Sampling Methods: From MCMC to Generative Modeling
Introduction and Reminders

Anna Korba

CREST, ENSAE, Institut Polytechnique de Paris

Self-introduction

- Graduated from ENSAE and ENS Cachan (MVA) in 2015
- PhD in Machine Learning at Telecom Paris Tech (2015-2018)
- Postdoc at University College London (2018-2020)
- Assistant Prof at ENSAE since 2020
- Active research in ML community, with a theoretical flavor. Attend & publish regularly in NeurIPS & ICML.
- Main interests: sampling, optimal transport, kernel methods.

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:

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

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

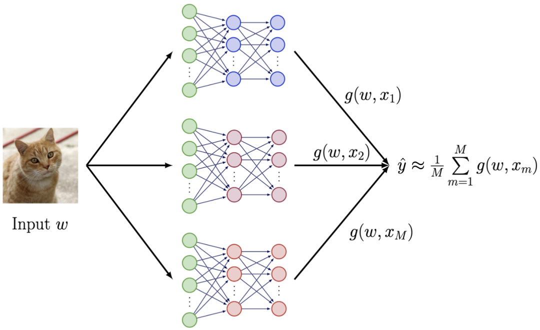
The posterior π is interesting for

- 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.

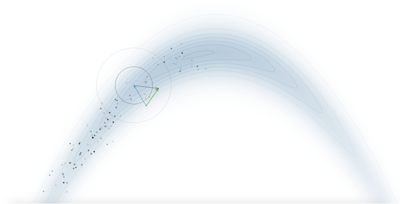


(Some, Non parametric) Sampling methods

(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)]

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



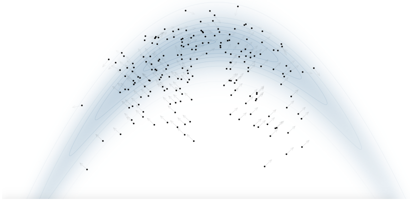
Picture from `https://chi-feng.github.io/mcmc-demo/app.html`.

(2) Interacting particle systems, whose empirical measure at stationarity approximates $\pi \propto \exp(-V)$

Example: Stein Variational Gradient Descent (SVGD) [Liu and Wang (2016)]

$$x^i_{m+1} = x^i_m - \frac{\gamma}{N}\sum_{j=1}^{N} \nabla V(x^j_m)k(x^i_m, x^j_m) - \nabla_2 k(x^i_m, x^j_m), \quad i = 1, \dots, N,$$

where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is a kernel (e.g. $k(x, y) = \exp(-\|x - y\|^2)$).

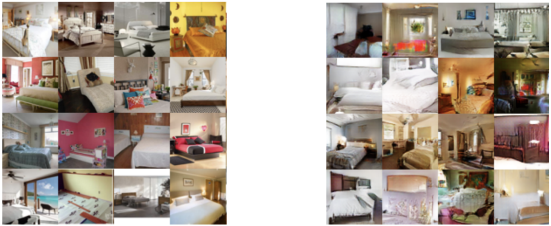


Picture from `https://chi-feng.github.io/mcmc-demo/app.html`.

Motivation for Sampling (2): Generative modeling

In this setting, we have a collection of samples (data) $x_1, \ldots, x_n \sim \pi$.

Goal of Generative Modeling: generate new samples that look like π.



LSUN bedroom samples vs MMD GAN [Li et al. (2017)].

The Sampling literature

Two different settings:

- **(1)** the "Bayesian inference" one, where $\pi \propto e^{-V}$
- **(2)** the "Generative Modeling" one, where $x_1, \ldots, x_n \sim \pi$

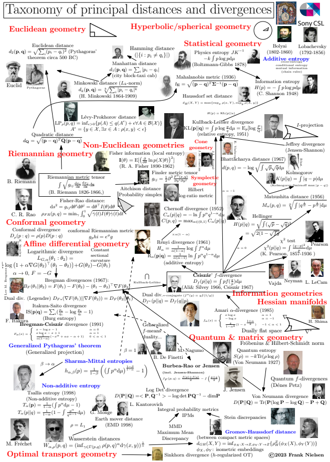
For **(1)**, you may have heard of: Importance Sampling, MCMC algorithms ...

For **(2)**, you may have heard of: Generative Adversarial Networks, Normalizing Flows, Diffusion Models...

There is no clear winner on the quality of approximation/computational complexity. Also, these methods are nowadays sometimes used jointly.

Assume that we have an algorithm that outputs a candidate probability distribution μ, we want to know how close it is from π.

One way is to pick a distance or divergence between probability distributions.



From https://franknielsen.github.io/.

Main families of divergences and distances

Let $\mathcal{P}(\mathbb{R}^d)$ denote the space of probability distributions over $\mathbb{R}^d$.

We will pick D a divergence, i.e. s.t. $\mathrm{D}(\mu||\pi) \geq 0$ for any $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\mathrm{D}(\mu||\pi) \Leftrightarrow \mu = \pi$; or a distance (i.e. satisfies triangle inequality).

Main families of divergences and distances are:

- f-divergences:
$$\int f\left(\frac{\mu}{\pi}\right) d\pi, \quad f \text{ convex, } f(1) = 0$$
defined for $\mu \ll \pi$ (μ absolutely continuous w.r.t. π)

- integral probability metrics (IPM):
$$\sup_{f \in \mathcal{G}} \left| \int f d\mu - \int f d\pi \right|$$
for $\mathcal{G}$ a class of functions "rich enough"

- optimal transport (OT) distances (cf Marco Cuturi or Austin Stromme's courses)

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*(π).

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 $\{P_\theta,\ \theta \in \mathbb{R}^p\}$ (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 ($\pi \propto e^{-V}$)
- Choosing D as the forward KL, i.e. $\mathrm{D}(\mu_\theta|\pi) = \mathrm{KL}(\pi|\mu_\theta)$ yields Maximum Likelihood, which is useful for fitting a model ($x_1, \dots, x_n \sim \pi$) since:

$$\min_\theta \mathrm{KL}(\pi|\mu_\theta) = \int \log\left(\frac{\pi}{\mu_\theta}\right) d\pi \Leftrightarrow \min_\theta - \int \log(\mu_\theta(x)) d\pi(x) \approx \frac{1}{n}\sum_{i=1}^n \log(\mu_\theta(x_i))$$

The Maximum Mean Discrepancy

When we have π (or an approximation) as a discrete measure, it is convenient to choose D as an IPM, i.e. integral probability metric (to approximate integrals).

For instance, D could be the MMD (Maximum Mean Discrepancy):

$$\begin{aligned}\mathrm{MMD}^2(\mu,\pi) &= \sup_{f\in\mathrm{H}_k,\|f\|_{\mathcal{H}_k}\le 1}\left|\int f d\mu - \int f d\pi\right| \\ &= \|m_\mu - m_\pi\|^2_{\mathrm{H}_k}, \quad \text{where } m_\mu = \int k(x,\cdot)d\mu(x) \\ &= \iint_{\mathbb{R}^d} k(x,y)d\mu(x)d\mu(y) \\ &\quad + \iint_{\mathbb{R}^d} k(x,y)d\pi(x)d\pi(y) - 2\iint_{\mathbb{R}^d} k(x,y)d\mu(x)d\pi(y).\end{aligned}$$

where $k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ is a p.s.d. kernel (e.g. $k(x,y)=e^{-\|x-y\|^2}$) and H_k is the RKHS associated to k:

$$\mathrm{H}_k = \overline{\left\{\sum_{i=1}^m \alpha_i k(\cdot,x_i);\ m\in\mathbb{N};\ \alpha_1,\dots,\alpha_m\in\mathbb{R};\ x_1,\dots,x_m\in\mathbb{R}^d\right\}}.$$

Example: Take $k(x,y) = e^{-\frac{\|x-y\|^2}{\sigma}}$, $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x^i}$, $\pi = \frac{1}{m}\sum_{j=1}^{m} \delta_{y^j}$.

$$\mathrm{MMD}^2(\mu,\pi) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x^i,x^j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(y^i,y^j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x^i,y^j).$$

Remark: scale carefully the bandwith σ. Or consider $k(x,y) = -\|x-y\|$, which is not p.s.d. but does not have scale issue, and the corresponding MMD is Energy Distance.

Wasserstein-p distances

Let $\mathcal{P}_p(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ with finite p moments, i.e. $\mathcal{P}_p(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int \|x\|^p d\mu(x) < \infty\}$.

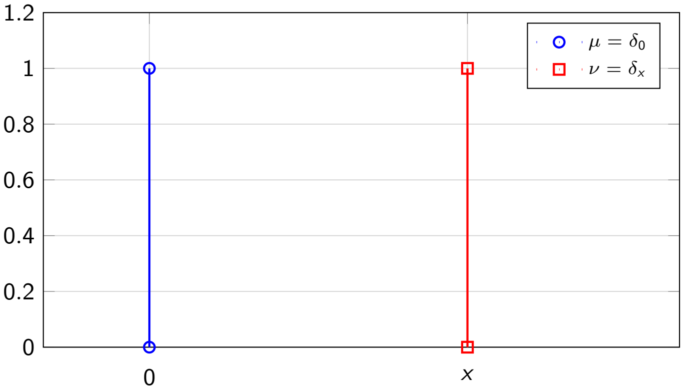
The Wasserstein-p distance from Optimal transport is defined as :

$$\forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d), \quad W_p^p(\mu, \nu) = \inf_{s \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, ds(x, y),$$

where $\Gamma(\mu, \nu)$ is the set of possible couplings between μ and ν (probabilities on $\mathbb{R}^d \times \mathbb{R}^d$ with first and second marginal equal to μ and ν). Most popular ones are:

- The W_2 (in many ways analog to an "euclidean distance" but on $\mathcal{P}_2(\mathbb{R}^d)$)
- The W_1, which interestingly can be written as an IPM:

$$W_1(\mu, \nu) = \inf_{f:\mathbb{R}^d \to \mathbb{R},\, f \text{ is 1-Lipschitz}} \left| \int f d\mu - \int f d\nu \right|$$



We have $\mathrm{KL}(\mu|\nu) = \mathrm{KL}(\nu|\mu) = +\infty$, and $W_2(\mu, \nu) = |x|$.

Outline

In this section we will recall some fundamental methods and principles from Simulation and Monte Carlo (see Nicolas Chopin's course).

We will consider the "Bayesian inference" setting, where the target π has a density that is known to be $\pi \propto e^{-V}$.

Recall that in this setting we are often interested in approximating:

$$\int f(x) d\pi(x) \quad \text{for some } f.$$

Importance Sampling (IS)

Let q be a proposal distribution such that $\mathrm{Supp}(\pi) \subset \mathrm{Supp}(q)$. Define for all $x \in \mathbb{R}^d$

$$w(x) = \frac{\pi(x)}{q(x)}$$

Define the Self-Normalized Importance Sampling (SNIS) estimator of the expectation of f as

$$\int f d\pi \approx \sum_{i=1}^{N} w_N^i f(X_i), \quad \text{where } w_N^i = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j)}$$

and $X_1, \ldots, X_N \sim q$.

Importance Sampling (IS)

Let q be a proposal distribution such that $\mathrm{Supp}(\pi) \subset \mathrm{Supp}(q)$. Define for all $x \in \mathbb{R}^d$

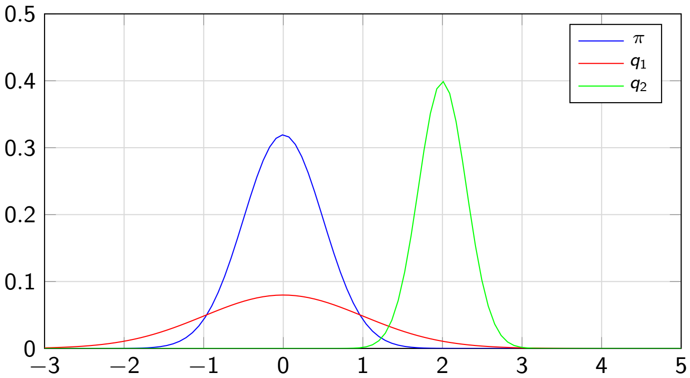
$$w(x) = \frac{\pi(x)}{q(x)}$$

Define the Self-Normalized Importance Sampling (SNIS) estimator of the expectation of f as

$$\int f d\pi \approx \sum_{i=1}^{N} w_N^i f(X_i), \quad \text{where } w_N^i = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j)}$$

and $X_1, \ldots, X_N \sim q$.

Remark: For IS to be effective, the proposal q must be close enough to π in χ^2-square distance (see Agapiou et al. (2017, Th1)), which makes IS also notably affected by the curse of dimensionality (e.g., Agapiou et al. (2017, Sec 2.4.1)).



- Designing a good proposal q is critical
- There is a huge literature on Adaptive Importance Sampling

Recall that a Markov kernel $Q(x, dy)$ is an application $\mathbb{R}^d \to \mathcal{P}(\mathbb{R}^d)$.

Let $Q(x, dy)$ a Markov kernel, such that $Q(x, dy) = q(x, y)dy$.

Metropolis-Hastings is a two-step iterative algorithm relying on the proposal Markov kernel Q.

Let x_m be the state at time m.

- **Step 1**: Sample a candidate $y \sim Q(x_m, dy)$
- **Step 2**: The next state is set according to the rule:

$$x_{m+1} = \begin{cases} y & \text{with probability } \mathrm{acc}(x_m, y) \\ x_m & \text{with probability } 1 - \mathrm{acc}(x_m, y) \end{cases}$$

where the acceptance probability is

$$\mathrm{acc}(x_m, y) = \min\left(1, \frac{q(y, x_m)\pi(y)}{q(x_m, y)\pi(x_m)}\right)$$

Examples of Markov kernels

- Gaussian random walk
$$y \sim \mathcal{N}(x, \Sigma)$$
- Langevin proposal (yields "MALA" i.e. Metropolis Adjusted Langevin Algorithm)
$$y \sim \mathcal{N}(x + \gamma \nabla \log \pi(x), 2\gamma \, \mathrm{Id})$$
Recall that if $\pi \propto e^{-V}$, i.e. $\pi = \tilde{\pi}/Z$ where $\tilde{\pi}$ is known and Z unknown, then $\nabla \log(\pi) = \frac{\nabla(\tilde{\pi}/Z)}{\tilde{\pi}/Z} = \nabla \log \tilde{\pi} = -\nabla V$.

References I

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431.

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.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems*, pages 2378–2386.

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