \pi^t$ can be implemented as an SDE / ODE

"velocity / drift field for particles" $\Leftrightarrow$ "path of law of iterates"

Absolutely continuous (AC) path $t \mapsto \pi^t$ means $\forall t$

$$\lim_{\delta \rightarrow 0} \frac{W_2(\pi^t, \pi^{t+\delta})}{\delta} < \infty$$

where

minimize over all couplings of X and Y
 which have marginals μ and ν

$$W_2(\mu, \nu) = \left(\inf_{\substack{\Gamma: \Gamma(X) = \mu \\ \Gamma(Y) = \nu}} \int \|X - Y\|^2 d\Gamma(X, Y) \right)^{1/2}$$

denotes the Wasserstein-2 distance

the squared cost of moving

In particular, the SDE/ODE, which is a description of a sample path

Samples: $dX = v_t(X)dt + \sqrt{2}dB_t$; $X(t) \sim p(t)$

has a corresponding as an evolution of measure.

Laws: $\dot{\rho} = -\nabla \cdot (v_t \cdot \rho) + \Delta \rho$

To what extent can we design algorithms / think about problems from the "Law" perspective?



What should this path be? How can we traverse it?
 How can we think about errors?

measures on $\mathbb{R}^d$

Let $t \mapsto \pi^t$, $\pi^t \in \mathcal{P}_2(\mathbb{R}^d)$ be such a path.

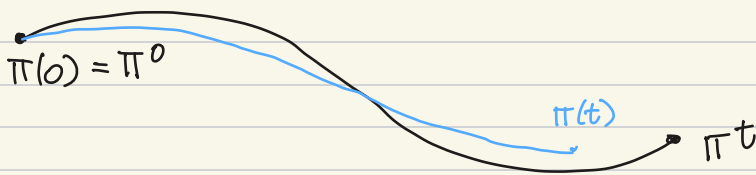
Annealed Langevin Dynamics is one way to try and follow this path

$$dX = \nabla \ln \pi^t(X) dt + \sqrt{2} dB_t$$

t changing target distribution.

At all times, running "Langevin" towards a changing distribution.

Depending on π^t , we may never really "converge"; there is nothing to converge to! Let's say we even initialize at $\text{Law}(X(0)) = \pi(0) = \pi^0$.



Can interpret this as trying to "track" π^t with $\pi(t)$.

To see how far we are at any time from where we "should be", let's try to bound $KL(\pi(t) \parallel \pi^t)$

if

$$\dot{p} = -\nabla \cdot (v_t \cdot p_t) \quad (1)$$

then

$$\|v_t\|_{L_2(p_t)} \geq \lim_{\delta \rightarrow 0} \frac{W_2(p_t, p_{t+\delta})}{\delta}$$

and in fact there is a unique field v_t satisfying (1)

$$\text{such that } \|v_t\|_{L_2(p_t)} = \lim_{\delta \rightarrow 0} \frac{W_2(p_t, p_{t+\delta})}{\delta}$$

so the optimal $\|v_t\|_{L_2(p_t)}$ is characterized by how fast p_t tries to move density around. This makes sense because v_t is a drift of an ODE which does precisely that. One way to interpret this is that the drift v_t implements many infinitesimal optimal transport maps.

Girsanov

Let $X(0) \sim p_1(0)$; $Y(0) \sim p_2(0)$ and

$$dX = v_t(X) dt + \sqrt{2} dB_t \Leftrightarrow \dot{p}_1 = -\nabla \cdot (v_t p) + \Delta p$$

$$dY = v'_t(Y) dt + \sqrt{2} dB_t \Leftrightarrow \dot{p}_2 = -\nabla \cdot (v'_t p) + \Delta p$$

then

$$KL(p_1(t) \| p_2(t)) \leq \int_0^t \frac{\|v_s - v'_s\|_{L_2(p_1)}^2}{2} ds + KL(p_1(0) \| p_2(0))$$

Now, to bound $KL(\pi^t \parallel \pi(t))$.

Let v_t generate π^t , so

$$\dot{\pi}^t = -\nabla \cdot (v_t \cdot \pi^t)$$

then we have

$$\dot{\pi}^t = -\nabla \cdot ((v_t + \nabla \ln \pi^t) \cdot \pi^t) + \Delta \pi^t$$

because

$$\begin{aligned} -\nabla \cdot (\nabla \ln \pi^t \cdot \pi^t) &= -\nabla \cdot \left(\frac{\nabla \pi^t}{\pi^t} \cdot \pi^t \right) \\ &= -\nabla \cdot (\nabla \pi^t) = -\Delta \pi^t \end{aligned}$$

Let $X(0) \sim \pi(0)$ and $Y(0) \sim \pi^0$ with

$$dX = (v_t + \nabla \ln \pi^t) dt + \sqrt{2} dB_t$$

$$\text{and } dY = \nabla \ln \pi^t dt + \sqrt{2} dB_t \quad \leftarrow \text{"Annealed LD"}$$

and denote $\text{Law}(Y(t)) = \pi(t)$

(note that $\text{Law}(X(t)) = \pi^t$ by definition)

so by Girsanov's

$$KL(\pi^t \parallel \pi(t)) \leq \int_0^t \|v_s\|_{L_2(\pi^s)}^2 ds + KL(\pi^0 \parallel \pi(0))$$

This quantity

$$\int_0^t \|v_s\|_{L_2(\Pi^s)}^2 ds$$

is called the "action" of path $s \mapsto \Pi^s$.

if $\|v_s\|_{L_2(\Pi^s)}$ is large, Π^t changes quickly at $t=s$ and so ALD "falls behind".

Interestingly, this bound scales with the speed at which we try to track Π^t .

let $t \rightarrow \Pi_K^t$ denote the same path, but with time moving K times slower.

$$\Pi_K^t = \Pi^{t/K}$$

then we have $\frac{d}{dt} \Pi_K^t = \frac{d}{dt} \Pi^{t/K} \cdot \frac{1}{K}$

so the focke-plank equation for Π_K^t reads

$$\begin{aligned} \frac{d}{dt} \Pi_K^t &= \frac{1}{K} \frac{d}{dt} \Pi^{t/K} = -\frac{1}{K} \nabla \cdot (v_{t/K} \Pi^{t/K}) \\ &= -\frac{1}{K} \nabla \cdot (v_{t/K} \Pi_K^t) \end{aligned}$$

so Π_k^t is generated by $\frac{1}{k} v_{t/k}$.

let $(v_k)_t$ denote the field that generates Π_k^t

$$\rightarrow (v_k)_t = \frac{1}{k} v_{t/k}$$

the action of this new path can be expressed in terms of the action of the old path.

let A denote the action $\int_0^T \|v_t\|_{L_2(\Pi^t)}^2 dt$

and let A_k denote the action $\int_0^{T/k} \|(v_k)_s\|^2 ds$

Let $s = t/k$. we have

$$\begin{aligned} \int_0^T \|v_t\|_{L_2(\Pi^t)}^2 dt &= \int_0^T \left\| \frac{1}{k} (v_k)_{t/k} \right\|^2 dt \\ &= \int_0^{T/k} \|(v_k)_s\|^2 ds \cdot \frac{k}{k^2} \\ &= \frac{1}{k} \int_0^{T/k} \|(v_k)_s\|^2 ds \end{aligned}$$

That is,

$$A = \frac{1}{k} A_k$$

so the KL distance between the algorithm and the trajectory can be reduced by simulating the trajectory slower. See Reference ① for details (including analysis of discretization error).

FIRST ORDER STATIONARITY IN SAMPLING

Reference ②

Recall that LD is Gradient flow of KL in Wasserstein.

Convergence of LD in KL is governed by LSI,

$$dx = \nabla \pi(x) dt + \sqrt{2} dB_t ; \text{ Law}(X(t)) = \pi(t)$$

$$\hookrightarrow \text{KL}(\pi(t) \parallel \pi) \leq e^{-t/C_{\text{LSI}}} \text{KL}(\pi(0) \parallel \pi)$$

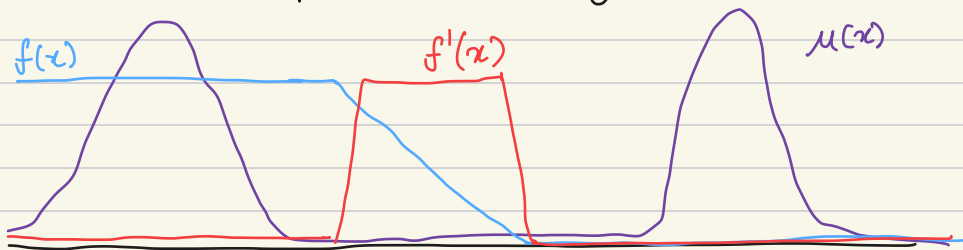
where $C_{\text{LSI}} \sim$ LSI constant. We also have the related:

$$\hookrightarrow \chi^2(\pi(t) \parallel \pi) \leq e^{-t/C_P} \chi^2(\pi(0) \parallel \pi)$$

where $C_P \sim$ Poincare constant

Poincare intuition

Poincare constant can be large for benign distributions like a well-separated mixture of Gaussians μ .

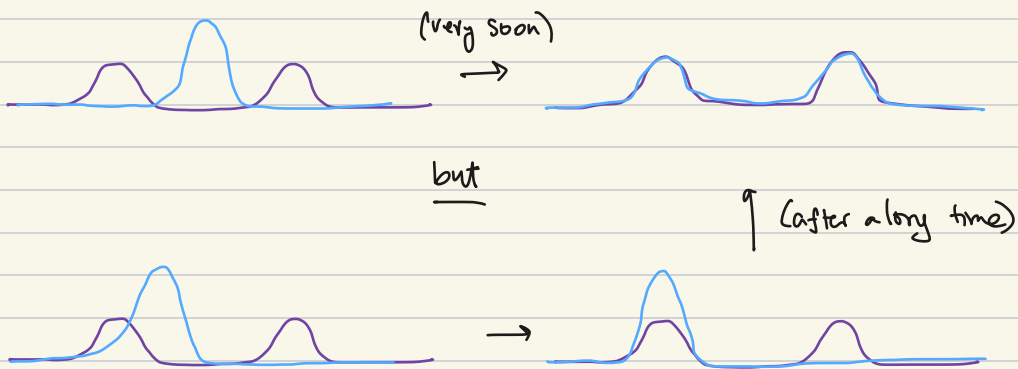


Here $\int f(x)^2 d\mu(x)$ is quite large but $\int (f'(x))^2 d\mu(x)$ is very small. The Poincaré constant of a distribution is small if it can explain the variance of any witness function by observing its derivatives.

Consider a mixture of two gaussians



LD doesn't mix well in this setting, since iterates can get stuck in one of the two modes for an exponentially long time (until a random kick sends them to the other mode)



depends a lot on initialization, or "warm start".

but there is a sense in which the iterate has "stabilized" quickly in both settings.

Fisher Divergence

We know from de Bruijn,

$$\frac{d}{dt} KL(\pi(t) \parallel \pi) = -FI(\pi(t) \parallel \pi)$$

$$\text{where } FI(\alpha \parallel \beta) = \mathbb{E}_{\alpha} \left\| \nabla \ln \frac{\alpha}{\beta} \right\|^2$$

An FI guarantee implies a type of "local" KL guarantee.

Let $B_{\varepsilon}(x)$ denote a ball of radius ε around x , and

let $\mu|_{B_{\varepsilon}(x)}$ denote the distribution μ conditioned on being in $B_{\varepsilon}(x)$

$$\text{So the density is } \frac{d\mu|_{B_{\varepsilon}(x_0)}}{dx} = \frac{1\{x \in B_{\varepsilon}(x_0)\}}{\mu(B_{\varepsilon}(x_0))}$$

$$\mathbb{E}_{x \sim \mu_1} KL(\mu_1|_{B_{\varepsilon}(x_0)} \parallel \mu_2|_{B_{\varepsilon}(x_0)}) \leftarrow \text{Average local KL guarantee}$$

$$\lesssim \varepsilon \sqrt{FI(\mu_1 \parallel \mu_2)} \leftarrow FI$$

In other words, convergence in FI $\Rightarrow$ conditioned on small neighborhoods, sampling is accurate in KL on average

$$\textcircled{1} \quad \mu|_{B_\varepsilon(x)}(x) \approx \frac{1}{\text{Vol}(B_\varepsilon(x))}$$

For smooth μ around x we have

$$\mu(y) = \mu(x) + (y-x)^T \nabla \mu(x) + O(\|y-x\|^2)$$

$$\Rightarrow \int_{y \in B_\varepsilon(x)} \mu(y) dy = \left(\mu(x) + O(\varepsilon^2) \right) \text{Vol}(B_\varepsilon(x))$$

and

$$\mu|_{B_\varepsilon(x)}(x) = \frac{\mu(x)}{\int_{y \in B_\varepsilon(x)} \mu(y) dy} \approx \frac{1}{\text{Vol}(B_\varepsilon(x))}$$

$\textcircled{2}$ let $\nabla_a f|_b$ denote the gradient w.r.t a evaluated at b . we have

$$\nabla_z \log \mu|_{B_\varepsilon(x)}(z) \Big|_{z=x} = \nabla \log \mu(x)$$

$$\nabla_z \log \mu|_{B_\varepsilon(x)}(z) \Big|_{z=x} = \nabla_z \log \mu(z) \Big|_{z=x} - \underbrace{\nabla_z \log \int_{y \in B_\varepsilon(x)} \mu(y) dy \Big|_{z=x}}_{\text{no } z}$$

$$= \nabla \log \mu$$

So

$$x_0 \sim \mu_1 \quad E_{KL}(\mu_1|_{B_\varepsilon(x_0)} \parallel \mu_2|_{B_\varepsilon(x_0)})$$

$$= E_{x_0 \sim \mu_1} E_{x_1 \sim \mu_1|_{B_\varepsilon(x_0)}} \left(\log \mu_1|_{B_\varepsilon(x_0)}(x_1) - \log \mu_2|_{B_\varepsilon(x_0)}(x_1) \right)$$

$$= E_{x_0 \sim \mu_1} E_{x_1 \sim \mu_1|_{B_\varepsilon(x_0)}} \left(\log \mu_1|_{B_\varepsilon(x_0)}(x_0 + (x_1 - x_0)) - \log \mu_2|_{B_\varepsilon(x_0)}(x_0 + (x_1 - x_0)) \right)$$

$$= E_{x_0 \sim \mu_1} E_{x_1 \sim \mu_1|_{B_\varepsilon(x_0)}} \left(\underbrace{\log \mu_1|_{B_\varepsilon(x_0)}(x_0) - \log \mu_2|_{B_\varepsilon(x_0)}(x_0)}_{\text{about the same}} + (x_1 - x_0)^T \left(\nabla \log \mu_1|_{B_\varepsilon(x_0)}(x_0) - \nabla \log \mu_2|_{B_\varepsilon(x_0)}(x_0) \right) \right)$$

order ε

$$\leq E_{x_0 \sim \mu_1} E_{x_1 \sim \mu_1|_{B_\varepsilon(x_0)}} \left(\underbrace{\left\| \nabla \log \mu_1|_{B_\varepsilon(x_0)}(x_0) - \nabla \log \mu_2|_{B_\varepsilon(x_0)}(x_0) \right\|^2}_{\text{no } x_1} \right)^{1/2}$$

$$\approx E_{x_0 \sim \mu_1} \left(\left\| \nabla \log \mu_1(x_0) - \nabla \log \mu_2(x_0) \right\|^2 \right)^{1/2}$$

$$= E \sqrt{FI(\mu_1 \parallel \mu_2)}$$

We can integrate de Bruijn to get the following:

$$\begin{aligned} -KL(\pi(T) \parallel \pi) + KL(\pi(0) \parallel \pi) &= \int_0^T \frac{d}{dt} KL(\pi(t) \parallel \pi) dt \\ &= \int_0^T F(\pi(t) \parallel \pi) dt \geq T F\left(\frac{1}{T} \int_0^T \pi(t) dt \parallel \pi\right) \end{aligned}$$

Here the last inequality follows by Jensen because

$F(\alpha \parallel \beta)$ is convex if α .

$$\text{So} \quad F\left(\frac{1}{T} \int_0^T \pi(t) dt \parallel \pi\right) \leq \frac{KL(\pi(0) \parallel \pi)}{T}$$

See [Reference ②](#) for an analysis of LMC.

Note that this is an "average iterate" guarantee,

that is, on the law of $\frac{1}{T} \int_0^T \pi(t) dt$

It will be useful for us to extend this analysis

to the setting of Annealed Langevin Dynamics

Now we have

$$\begin{aligned} \frac{d}{dt} KL(\pi(t) \parallel \pi^t) &= \frac{d}{ds} KL(\pi(t) \parallel \pi^s) \Big|_{s=t} \\ &\quad + \frac{d}{ds} KL(\pi(s) \parallel \pi^t) \Big|_{s=t} \end{aligned}$$

we have from before

$$\frac{d}{ds} \text{KL}(\pi(s) \parallel \pi^t) \Big|_{s=t} = -F_1(\pi(t) \parallel \pi^t)$$

and

$$\begin{aligned} \frac{d}{ds} \text{KL}(\pi(t) \parallel \pi^s) \Big|_{s=t} &= \frac{d}{ds} \int \ln \frac{\pi(t)}{\pi^s} d\pi(t) \Big|_{s=t} \\ &= \int \frac{d}{ds} \ln \frac{\pi(t)}{\pi^s} d\pi(t) \Big|_{s=t} \\ &= \int \frac{d}{ds} \ln \pi^s \Big|_{s=t} d\pi(t) \\ &= \mathbb{E}_{\pi(t)} \left[\frac{d}{ds} \ln \pi^s \right] \end{aligned}$$

so integrating

$$\begin{aligned} \text{KL}(\pi(0) \parallel \pi^0) + \int_0^T \mathbb{E}_{\pi(t)} \left[\frac{d}{dt} \ln \pi^t \right] dt &\geq \int F_1(\pi(t) \parallel \pi^t) dt \\ &\geq T F_1 \left(\frac{1}{T} \int \pi(t) dt \parallel \pi^t \right) \end{aligned}$$

$$F_1 \left(\frac{1}{T} \int_0^T \pi(t) dt \parallel \pi \right) \leq \underbrace{\text{KL}(\pi(0) \parallel \pi^0)}_{\text{initialization error}} + \underbrace{\int_0^T \mathbb{E}_{\pi(t)} \left[\frac{d}{dt} \ln \pi^t \right] dt}_{\text{tracking a moving path.}}$$

Not mentioned: discretization error. can be handled

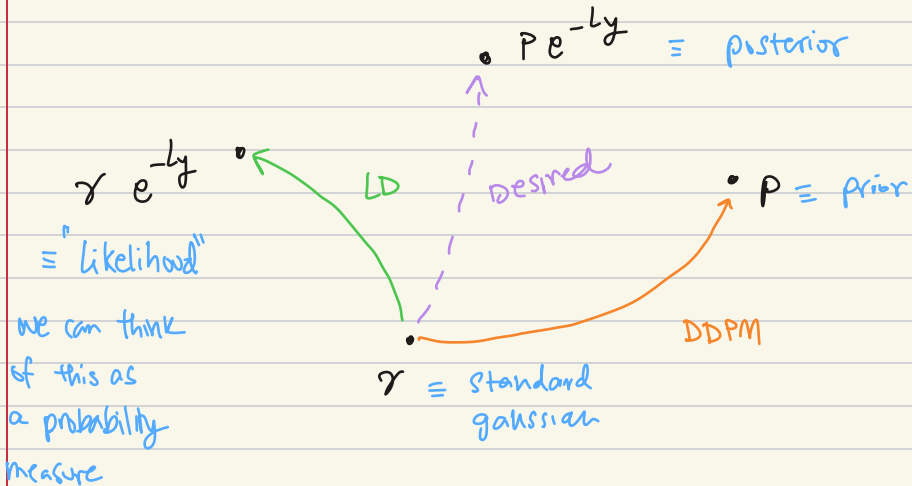
similarly to Reference ②

Now we have some handle on KL and F_1 for ALD.

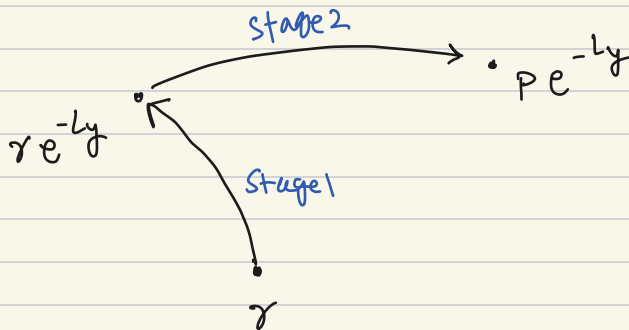
I have dropped all normalizing factors. Anything that looks like a measure is a probability measure.

Guarantees for posterior sampling

We have the following picture. (ignoring normalizing factors)



How do we combine the orange and green paths to lead us to $P e^{-Ly}$?



Stage 1: sample from γe^{-Ly} using LD

$$dX = (-\nabla Ly(X) - X) dt + \sqrt{2} dB_t$$

Because γe^{-Ly} is strongly log-concave
we can sample from it efficiently.

Stage 2: Travel from γe^{-Ly} to $p e^{-Ly}$ along
the path $t \mapsto P_{T-t} e^{-Ly}$ for some sufficiently large
 T . We will use μ_t to denote $P_t e^{-Ly}$.

We see that for large enough T , $P_T \approx \gamma \Rightarrow P_T e^{-Ly} \approx \gamma e^{-Ly}$

μ_{T-t} can be thought of as the posterior of the noised prior

P_{T-t} . As $t \rightarrow T$, this is the true posterior.

We will implement this transport using Annealed Langevin
Dynamics.

$$\begin{aligned} \text{Note that } \nabla \ln \mu_t(x) &= \nabla \ln \frac{P_t(x) e^{-Ly(x)}}{Z} \quad \leftarrow \text{not a function of } x \\ &= \nabla \ln P_t(x) - \nabla Ly(x) \end{aligned}$$

So LD towards μ_{T-t} is

$$dX = \left(\nabla \ln P_{T-t}(x) - \nabla Ly(x) \right) dt + \sqrt{2} dB_t$$

We will use P_t to denote the law of the
iterate $X(t)$.

Parameters

T_1 : how long to run stage 1, should be around $\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}$
to get $KL(re^{-Ly}, \text{Law}(X_{T_1})) < \varepsilon$

T_2 : where to start the "reverse process" stage 2;
should be around $\log \frac{d}{\varepsilon}$ to ensure that
 $KL(p_{T_2} e^{-Ly}, \text{Law}(X_{T_1})) < \varepsilon$

K : speed factor to improve errors

η : step size for discretization; set to $K^{-1/4}$

Algorithm:

Sample a standard gaussian

Take $X_0 \sim \mathcal{N}(0, I_d)$

Stage 1

for $i \in [T_1]$

iterate LMC for re^{-Ly}

$$X_i = X_{i-1} + (-\nabla \ln Ly(X_{i-1}) - X_{i-1})\eta + \sqrt{2\eta} \cdot \varepsilon$$

$$X_0 = X_{T_1}$$

Stage 2

for $i \in T_2 \cdot \frac{K}{2} \rightarrow 0$

iterate Annealed Langevin Monte Carlo

$$X_i = X_{i-1} + (s(X_{i-1}, i/\frac{K}{2}) - \nabla Ly(X_{i-1}))\eta + \sqrt{2\eta} \cdot \varepsilon$$

Assumptions

- ① P , the prior over $\mathbb{R}^d$ is m -subgaussian.
- ② The score of the prior $\nabla \ln p(x)$, is L -Lipschitz.
- ③ The loss $L_y(x)$, is R -smooth and convex

SKETCH

- ① The results about ALD allow us to track within KL provided we bound the action of the path.

$$KL(\mu_{T-t} \| p_t) \lesssim \int_0^t \|v_s\|_{L_2(\mu_{T-s})}^2 ds$$

↑ not exact inequality because of discretization errors

where v_t generates μ_{T-t}

$$\frac{d}{dt} \mu_{T-t} = -\nabla \cdot (v_t \cdot \mu_{T-t})$$

Once we have this, we can use the speed K to control the error in KL.

- ② The results about "approximate first order stationarity" allow us to track in FI provided we bound

$$\int_{p_t} E \left[\frac{d}{dt} \ln \mu_{T-t} \right] dt$$

Need to bound two quantities that depend on how fast μ_{T-t} changes.

- (1) The action $\int_0^T \|v_s\|_{L_2(\mu_{T-s})}^2 ds$ requires some "metric" understanding, depends on a Wasserstein limit — how fast does mass move under $t \mapsto \mu_{T-t}$?
- (2) The quantity $E_{P_t} \frac{d}{dt} \ln \mu_{T-t}$ is more of an "information" based quantity, like a KL distance.

The key to both of these is the following lemma

(we know that $\mu_t \rightarrow \gamma$, this characterizes how fast, pointwise)

$$\frac{d}{dt} \ln \mu_t(x) \leq \frac{e^{-t}}{1 - e^{-2t}} \text{poly}(m, d, R, d) (1 + \|x\| + \|x\|^2)$$

$\uparrow$
 $\sim e^{-t}$ for large t
 $\sim O(\frac{1}{t})$ for small t

$\uparrow$
dimension d

$\uparrow$
at most quadratic in x

Various constants, subgaussian norm of the prior, Lipschitz constant of the score, smoothness of the log likelihood, which we will assume for readability are all 1 for the next few pages.

Once we have these, we can show that

Theorem

Let p_t denote the law of X_t in stage 2 of the alg.

There exists $t_{\#} \leq \text{poly}(1/k)$

satisfying both

$$\rightarrow \text{KL}(\mu_{t_{\#}}, p_{t_{\#}}) \leq \text{poly}(d, 1/k)$$

$$\rightarrow \text{FI}(p_{t_{\#}}, \mu_0) \leq \text{poly}(d/k)$$

Notes

1) In comparison to LD guarantees, results based on

Girsanov switch the arguments in the KL. You can sweep this

under the rug by giving a TV guarantee, because

$$\text{TV} < \sqrt{\text{KL}}$$

and TV is symmetric

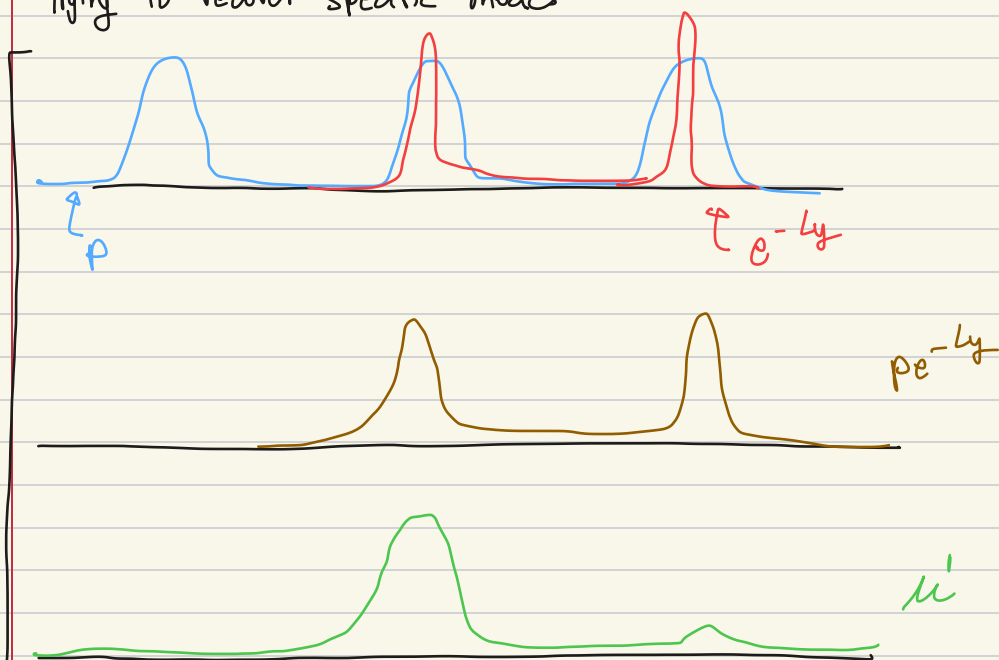
2) The Fisher guarantee by itself would actually follow

from **Reference ②** if we ran LMC towards μ_0 , but

we will see that this is not useful.

lets assume we
have this as SC: :o)

In an instance consisting of a mixture of Gaussians, one can think of almost a canonical posterior sampling problem as trying to recover specific modes



In this figure, the top row describes the prior and likelihood. The second row describes the ideal posterior, while the final row offers a distribution μ' satisfying $FI(\mu', pe^{-L_y})$ small. This is sometimes called "mode collapse".

The benefit of using Annealed LMC as we describe is that it naturally specifies a "warm start" schedule for distributions that get closer to the true posterior.

The KL guarantee for μ_{t*} serves as a regularizer that prevents the type of mode collapse referenced previously.

- 3) In comparison with Split Cribbs, this method gives us a way to tune a hyperparameter K that directly polynomially controls the errors in KL and $\mathcal{F}$.