

Discretized LMC — A Girsanov Theorem Approach

Refs:

- ① S. Chewi, Log-concave sampling, (book draft), Nov. 2024.
2. A. Wibisono, Sampling as optimization in the space of measures: the Langevin Dynamics as a composite optimization problem. COLT 2018.
- ③ A. Dalalyan and A. Tsybakov, Sparse regression learning by aggregation and Langevin Monte Carlo, Journal of Comput. System Science, 2012.
4. S. Vempala and A. Wibisono, Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices, NeurIPS 2019.
5. S. Chewi, M. Erdogdu, M. Li, R. Shen and M. Zhang, Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev, arXiv: 2112.12662, 2021.
- ⑥ B. Oksendal, Stochastic Differential Equations, Springer 1998.

(Notes closely follows [1], with [6] for Girsanov Thm)

II

Proof via Girsanov Theorem

Unlike the Wasserstein coupling proof, in this case, we only assume that $V(\cdot)$ is β -smooth (no strong convexity). However, the results will be correspondingly weaker. We closely follow [1], which in turn, builds on [3].

The approach below is based on Girsanov Theorem, which we restate here from Note 9.

Girsanov's Theorem (Thm 8.6.4 in [6])

Let $\gamma_t \in \mathbb{R}^d$ be a Ito process of the form:

$$d\gamma_t = B_t(\gamma_t) dt + \sigma_t(\gamma_t) dB_t, \quad t \in [0, T]$$

$B_t \in \mathbb{R}^{d'}$ is a Brownian motion, over $(\Omega, \mathcal{F}, \mathbb{P}_T)$,
 $B_t(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^d$, and $\sigma_t(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^{d \times d'}$ (a $d \times d'$ matrix).

Suppose that $\exists u_t(\cdot), \alpha_t(\cdot)$ (both $\mathbb{R}^d \mapsto \mathbb{R}^d$)

$$\text{s.t. } \sigma_t(\gamma_t) u_t(\gamma_t) = \beta_t(\gamma_t) - \alpha_t(\gamma_t),$$

$$\text{and } E_{\mathbf{P}_T} \left[e^{\frac{1}{2} \int_0^T u_s^2(\gamma_s) ds} \right] < \infty \quad \left(\text{Novikov's condition} \right)$$

$$\text{Let } M_t = e^{\left(- \int_0^t u_s(\gamma_s) dB_s - \frac{1}{2} \int_0^t u_s^2(\gamma_s) ds \right)} \quad 0 \leq t \leq T,$$

and a new measure $\mathbb{Q}_T$ that is absolutely continuous w.r.t $\mathbf{P}_T$, with M_t being the Radon-Nikodym derivative, i.e.,

$$d\mathbb{Q}_T(\omega) = M_T(\omega) d\mathbf{P}_T(\omega) \quad \text{over } \mathcal{F}.$$

Then, we have :

$$\textcircled{1} \quad \widehat{B}_t = \int_0^t u_s(\gamma_s) ds + B_t \quad \text{is a B.M. over } (\Omega, \mathcal{F}, \mathbb{Q}_T)$$

$\textcircled{2}$ The process γ_t has the Stochastic Integral

representation: $d\gamma_t = \alpha_t(\gamma_t) dt + \sigma_t(\gamma_t) d\hat{B}_t$

Remark: Summarizing, we have

$$\begin{aligned} d\gamma_t &= \beta_t(\gamma_t) dt + \sigma_t(\gamma_t) dB_t \quad \text{BM over } (\Omega, \mathcal{F}, P_T) \\ &= \alpha_t(\gamma_t) dt + \sigma_t(\gamma_t) d\hat{B}_t \quad \text{BM over } (\Omega, \mathcal{F}, Q_T) \end{aligned}$$

Now, let us define a new process $\tilde{\gamma}_t$ as:

$$d\tilde{\gamma}_t = \alpha_t(\tilde{\gamma}_t) dt + \sigma_t(\tilde{\gamma}_t) dB_t \quad \text{BM over } (\Omega, \mathcal{F}, P_T)$$

Thus, $(\gamma_t, \tilde{\gamma}_t)$ are a pair of processes that are coupled by the same B.M. B_t , but with different drift terms.

Also, observe that this trajectory γ_t driven by $\hat{B}_t$ (under measure Q_T) is the same in distribution as $\tilde{\gamma}_t$ driven by B_t (under measure P_T).

Above, (P_T, Q_T) are measures over the entire path $[0, T]$. Let the marginal dist. associated with these path measures be P_t, q_t respectively, i.e., for any $t \in [0, T]$,

$$P_t(A) = P_T(\gamma_t \in A)$$

Below immediately generalizes to any joint dist., e.g.,
 $P_{t_1, t_2}(A, B) = P_T(\gamma_{t_1} \in A, \gamma_{t_2} \in B)$

Further, suppose $d = d'$ (i.e., the BM is the same dim. as the ambient dim.), and $\sigma_t(\cdot) = c \cdot I_d$ (this condition can be removed, but we keep this to keep notation simple for our use case.).

In this setting, we have

$$u_t(\gamma_t) = \frac{1}{c} (B_t(\gamma_t) - \alpha_t(\gamma_t)).$$

Corollary 1: With setting and notation as above, we have

$$KL(Q_t \| P_t) \leq \frac{1}{2c^2} \int_0^T E_{Q_T} [\| \beta_s(\gamma_s) - \alpha_s(\gamma_s) \|^2] ds$$

(The above result holds for any joint dist. $q_{t_1, t_2, \dots}(\cdot)$ and $p_{t_1, t_2, \dots}(\cdot)$, with no change in the proof.)

Proof: From the data-processing inequality (KL divergence increases when distributions are lifted to a higher-dim. joint dist.), we have

$$KL(Q_t \| P_t) \leq KL(Q_T \| P_T)$$

$$\text{But } KL(Q_T \| P_T) = E_{Q_T} \left[\ln \frac{dQ_T}{dP_T} \right]$$

$$= E_{Q_T} [\ln M_T]$$

Recall from Girsanov's Thm:

$$\underbrace{\sigma_t^*(\gamma(t))}_{= cI_d} u_t(\gamma(t)) = \beta_t(\gamma(t)) - \alpha_t(\gamma(t)),$$

$$\frac{dQ_T}{dP_T} = M_t = e^{\left(- \int_0^t u_s(\gamma(s)) dB(s) - \frac{1}{2} \int_0^t u_s^*(\gamma(s)) ds \right)}$$

$$= E_{Q_T} \left[-\frac{1}{c} \int_0^T (\beta_s(\gamma_s) - \alpha_s(\gamma_s)) dB_s - \frac{1}{2c^2} \int_0^T \|\beta_s(\gamma_s) - \alpha_s(\gamma_s)\|^2 ds \right]$$

note B_s is not a BM under Q_T ;

however, $\hat{B}_s$ is a BM

→ (A)

Next, recall that:

$$\hat{B}_t = B_t + \frac{1}{c} \int_0^t (\beta_s(\gamma_s) - \alpha_s(\gamma_s)) ds$$

$$\therefore dB_t = d\hat{B}_t - \frac{1}{c} (\beta_t(\gamma_t) - \alpha_t(\gamma_t)) dt$$

Substituting this in (A) above, and using the fact that any stochastic integral has 0-mean ($E_{Q_T}[S \dots d\hat{B}(s)] = 0$), we have:

$$KL(Q_T \| P_T) = \frac{1}{2c^2} \int_0^T E_{Q_T} [\|\beta_t(\gamma_t) - \alpha_t(\gamma_t)\|^2] dt$$

□

LMC Discretization

Setting and Notation: $\pi_*(x) \sim e^{-V(x)}$, $x \in \mathbb{R}^d$

$$dX_t = -\nabla V(X_t) dt + \sqrt{2} dB_t \quad (1D)$$

Langevin Dynamics

continuous-time evolution

$$\tilde{X}_{(k+1)h} = \tilde{X}_{kh} - h \nabla V(\tilde{X}_{kh}) + \sqrt{2} \underbrace{(B_{(k+1)h} - B_{kh})}_{\sim N(0, h), \text{ indep. of } \tilde{X}_{kh}}$$

Langevin Monte Carlo (LMC)

discrete-time evolution.

$\sim N(0, h)$,
indep. of $\tilde{X}_{kh}$

We need to embed this discrete-time evolution back into continuous time. The mapping ensures that:

- (i) At each continuous time when $t = kh$, the discrete evolution equals the value of the embedding, and
- (ii) Between timesteps, the drift term is linearly interpolated and the Brownian motion naturally

evolves as below:

$$\tilde{X}_t = \tilde{X}_{\lfloor \frac{t}{h} \rfloor h} - (t - \lfloor \frac{t}{h} \rfloor h) \nabla V(\tilde{X}_{\lfloor \frac{t}{h} \rfloor h}) + \sqrt{2} (B_t - B_{\lfloor \frac{t}{h} \rfloor h})$$

continuous-time embedding of LMC.

Theorem 1: (Thm 4.4.1 in [1]) Setting as above, and assume $\nabla V(\cdot)$ is β -Lipschitz. Let μ_{kh} be the law of LMC at times $t = kh$, $k=0,1,\dots$ and let π_t be the law of LD with initial condition $X_0 \sim \mu_0$. Then, for $h \in [0, \frac{1}{\beta}]$, and $T = Nh$,

$$KL(\mu_T \| \pi_T) \leq \beta^2 h^3 \sum_{k=0}^{N-1} E_{Z_k \sim \mu_{kh}} [\|\nabla V(Z_k)\|^2] + \beta^2 dh T$$

Proof: Let $\{B_t, 0 \leq t \leq T\}$ be a Brownian

motion under the prob. triplet $(\Omega, \mathcal{F}, \mathbb{P}_T)$.

Note: The random variables here are continuous path-valued, i.e., a map from Ω to $C[0, T]$ (cont. functions, endowed with the sup-norm topology). Distributions for r.v.s / random vectors at any finite set of times are defined by appropriate marginalization (formally, use measures over appropriate cylinder sets).

Then, with measure $\mathbb{P}_T$, we have that:

$$dX_t = -\nabla V(X_t)dt + \sqrt{2} dB_t, \quad X_0 \sim \mu_0$$

is the standard LD.

Further, with respect to the same Brownian motion, we have the embedded ZMC as:

$$\begin{aligned} \tilde{X}_t &= \tilde{X}_{\lfloor \frac{t}{h} \rfloor h} - (t - \lfloor \frac{t}{h} \rfloor h) \nabla V(\tilde{X}_{\lfloor \frac{t}{h} \rfloor h}) \\ &\quad + \sqrt{2} (B_t - B_{\lfloor \frac{t}{h} \rfloor h}) \end{aligned}$$

In differential form, we have:

$$d\tilde{X}_t = -\nabla V(\tilde{X}_{\lfloor \frac{t}{h} \rfloor h}) dt + \sqrt{2} dB_t$$

Thus, from Corollary 1, we have:

$$KL(Q_T \| P_T) = \frac{1}{4} \int_0^T E_{Q_T} [\|\nabla V(\tilde{X}_{\lfloor \frac{t}{h} \rfloor h}) - \nabla V(X_t)\|^2] dt$$

∇V is β -Lipschitz

$$\leq \frac{\beta^2}{4} \int_0^T E_{Q_T} [\|\tilde{X}_{\lfloor \frac{t}{h} \rfloor h} - X_t\|^2] dt$$

$$\leq \frac{\beta^2}{4} \int_0^T \left((t - \lfloor \frac{t}{h} \rfloor h)^2 E_{Q_T} [\|\nabla V(\tilde{X}_{\lfloor \frac{t}{h} \rfloor h})\|^2] + \underbrace{2d(t - \lfloor \frac{t}{h} \rfloor h)}_{\text{From B.M. Variance.}} \right) dt$$

$$\leq \frac{\beta^2 h^3}{12} \sum_{k=0}^{N-1} E_{Q_T} [\|\nabla V(\tilde{X}_{kh})\|^2] + \frac{\beta^2 d h^2 N}{4}.$$

□

The above gives a guarantee on $KL(\mu_{Nh} \parallel \pi_{Nh})$. We would like a guarantee with respect to π_* , the stat. measure of the LD. Since KL does not satisfy a Δ^c inequality, we use $\sqrt{KL} \geq \|\cdot\|_{TV}$ (total variation) from Pinsker's inequality.

Corollary 2 (Cor. 4.4.2 in [1]): Assume ∇V is β -Lipschitz and also that π_* is $\frac{1}{\alpha}$ -LSI. Then

for small enough $\varepsilon > 0$, and h chosen appropriately,

$$\|\mu_{Nh} - \pi_*\|_{TV} \leq \varepsilon \quad \text{if} \quad N = \tilde{O} \left(\frac{\kappa^2 d}{\varepsilon^2} \left(\ln \frac{KL(\mu_0 \parallel \pi_*)^2}{\varepsilon^2} \right) \right)$$

We refer to [1] for the proof. We comment that the α -LSI assumption is used to bound

$$\sqrt{KL(\pi_{Nh} \parallel \pi_*)} < \frac{\varepsilon}{2} \quad \text{for}$$

$$T = Nh \approx \frac{1}{\alpha} \ln \frac{KL(\mu_0 \parallel \pi_*)}{\varepsilon^2},$$

because the LD converges exponentially fast under the $(\frac{1}{\alpha})$ -LSI condition.

Please read [17] for the proof.

