Sampling Methods: From MCMC to Generative Modeling
Bayesian learning and Langevin algorithm

Anna Korba

CREST, ENSAE, Institut Polytechnique de Paris

Outline

Motivation for Sampling (1): Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

Motivation for Sampling (1): Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

(1) Let $\mathcal{D} = (w_i, y_i)_{i=1}^p$ a dataset of i.i.d. examples with features w, label y.

(2) Assume an underlying model parametrized by $x \in \mathbb{R}^d$, e.g.:

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{Id}).$$

Motivation for Sampling (1): Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

(1) Let $\mathcal{D} = (w_i, y_i)_{i=1}^p$ a dataset of i.i.d. examples with features w, label y.

(2) Assume an underlying model parametrized by $x \in \mathbb{R}^d$, e.g.:

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{Id}).$$

Step 1. Compute the Likelihood:

$$p(\mathcal{D}|x) \overset{(1)}{\propto} \prod_{i=1}^{p} p(y_i|x, w_i) \overset{(2)}{\propto} \exp(-\frac{1}{2} \sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2).$$

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp(-\frac{\|x\|^2}{2}).$$

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp(-\frac{\|x\|^2}{2}).$$

Step 3. Bayes' rule yields the formula for the posterior distribution over the parameter x:

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p_0(x)}{Z} \quad \text{where} \quad Z = \int_{\mathbb{R}^d} p(\mathcal{D}|x)p_0(x)dx$$

is called the normalization constant and is **intractable**.

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp(-\frac{\|x\|^2}{2}).$$

Step 3. Bayes' rule yields the formula for the posterior distribution over the parameter x:

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p_0(x)}{Z} \quad \text{where} \quad Z = \int_{\mathbb{R}^d} p(\mathcal{D}|x)p_0(x)dx$$

is called the normalization constant and is **intractable**.

Denoting $\pi := p(\cdot|\mathcal{D})$ the posterior on parameters $x \in \mathbb{R}^d$, we have:

$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \frac{1}{2}\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2 + \frac{\|x\|^2}{2}.$$

i.e. π's density is known "up to a normalization constant".
π is a probability distribution over parameters of a model.

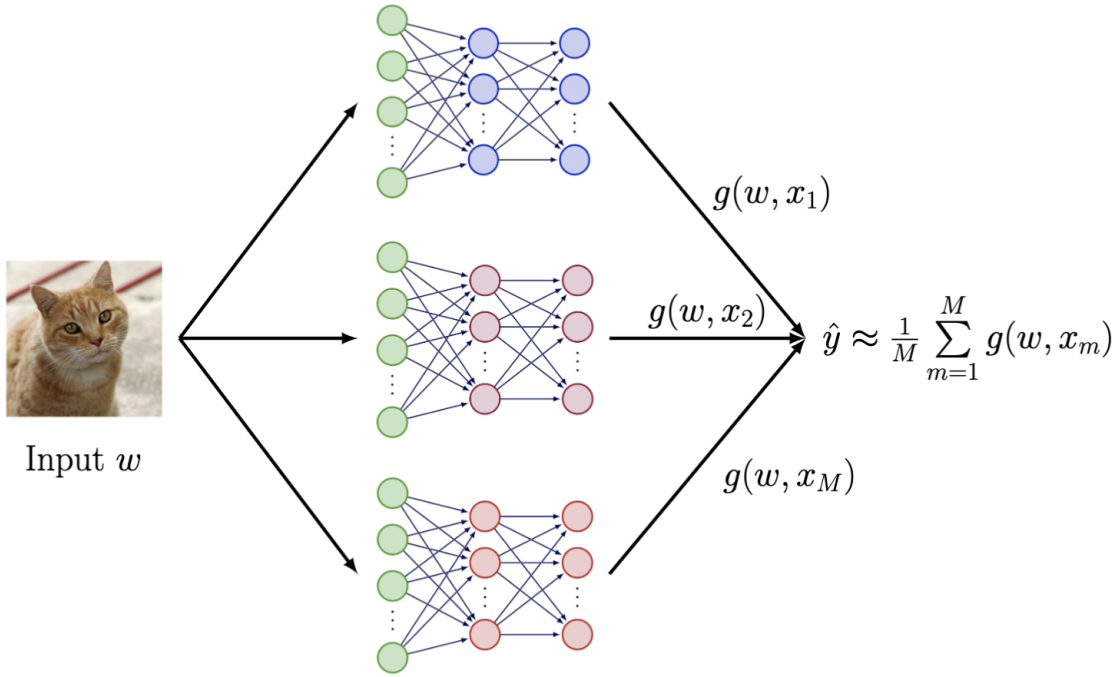
The posterior π is interesting for

- measuring uncertainty on prediction through the distribution of $g(w, \cdot)$, $x \sim \pi$.
- prediction for a new input w:

$$\hat{y} = \underbrace{\int_{\mathbb{R}^d} g(w, x) d\pi(x)}_{\text{"Bayesian model averaging"}}$$

  i.e. predictions of models parametrized by $x \in \mathbb{R}^d$ are reweighted by $\pi(x)$.

Here, Sampling methods construct an approximation $\mu_M = \frac{1}{M}\sum_{m=1}^{M} \delta_{x_m}$ of π.



Sampling as Optimization

Actually, in many cases (e.g. it is underlying many algorithms), the sampling problem (approximating π) can be viewed as optimization over $\mathcal{P}(\mathbb{R}^d)$:

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{D}(\mu|\pi)$$

where D is a divergence or distance, hence that is minimized for $\mu = \pi$.

The Kullback-Leibler divergence

D could be the (reverse) Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(\mu|\pi) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) & \text{if } \mu \ll \pi \\ +\infty & \text{otherwise.} \end{cases}$$

We recognize a f-divergence $\int f\left(\frac{\mu}{\pi}\right) d\pi$ where $f(x) = x\log(x)$. Taking $f(x) = -\log(x)$ yields the (forward) KL i.e. $\mathrm{KL}(\pi|\mu)$.

The Kullback-Leibler divergence

D could be the (reverse) Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(\mu|\pi) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) & \text{if } \mu \ll \pi \\ +\infty & \text{otherwise.} \end{cases}$$

We recognize a f-divergence $\int f\left(\frac{\mu}{\pi}\right) d\pi$ where $f(x) = x\log(x)$. Taking $f(x) = -\log(x)$ yields the (forward) KL i.e. $\mathrm{KL}(\pi|\mu)$.

The (reverse) KL as an objective is convenient when the unnormalized density of π is known since it **does not depend on the normalization constant!**

Indeed writing $\pi(x) = e^{-V(x)}/Z$ we have:

$$\mathrm{KL}(\mu|\pi) = \int_{\mathbb{R}^d} \log\left(\frac{\mu}{e^{-V}}(x)\right) d\mu(x) + \log(Z).$$

But, it is not convenient when μ or π are discrete, because the KL is $+\infty$ unless $supp(\mu) \subset supp(\pi)$.

Examples with parametric models

Consider the sampling optimization objective:

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{D}(\mu|\pi)$$

But now (in this slide) assume we restrict the search space to a parametric families $\{\mu_\theta,\ \theta \in \mathbb{R}^p\}$ (ex: Gaussian with diagonal covariance matrices can be parametrized by $\theta = (m, \sigma) \in \mathbb{R}^{2d}$). The problem rewrites as a finite-dimensional optimization problem (i.e. over $\mathbb{R}^p$):

$$\min_{\theta \in \mathbb{R}^p} \mathrm{D}(\mu_\theta|\pi)$$

- Choosing D as the reverse KL, i.e. $\mathrm{D}(\mu_\theta|\pi) = \mathrm{KL}(\mu_\theta|\pi)$ yields Variational Inference [Blei et al. (2017)] which is useful for Bayesian Inference ($\pi \propto e^{-V}$)
- Here, we could also use normalizing flows to construct a family $\mu_\theta = f_{\theta\#}p$ and optimize the previous objective[1].

[1] Rezende, D., Mohamed, S. (2015, June). Variational inference with normalizing flows. In International conference on machine learning.

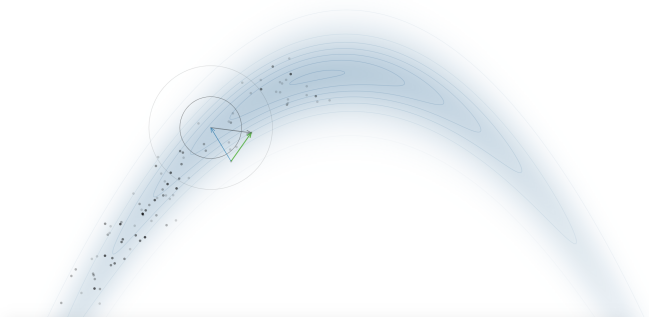
Next we will focus on a non-parametric scheme.

Langevin Monte Carlo

(1) Markov Chain Monte Carlo (MCMC) methods: generate a Markov chain in $\mathbb{R}^d$ whose law converges to $\pi \propto \exp(-V)$

Example: Langevin Monte Carlo (LMC) [Roberts and Tweedie (1996)]

$$x_{m+1} = x_m + \gamma \nabla \log \pi(x_m) + \sqrt{2\gamma}\eta_m, \quad \eta_m \sim \mathcal{N}(0, \mathrm{Id}).$$



Picture from `https://chi-feng.github.io/mcmc-demo/app.html`.

Note that in the Bayesian inference setting, where $\pi = \frac{\exp(-V)}{Z}$, it is easily implementable since the **score** $\nabla_x \log \pi(x) = -\nabla_x (V(x) + \log(Z)) = -\nabla V(x)$ since $\nabla_x \log(Z) = 0$.

Outline

Langevin diffusion

Langevin diffusion is the Stochastic Differential Equation (SDE):

$$\mathrm{d}x_t = -\nabla V(x_t)dt + \sqrt{2}\mathrm{d}B_t, \quad x_t \sim p_t$$

where B_t denotes the standard Brownian motion in $\mathbb{R}^d$, defined as:

- $B_0 = 0$ almost surely;
- For any $t_0 < t_1 < \cdots < t_N$, the increments $B_{t_n} - B_{t_{n-1}}$ are independent, $n = 1, 2, \ldots, N$;
- The difference $B_t - B_s$ and B_{t-s} have the same distribution: $\mathcal{N}(0, (t-s)\,\mathrm{Id})$ for $s < t$;
- B_t is continuous almost surely.

Langevin diffusion defines a *Markov process* as follows:

$$x_t = x_0 - \int_0^t \nabla V(x_s)ds + \sqrt{2}B_t,$$

where θ_0 is some initialization.

(discrete time) **Langevin Monte Carlo** (LMC) or **Unadjusted Langevin Algorithm** (ULA) (Roberts and Tweedie, 1996)

$$x_{t+1} = x_t - \gamma \nabla V(x_t) + \sqrt{2\gamma}\eta_t, \quad \eta_t \sim \mathcal{N}(0, \mathrm{Id}). \tag{1}$$

It's the Euler-Maruyama time-discretization that is obtained as follows:

$$\begin{aligned} x_\gamma &\approx x_0 - \int_0^\gamma \nabla V(x_0)\, dt + \sqrt{2\gamma}\, \eta \\ &= x_0 - \left(\int_0^\gamma dt \right) \nabla V(x_0) + \sqrt{2\gamma}\, \eta \\ &= x_0 - \gamma \nabla V(x_0) + \sqrt{2\gamma}\, \eta. \end{aligned}$$

We can now iterate this approach k times, which gives us a recursion, which can be easily implementable on a computer:

$$x_{k\gamma} \approx x_{(k-1)\gamma} - \gamma \nabla V(x_{(k-1)\gamma}) + \sqrt{2\gamma}\, \eta_k,$$

where $\eta_k \sim \mathcal{N}(0, \mathrm{Id})$ for all k. Dropping the dependency on γ in the indices yields the scheme(1).

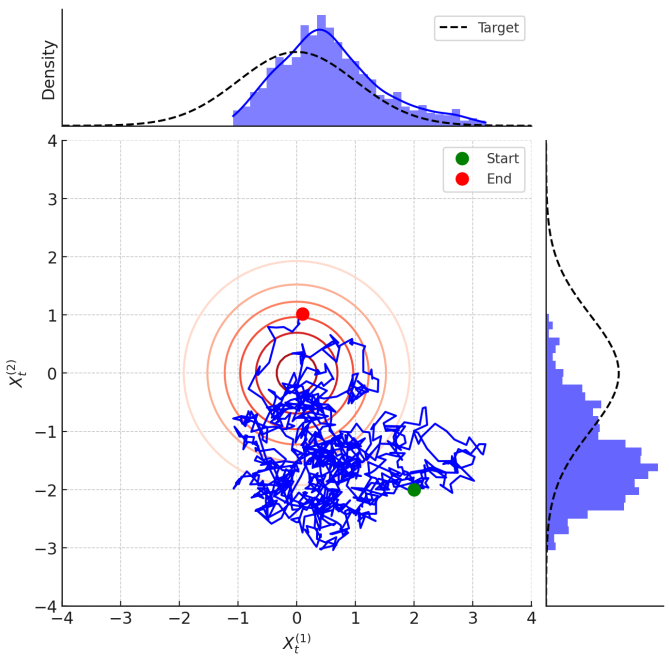
Ornstein-Uhlenbeck

Example: $\pi \propto \exp(-\frac{\|x\|^2}{2})$, $\log \pi(x) = -V(x) = -\frac{\|x\|^2}{2}$, $\nabla \log \pi - x$.

(continuous time) **Langevin diffusion** = Ornstein-Uhlenbeck process:

$$\mathrm{d}x_t = -x_t + \mathrm{d}B_t.$$

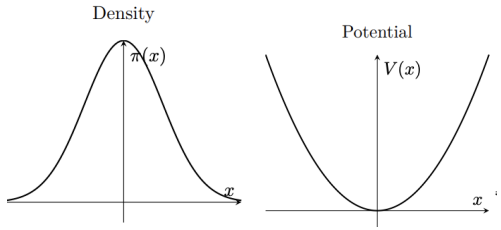
(discrete time) $x_{t+1} = x_t - \gamma x_t + \sqrt{2\gamma}\eta_t, \quad \eta_t \sim \mathcal{N}(0, \mathrm{Id})$.



Recall above we plot $x_{t+1} = x_t + \gamma \nabla \log \pi(x_t) + \sqrt{2\gamma}\eta_t$ for $\pi \propto \exp(-\frac{\|x\|^2}{2})$.

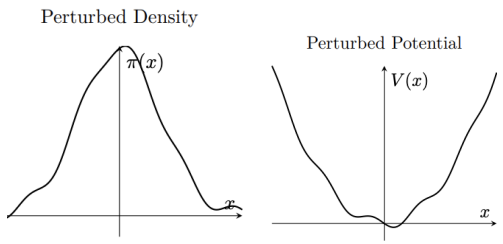
When does Langevin diffusion's law converges (fast) to π?

- Consider a standard Gaussian distribution $\pi(x) \propto \exp(-\frac{\|x\|^2}{2})$, i.e. $\pi \propto \exp(-V)$ with V 1-strongly convex, i.e. π is (1-)strongly log-concave.



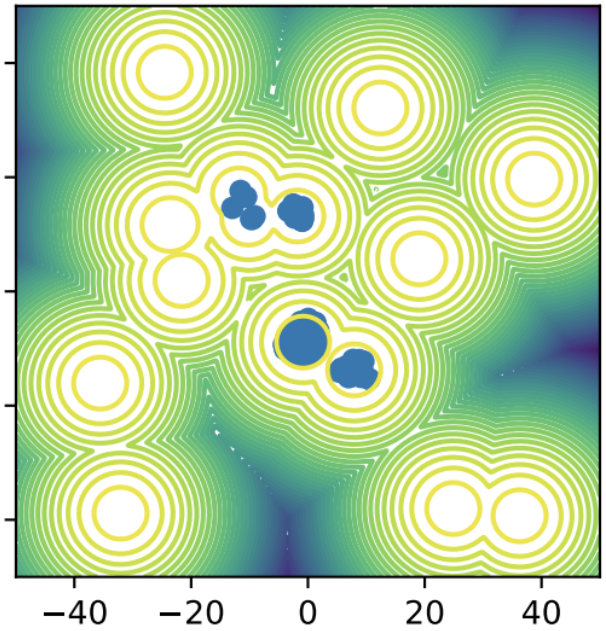
Then $\mathrm{KL}(p_t|\pi) = \exp(-2t)\,\mathrm{KL}(p_0|\pi)$.

- If π is a perturbation of a strongly-log-concave distribution, then the rate degrades with the size of the perturbation.



(see Holley–Stroock theorem and log-Sobolev inequalities, (Bakry et al., 2014)).

Langevin in the multimodal case



Mixture of equally weighted 16 Gaussians with unit variance and uniformly chosen centers in $[-40, 40]^2$, a standard sampling benchmark. ULA was initialized with $\mathcal{N}(0, I_2)$, step-size $h = 0.01$. ULA was run with 5.10^4 steps (one minute run).

The Fokker-Planck equation

For simplicity, let us assume $d = 1$, so that Langevin diffusion becomes:

$$\mathrm{d}x_t = -\partial_x V(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

To understand how $p(x, t)$ evolves, we will use the Fokker–Planck equation, which governs the evolution of $p(x, t)$ through the following partial differential equation (PDE):

$$\partial_t p(x, t) = \partial_x \left[\partial_x V(x) p(x, t)\right] + \partial_x^2 p(x, t).$$

This equation characterizes how the "change" in $p(\cdot, t)$ behaves, i.e., $\partial_t p(x, t)$.

The Fokker-Planck equation

Now, the idea is: if $p(\cdot, t)$ converges to a distribution as $t \to \infty$, then whenever this limit is reached, there should not be any more changes in p. In other words, whenever $p(\cdot, t)$ hits its limit, $\partial_t p(x, t)$ has to be equal to 0.

Therefore, we can simply "check" if π is a limit of $p(\cdot, t)$ by replacing $p(x, t)$ with $\pi(x)$ in the Fokker–Planck equation and observing whether the right-hand side is equal to 0 or not. Let us apply this procedure:

$$\begin{aligned}\partial_x \left[\partial_x V(x)\pi(x)\right] + \partial_x^2 \pi(x) &= \partial_x \left[\partial_x V(x)\pi(x) + \partial_x \pi(x)\right] \\ &= \partial_x \left[\partial_x V(x)\pi(x) - \partial_x V(x)\pi(x)\right] \\ &= 0,\end{aligned}$$

where we used the fact that

$$\partial_x V(x) = -\partial_x \log \pi(x) = -\frac{1}{\pi(x)} \partial_x \pi(x),$$

hence

$$\partial_x \pi(x) = -\pi(x) \partial_x V(x).$$

Conclusion: π is an equilibrium for the FP equation !

References I

Bakry, D., Gentil, I., Ledoux, M., et al. (2014). *Analysis and geometry of Markov diffusion operators*, volume 103. Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.