CSC/DSCC 574 Continuous Algorithms for Optimization and Sampling Lecture 9

Proximal Sampling

Lecturer: Jiaming Liang *March 26, 2024*

1 Alternating sampling framework

Our goal in this lecture is again to sample from $\nu(x) \propto \exp(-f(x))$. Recall that one step of Langevin Monte Carlo (LMC) is

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{2\eta} z, \quad z \sim \mathcal{N}(0, I)$$

and is equivalent to sampling $x_{k+1} \sim p(y|x_k)$ where

$$p(y|x_k) \propto \exp\left(-\frac{1}{2\eta}\|x - (x_k - \eta \nabla f(x_k))\|^2\right).$$

We have also seen that LMC is biased, namely $\rho_k \to \bar{\rho}$ as $k \to \infty$, but $\mathsf{KL}(\bar{\rho}|\nu) > 0$. A natural way to fix this bias issue is to use the Metropolis-Hastings filter. Given x_k, we take $p(\cdot|x_k)$ as a proposal density, draw $y_k \sim p(y_k|x_k)$, and accept y_k with probability

$$\min\left\{1, \frac{\nu(y_k)p(x_k|y_k)}{\nu(x_k)p(y_k|x_k)}\right\}.$$

This is the so-called Metropolis-adjusted Langevin algorithm. Since the Metropolis-Hastings filter makes the Markov chain reversible and hence ν is the stationary distribution.

We will explore another idea in this lecture, namely the Gibbs sampling, to generate an unbiased sample. Consider a joint distribution

$$\pi(x, y) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x - y\|^2\right)$$

where $\eta > 0$ is a regularization parameter (or the stepsize of Algorithm 1 below), and apply Gibbs sampling on $\pi(x, y)$, i.e., alternatively sampling from conditional distributions $\pi^{Y|X}$ and $\pi^{X|Y}$.

Algorithm 1 Alternating Sampling Framework

1. Sample $y_k \sim \pi^{Y|X}(y \mid x_k) \propto \exp\left(-\frac{1}{2\eta}\|x_k - y\|^2\right)$
2. Sample $x_{k+1} \sim \pi^{X|Y}(x \mid y_k) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x - y_k\|^2\right)$

Observe that the X-marginal of π is the target distribution ν, whereas the conditional distribution of Y given X is Gaussian: $\pi^{Y|X}(\cdot \mid x) = \mathcal{N}(x, \eta I)$. Therefore, the Y-marginal is the convolution of π^X with a Gaussian, $\pi^Y = \pi^X * \mathcal{N}(0, \eta I)$.

It is known from Gibbs sampling that $\{x_k, y_k\}_{k\geq 1}$ form a reversible Markov chain with stationary distribution $\pi(x, y)$, whose X-marginal is $\nu(x)$. Therefore, assuming we can exactly implement Steps 1 and 2 of Algorithm 1, then we eventually generate an unbiased sample from the target $\nu(x)$.

1.1 Restricted Gaussian oracle

The conditional distribution of X given Y in Step 2 of Algorithm 1 is the "regularized" distribution

$$\pi^{X|Y}(x \mid y) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x - y\|^2\right).$$

Definition 1. *Given a point $y \in \mathbb{R}^d$ and stepsize $\eta > 0$, the restricted Gaussian oracle (RGO) for $f : \mathbb{R}^d \to \mathbb{R}$ is a sampling oracle that returns a random sample from a distribution proportional to $\exp(-f(\cdot) - \|\cdot - y\|^2/(2\eta))$.*

In view Definition 1, Step 2 of Algorithm 1 is an RGO, whose implementation is nontrivial and needs a subroutine for the exact realization.

Algorithm 2 Exact Implementation of the RGO

1. Compute $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} \left\{f_y^\eta(x) := f(x) + \frac{1}{2\eta}\|x - y\|^2\right\}$;
2. Generate $X \sim \exp(-h_1(x))$ where $h_1(x) := \frac{1}{2\eta}\|x - x^*\|^2 + f_y^\eta(x^*)$;
3. Generate $U \sim \mathcal{U}[0, 1]$;
4. If
$$U \leq \frac{\exp(-f_y^\eta(X))}{\exp(-h_1(X))}, \tag{1}$$
then accept X; otherwise, reject X and go to step 2.

Lemma 1. *Assume f is convex and M-Lipschitz continuous. Define*

$$h_2 := \frac{1}{2\eta}\|\cdot - x^*\|^2 + 2M\|\cdot - x^*\| + f_y^\eta(x^*).$$

Then, for every $x \in \mathbb{R}^d$, we have $h_1(x) \leq f_y^\eta(x) \leq h_2(x)$.

Proof. The first inequality follows from Step 1 of Algorithm 1 and the strong convexity of $f_y^\eta(x)$

$$f_y^\eta(x) \geq f_y^\eta(x^*) + \frac{1}{2\eta}\|x - x^*\|^2 = h_1(x).$$

Also, noting from Step 1 of Algorithm 1 that

$$\frac{y - x^*}{\eta} \in \partial f(x^*). \tag{2}$$

It follows from the assumption that f is M-Lipschitz continuous that

$$f(x) - \ell_f(x; x^*) \leq 2M\|x - x^*\|.$$

Using the above inequality and (2), we have

$$\begin{aligned} f_y^\eta(x) &\leq \ell_f(x; x^*) + \frac{1}{2\eta}\|x - y\|^2 + 2M\|x - x^*\| \\ &\overset{(2)}{=} f(x^*) + \langle f'(x^*), x - x^*\rangle + \frac{1}{2\eta}\|x - y\|^2 + 2M\|x - x^*\| \\ &= \frac{1}{2\eta}\|\cdot - x^*\|^2 + 2M\|\cdot - x^*\| + f_y^\eta(x^*) = h_2(x). \end{aligned}$$

□

The following lemma summarizes basic properties of Algorithm 2.

Lemma 2. *The sample X generated by Algorithm 2 follows the distribution $\pi^{X|Y}$. Let S denote the event that (1) happens. Then,*

$$\mathbb{P}(S) = \frac{\int \exp(-f_y^\eta(x))\mathrm{d}x}{\int \exp(-h_1(x))\mathrm{d}x}.$$

Proof. Let $k(x|S)$ denote the conditional density of X given S. We have

$$k(x|S) = \frac{\mathbb{P}(S|X = x)q(x)}{\mathbb{P}(S)}, \quad \mathbb{P}(S|X = x) = \frac{\exp(-f_y^\eta(x))}{\exp(-h_1(x))}, \quad q(x) = \frac{\exp(-h_1(x))}{\int \exp(-h_1(x))\mathrm{d}x}$$

Also,

$$\mathbb{P}(S) = \int \mathbb{P}(S|X = x)q(x)\mathrm{d}x = \int \frac{\exp(-g^\eta(x))}{\exp(-h_1(x))}\frac{\exp(-h_1(x))}{\int \exp(-h_1(x))\mathrm{d}x}dx = \frac{\int \exp(-f_y^\eta(x))\mathrm{d}x}{\int \exp(-h_1(x))\mathrm{d}x}.$$

So,

$$k(x|S) = \frac{\mathbb{P}(S|X = x)q(x)}{\mathbb{P}(S)} = \frac{\exp(-f_y^\eta(x))}{\int \exp(-f_y^\eta(x))\mathrm{d}x} = \pi^{X|Y}(x|y),$$

where $\pi^{X|Y}(x|y)$ is the nomalized density of $\exp(-f_y^\eta(x)$. Therefore, we verify $X \sim \pi^{X|Y}$. □

We need the following technical lemma about Gaussian integrals.

Proposition 1. *The following statements hold:*

a) $\int \exp\left(-\|x\|^2/(2\eta)\right) \mathrm{d}x = (2\pi\eta)^{d/2}$ *for every* $\eta > 0$*;*

b) let $a \geq 0$ *and* $d \geq 1$*, if*

$$2a(\eta d)^{1/2} \leq 1, \tag{3}$$

then

$$\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\eta}\|x\|^2 - a\|x\|\right) \mathrm{d}x \geq \frac{(2\pi\eta)^{d/2}}{\sqrt{e}}. \tag{4}$$

Proof. a) This statement is a well-known result.

b) First note the fact that

$$a\|x\| \leq a^2\|x\|^2 + \frac{1}{4}.$$

This inequality and a) imply that

$$\begin{aligned}\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\eta}\|x\|^2 - a\|x\|\right)\mathrm{d}x &\geq \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\eta}\|x\|^2 - a^2\|x\|^2 - \frac{1}{4}\right) \mathrm{d}x \\ &= \exp\left(\frac{-1}{4}\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\tilde{\eta}}\|x\|^2\right) \mathrm{d}x = \exp\left(\frac{-1}{4}\right) (2\pi\tilde{\eta})^{\frac{d}{2}}\end{aligned}$$

where $\tilde{\eta} = (1/\eta + 2a^2)^{-1}$. It follows from (3) that

$$\tilde{\eta} \geq \left(1 + \frac{1}{2d}\right)^{-1} \eta.$$

Combining the above inequalities, we have

$$\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\eta}\|x\|^2 - a\|x\|\right) \mathrm{d}x \geq (2\pi\eta)^{\frac{d}{2}} \left(1 + \frac{1}{2d}\right)^{-\frac{d}{2}} \exp\left(\frac{-1}{4}\right).$$

Hence, (4) holds due to the fact that

$$\left(1 + \frac{1}{2d}\right)^{\frac{d}{2}} \leq \exp\left(\frac{1}{4}\right).$$

□

Finally, we are ready to present the complexity of Algorithm 2.

Proposition 2. *Assume* f *is convex and* M*-Lipschitz continuous. If*

$$\eta \leq \frac{1}{16M^2 d}, \tag{5}$$

then the expected number of iterations in the rejection sampling of Algorithm 2 is at most $\sqrt{e}$*.*

Proof. If follows directly from the definition of h_2 that

$$\int_{\mathbb{R}^d} \exp(-h_2(x))\mathrm{d}x = \exp(-f_y^\eta(x^*)) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\eta}\|x-x^*\|^2 - 2M\|x-x^*\|\right)\mathrm{d}x$$

Applying Proposition 1(b) to the above yields

$$\int_{\mathbb{R}^d} \exp(-h_2(x))\mathrm{d}x \geq \exp(-f_y^\eta(x^*))\frac{(2\pi\eta)^{d/2}}{\sqrt{e}}.$$

Note that the condition (3) in Proposition 1 holds thanks to (5). By Lemma 1, the above inequality leads to

$$\int_{\mathbb{R}^d} \exp(-f_y^\eta(x))\mathrm{d}x \geq \int_{\mathbb{R}^d} \exp(-h_2(x))\mathrm{d}x \geq \exp(-f_y^\eta(x^*))\frac{(2\pi\eta)^{d/2}}{\sqrt{e}}. \tag{6}$$

Using the definition of h_1 and Proposition 1(a), we have

$$\int_{\mathbb{R}^d} \exp(-h_1(x))\mathrm{d}x = \exp(-f_y^\eta(x^*))(2\pi\eta)^{d/2}.$$

Using Lemma 2, (6), and the above identity, we conclude that $\mathbb{P}(S) \geq \frac{1}{\sqrt{e}}$, and the expected number of the iterations is $\frac{1}{\mathbb{P}(S)} \leq \sqrt{e}$. □

2 Convergence of ASF

Step 1 of Algorithm 1 is a forward step: starting from $x_k \sim \nu = \pi^X$, the law of y_k is $\pi^Y = \int \pi^X \pi^{Y|X}\mathrm{d}x = \pi^X * \mathcal{N}(0, \eta I)$. Step 2 of Algorithm 1 is a backward step: starting from $y_k \sim \pi^Y$, the law of x_{k+1} is $\pi^X = \int \pi^Y \pi^{X|Y}\mathrm{d}y$.

The perspective that we adopt in the convergence analysis is that: in the forward step, π^Y is obtained by evolving π^X along the heat flow for time η; in the backward step, π^X is obtained by reversing the heat flow starting from π^Y for time η.

Example: time reversal of ODE Consider $\dot{x}_t = b(x_t)$ and $y_t = x_{T-t}$ for $0 \leq t \leq T$. The reversed ODE for $0 \leq t \leq T$ is

$$\dot{y}_t = -\dot{x}_{T-t} = -b(x_{T-t}) = -b(y_t).$$

Now we want to reverse the heat flow $X_t = X_0 + W_t$, $\mathrm{d}X_t = \mathrm{d}W_t$,

$$\frac{\partial \nu_t}{\partial t} = \frac{1}{2}\Delta\nu_t.$$

Define $Y_t = X_{T-t}$ for $0 \leq t \leq T$. The law of Y_t is $Y_t \sim \nu_t^- = \nu_{T-t}$, and hence $\nu_0^- = \nu_T$ and $\nu_T^- = \nu_0$. The backward heat flow for $0 \leq t \leq T$ is

$$\frac{\partial \nu_t^-}{\partial t} = -\frac{\partial \nu_{T-t}}{\partial t} = -\frac{1}{2}\Delta\nu_{T-t} = -\frac{1}{2}\Delta\nu_t^-.$$

It is equivalent to

$$\frac{\partial \nu_t^-}{\partial t} = -\Delta \nu_t + \frac{1}{2}\Delta \nu_t = -\nabla \cdot \left(\nu_t^- \nabla \log \nu_t^-\right) + \frac{1}{2}\Delta \nu_t = -\nabla \cdot \left(\nu_t^- \nabla \log \nu_{T-t}\right) + \frac{1}{2}\Delta \nu_t,$$

which is also the Fokker-Plank equation for the following SDE

$$\mathrm{d}Y_t = \nabla \log \nu_{T-t}(Y_t)\mathrm{d}t + \mathrm{d}W_t. \tag{7}$$

So if $Y_0 \sim \nu_0^- = \nu_T$, then $Y_t \sim \nu_t^- = \nu_{T-t}$ and $Y_T \sim \nu_T^- = \nu_0$.

Now, we consider two simultaneous flows: there are two input distributions $\nu_0 = \pi^X$ and $\rho_0 = \rho_k^X$ (which is the law of x_k) to the Gaussian channel $\pi^{Y|X}$, that is they evolve along the same heat flow, and their corresponding output distributions are $\nu_\eta = \pi^Y$ and $\rho_\eta = \rho_k^Y$. We also have the backward simutaneous heat flow: the outputs of the Gaussian channel are the input distributions $\nu_0^- = \pi^Y$ and $\rho_0^- = \rho_k^Y$ to the RGO, that is they evolve along the same *backward* heat flow, and their corresponding output distributions are $\nu_\eta^- = \pi^X$ and $\rho_\eta = \rho_{k+1}^X$. Note that after simultaneous forward and backward heat flows, ν_0 gets back to itself, but ρ_0 starts from ρ_k^X and ends at $\rho_\eta^- = \rho_{k+1}^X$, which will be shown to be closer to the target ν then ρ_k^X. The process is as in Figure 1.

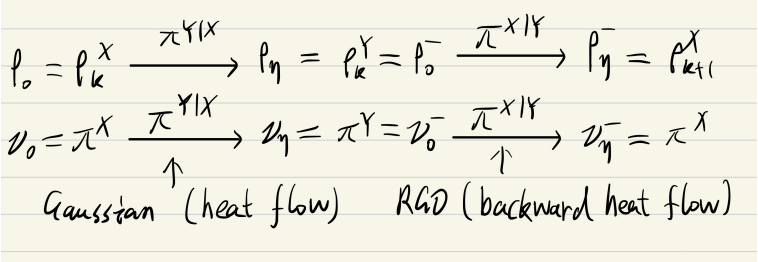

Figure 1: Simutaneous flows

Next, following a similar argument as in the analysis of LMC, we derive the rate of change of KL divergence along simultaneous heat flows, which is a de Bruijn's identity.

Lemma 3. *Let ρ_t and ν_t follow two simultaneous heat flows for $t \geq 0$,*

$$\frac{\partial \rho_t}{\partial t} = \frac{1}{2}\Delta \rho_t\,, \quad \frac{\partial \nu_t}{\partial t} = \frac{1}{2}\Delta \nu_t\,. \tag{8}$$

Then, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \,\|\, \nu_t) = -\frac{1}{2}\mathsf{FI}(\rho_t \,\|\, \nu_t)\,. \tag{9}$$

Proof. Taking the time derivative of $\mathsf{KL}(\rho_t \,\|\, \nu_t)$ and using (8), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \,\|\, \nu_t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int \rho_t \log\frac{\rho_t}{\nu_t}\mathrm{d}x = \int \frac{\partial \rho_t}{\partial t}\log\frac{\rho_t}{\nu_t}\mathrm{d}x + \int \rho_t \frac{\nu_t}{\rho_t}\frac{\partial}{\partial t}\left(\frac{\rho_t}{\nu_t}\right)\mathrm{d}x \\ &= \int \frac{\partial \rho_t}{\partial t}\log\frac{\rho_t}{\nu_t}\mathrm{d}x + \int \nu_t\left(\frac{1}{\nu_t}\frac{\partial \rho_t}{\partial t} - \frac{\rho_t}{\nu_t^2}\frac{\partial \nu_t}{\partial t}\right)\mathrm{d}x \\ &= \int \frac{\partial \rho_t}{\partial t}\log\frac{\rho_t}{\nu_t}\mathrm{d}x - \int \frac{\rho_t}{\nu_t}\frac{\partial \nu_t}{\partial t}\mathrm{d}x \overset{(8)}{=} \frac{1}{2}\int \Delta\rho_t \log\frac{\rho_t}{\nu_t}\mathrm{d}x - \frac{1}{2}\int \frac{\rho_t}{\nu_t}\Delta\nu_t \mathrm{d}x. \end{aligned} \tag{10}$$

Using integration by parts and Fact 1 with $h = \rho_t/\nu_t$, we have

$$\begin{aligned}\int \Delta\rho_t \log\frac{\rho_t}{\nu_t}\mathrm{d}x &= \int \rho_t \Delta\log\frac{\rho_t}{\nu_t}\mathrm{d}x = \int \rho_t\left(\frac{\nu_t}{\rho_t}\Delta\frac{\rho_t}{\nu_t} - \left\|\nabla\log\frac{\rho_t}{\nu_t}\right\|^2\right)\mathrm{d}x \\ &= \int \nu_t \Delta\frac{\rho_t}{\nu_t}\mathrm{d}x - \int \rho_t\left\|\nabla\log\frac{\rho_t}{\nu_t}\right\|^2\mathrm{d}x = \int \frac{\rho_t}{\nu_t}\Delta\nu_t\mathrm{d}x - \mathsf{FI}(\rho_t \,\|\, \nu_t),\end{aligned}$$

where we use integration by parts again in the last identity. The above relation, (10), and Fact 1 imply that (9) holds. □

Fact 1. *For any twice differentiable function h such that $h(x) \neq 0$ for any $x \in \mathbb{R}^d$, we have*

$$\nabla \log h = \frac{\nabla h}{h}, \quad \Delta \log h = \frac{\Delta h}{h} - \|\nabla \log h\|^2 .$$

Indeed, we have that,

$$\Delta \log h = \nabla\cdot(\nabla \log h) = \nabla\cdot\left(\frac{\nabla h}{h}\right) = \frac{\Delta h}{h} - \frac{\langle \nabla h, \nabla h\rangle}{h^2} = \frac{\Delta h}{h} - \left\|\frac{\nabla h}{h}\right\|^2 = \frac{\Delta h}{h} - \|\nabla \log h\|^2 .$$

We also derive the de Bruijn's identity for simultaneous backward heat flows.

Lemma 4. *Consider the backward SDE (7) for the heat equation and let ν_t^- and ρ_t^- denote the laws of Y_t starting from $\nu_0^-(=\pi^Y)$ and $\rho_0^-(=\rho_k^Y)$, respectively. We know*

$$\begin{aligned}\partial_t \nu_t^- &= -\nabla\cdot\left(\nu_t^- \nabla\log\nu_t^-\right) + \frac{1}{2}\Delta\nu_t^- = -\frac{1}{2}\Delta\nu_t^-, \\ \partial_t \rho_t^- &= -\nabla\cdot\left(\rho_t^- \nabla\log\nu_t^-\right) + \frac{1}{2}\Delta\rho_t^- = \nabla\cdot\left(\rho_t^-\nabla\log\frac{\rho_t^-}{\nu_t^-}\right) - \frac{1}{2}\Delta\rho_t^- .\end{aligned}$$

Then, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t^- \,\|\, \nu_t^-) = -\frac{1}{2}\mathsf{FI}(\rho_t^- \,\|\, \nu_t^-) . \tag{11}$$

Proof. Following a similar argument as in the proof of Lemma 3, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t^- \parallel \nu_t^-) &= \frac{\mathrm{d}}{\mathrm{d}t}\int \rho_t^- \log\frac{\rho_t^-}{\nu_t^-}\mathrm{d}x = \int \frac{\partial \rho_t^-}{\partial t}\log\frac{\rho_t^-}{\nu_t^-}\mathrm{d}x + \int \rho_t^- \frac{\nu_t^-}{\rho_t^-}\frac{\partial}{\partial t}\left(\frac{\rho_t^-}{\nu_t^-}\right)\mathrm{d}x \\ &= \int \frac{\partial \rho_t^-}{\partial t}\log\frac{\rho_t^-}{\nu_t^-}\mathrm{d}x + \int \nu_t^-\left(\frac{1}{\nu_t^-}\frac{\partial \rho_t^-}{\partial t} - \frac{\rho_t^-}{(\nu_t^-)^2}\frac{\partial \nu_t^-}{\partial t}\right)\mathrm{d}x \\ &= \int \frac{\partial \rho_t^-}{\partial t}\log\frac{\rho_t^-}{\nu_t^-}\mathrm{d}x + \frac{\partial}{\partial t}\int \rho_t^-\mathrm{d}x - \int \frac{\rho_t^-}{\nu_t^-}\frac{\partial \nu_t^-}{\partial t}\mathrm{d}x \\ &= \int \frac{\partial \rho_t^-}{\partial t}\log\frac{\rho_t^-}{\nu_t^-}\mathrm{d}x - \int \frac{\rho_t^-}{\nu_t^-}\frac{\partial \nu_t^-}{\partial t}\mathrm{d}x,\end{aligned}$$

and then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t^- \parallel \nu_t^-) &= \int \left[\nabla\cdot\left(\rho_t^-\nabla\log\frac{\rho_t^-}{\nu_t^-}\right) - \frac{1}{2}\Delta\rho_t^-\right]\log\frac{\rho_t^-}{\nu_t^-}\mathrm{d}x + \frac{1}{2}\int \frac{\rho_t^-}{\nu_t^-}\Delta\nu_t^-\mathrm{d}x \\ &= -\int \rho_t^-\left\langle\nabla\log\frac{\rho_t^-}{\nu_t^-}, \nabla\log\frac{\rho_t^-}{\nu_t^-}\right\rangle\mathrm{d}x - \frac{1}{2}\int \Delta\rho_t^-\log\frac{\rho_t^-}{\nu_t^-}\mathrm{d}x + \frac{1}{2}\int \frac{\rho_t^-}{\nu_t^-}\Delta\nu_t^-\mathrm{d}x \\ &= -\mathsf{FI}(\rho_t^- \parallel \nu_t^-) + \frac{1}{2}\mathsf{FI}(\rho_t^- \parallel \nu_t^-) = -\frac{1}{2}\mathsf{FI}(\rho_t^- \parallel \nu_t^-).\end{aligned}$$

□

Fact 2. *If ν satisfies α-LSI, then $\nu_t = \nu * \mathcal{N}(0,\eta I)$ satisfies α_t-LSI with $\alpha_t = (1/\alpha + t)^{-1}$ and $\nu_t^- = \nu * \mathcal{N}(0,(\eta) - tI)$ satisfies $\alpha_{\eta-t}$-LSI.*

Theorem 1. *Suppose ν satisfies α-LSI. Along ASF, we have*

$$\mathsf{KL}(\rho_k^X \parallel \pi^Y) \le \frac{\mathsf{KL}(\rho_k^X \parallel \pi^X)}{1+\alpha\eta}, \quad \mathsf{KL}(\rho_{k+1}^X \parallel \pi^X) \le \frac{\mathsf{KL}(\rho_k^Y \parallel \pi^Y)}{1+\alpha\eta}. \tag{12}$$

Therefore, we have

$$\mathsf{KL}(\rho_k^X \parallel \nu) \le \frac{\mathsf{KL}(\rho_0^X \parallel \nu)}{(1+\alpha\eta)^{2k}}.$$

Proof. For the forward step, it follows from Lemma 3 and Fact 2 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \parallel \nu_t) = -\frac{1}{2}\mathsf{FI}(\rho_t \parallel \nu_t) \le -\alpha_t\mathsf{KL}(\rho_t \parallel \nu_t) = -\frac{\alpha}{1+\alpha t}\mathsf{KL}(\rho_t \parallel \nu_t).$$

Integrating gives the first inequality in (12). For the backward step, it follows from Lemma 4 and Fact 2 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t^- \parallel \nu_t^-) = -\frac{1}{2}\mathsf{FI}(\rho_t^- \parallel \nu_t^-) \le -\alpha_{\eta-t}\mathsf{KL}(\rho_t^- \parallel \nu_t^-) = -\frac{\alpha}{1+\alpha(\eta-t)}\mathsf{KL}(\rho_t^- \parallel \nu_t^-).$$

Therefore, just as in the forward step, integration yields the second inequality in (12). □

A channel in information theory is a conditional distribution $P^{Y|X}$ taking input distribution ρ^X and generating output distrinution $\rho^Y = \int \rho^X(x)P^{Y|X}(y|x)\mathrm{d}x$.

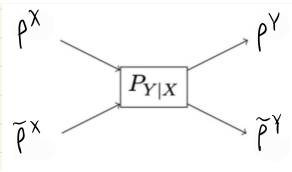

Figure 2: Channel

Lemma 5 (Data processing inequality)**.** *Consider a channel that produces Y given X based on the law $P_{Y|X}$ (see Figure 2). If ρ^Y (resp., $\tilde{\rho}^Y$) is the distribution of Y when X is generated by ρ^X (resp., $\tilde{\rho}^X$), then for any ϕ-divergence $D(\cdot\|\cdot)$ (where ϕ is a convex function),*

$$D(\rho^Y \,\|\, \tilde{\rho}^Y) \le D(\rho^X \,\|\, \tilde{\rho}^X).$$

Proof. By the definition of ϕ-divergence and Jensen's inequality, we have

$$\begin{aligned} D(\rho^X \,\|\, \tilde{\rho}^X) &= \mathbb{E}_{\tilde{\rho}^X}\left[\phi\left(\frac{\rho^X}{\tilde{\rho}^X}\right)\right] = \mathbb{E}_{\tilde{\rho}^{XY}}\left[\phi\left(\frac{\rho^{XY}}{\tilde{\rho}^{XY}}\right)\right] = \mathbb{E}_{\tilde{\rho}^Y}\left[\mathbb{E}_{\tilde{\rho}^{X|Y}}\left[\phi\left(\frac{\rho^{XY}}{\tilde{\rho}^{XY}}\right)\right]\right] \\ &\ge \mathbb{E}_{\tilde{\rho}^Y}\left[\phi\left(\mathbb{E}_{\tilde{\rho}^{X|Y}}\left[\frac{\rho^{XY}}{\tilde{\rho}^{XY}}\right]\right)\right] = \mathbb{E}_{\tilde{\rho}^Y}\left[\phi\left(\mathbb{E}_{\rho^{X|Y}}\left[\frac{\rho^{Y}}{\tilde{\rho}^{Y}}\right]\right)\right] \\ &= \mathbb{E}_{\tilde{\rho}^Y}\left[\phi\left(\frac{\rho^Y}{\tilde{\rho}^Y}\right)\right] = D(\rho^Y \,\|\, \tilde{\rho}^Y). \end{aligned}$$

□

If the channel is deterministic (e.g., $Y = g(X)$ and g is invertible), then "=" holds. If the strict inequality holds, it is called a strong data processing inequality.

In our case, ϕ-divergence is KL divergence, and we have

$$\mathsf{KL}(\rho^Y \,\|\, \tilde{\rho}^Y) \le \mathsf{KL}(\rho^X \,\|\, \tilde{\rho}^X).$$

So Theorem 1 gives a strong data processing inequality with a contraction factor $(1+\alpha\eta)^{-1}$. In the forward step, the channel is Gaussian, and in the backward step, the channel is RGO.

3 LMC as an approximate implementation of ASF

We prove that LMC is indeed an instance of ASF whose implementation of RGO is inexact.

Assume f in the target distribution $\pi \propto \exp(-f)$ is convex and smooth and recall that the iterative step in LMC can be described as

$$y_{k+1} = y_k - \eta\nabla f(y_k) + \sqrt{2\eta}z, \quad z \sim \mathcal{N}(0, I). \tag{13}$$

We claim that the following algorithm gives an equivalent form of LMC (13) from the proximal sampling perspective.

Algorithm 3 Langevin Monte Carlo

1. Sample $y_k \sim \pi^{Y|X}(y \mid x_k) \propto \exp\left(-\frac{1}{2\eta}\|x_k - y\|^2\right)$
2. Sample $x_{k+1} \sim \exp\left(-\frac{1}{2\eta}\|x - y_k + \eta\nabla f(y_k)\|^2\right)$

Indeed, steps 1 and 2 above can be equivalently written as

$$x_{k+1} = y_k - \eta\nabla f(y_k) + \sqrt{\eta}z_k, \quad z_k \sim N(0, I),$$
$$y_{k+1} = x_{k+1} + \sqrt{\eta}z'_k, \quad z'_k \sim \mathcal{N}(0, I),$$

where y_{k+1} is the sample from step 1 in the next iteration. Combining the above identities, we have

$$y_{k+1} = y_k - \eta\nabla f(y_k) + \sqrt{\eta}(z_k + z'_k) \overset{d}{=} y_k - \eta\nabla f(y_k) + \sqrt{2\eta}z, \quad z \sim \mathcal{N}(0, I).$$

Moreover, LMC and ASF share the same step 1, and step 2 of LMC equivalently generates x_{k+1} from $\exp(-h_1(x))$ where

$$h_1(x) := f(y_k) + \langle\nabla f(y_k), x - y_k\rangle + \frac{1}{2\eta}\|x - y_k\|^2. \tag{14}$$

Using the definition of h_1 in (14) and the convexity of f, we have

$$h_1(x) \le f(x) + \frac{1}{2\eta}\|x - y_k\|^2 = f^\eta_{y_k}(x).$$

Note that $f^\eta_{y_k}(x)$ is the potential function of the RGO in step 2 of ASF. Hence, step 2 of LMC can be interpreted as an RGO implementation with the proposal density $\exp(-h_1(x))$ but without rejection. As a result, LMC is an approximate implementation of ASF and thus LMC is biased.

More generally, from the proximal sampling perspective, the basic idea behind LMC (or its inexact RGO implementation) is to approximate $f(x)$ by its linearization $f(y_k) + \langle\nabla f(y_k), x - y_k\rangle$, and this idea has been widely used to implement the proximal mapping of f in optimization methods such as gradient descent and accelerated gradient descent.