Efficient Sampling Techniques

Alexey Naumov

HDI Lab,

HSE University

June 20-25, 2022

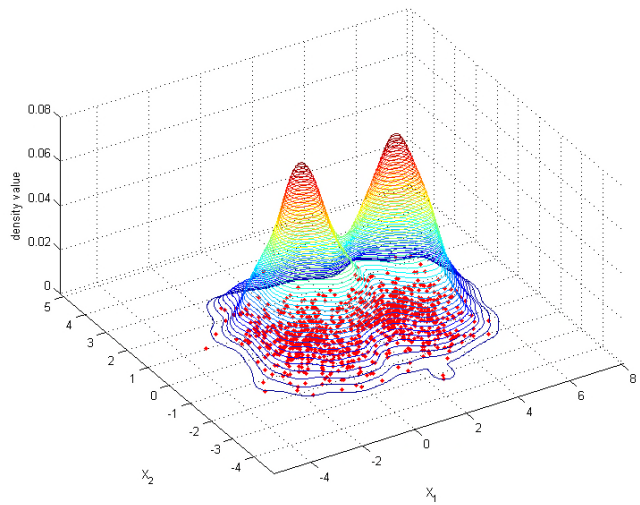
Summary

Density estimation

- Classical statistical problem:
  1. We have a sample $X_1, \ldots, X_n \in \mathbb{R}^d$ from a density $p_{\text{data}}(x)$.
  2. Aim: estimate $p_{\text{data}}(x)$ and sample from it
- Classical solution: kernel density estimation

$$\pi(x) = \frac{1}{n} \sum_{j=1}^{n} K_h(X_j - x),$$

where K_h – kernel, h – bandwidth.



- This approach work when $d = 1, 2, 3$.

Density estimation

- High dimension $d > 3$.
- Black and white pictures 1024×1024 pixels, dim $d = 2^{20} > 10^6$.
- Other object of interest: video, protein structure, ...
- We need other methods (e.g. GANs)
- How to sample from π?

Motivation

- Bayesian inference and learning. Let $\theta \in \Theta$ be an unknown variable (parameter) and $\mathbf{X} = (X_1, \ldots, X_N) \in \mathsf{X}$ be a data.
  1. Posterior distribution: given the prior $\mathrm{p}_0(\theta)$ and likelihood $p(X_i|\theta)$

$$\pi(\theta|\mathbf{X}) = \frac{\prod_{i=1}^{N} p(X_i|\theta)\,\mathrm{p}_0(\theta)}{\int\limits_{\Theta} \prod_{i=1}^{N} p(X_i|\theta)\,\mathrm{p}_0(\theta)\mathrm{d}\theta}$$

  2. Expectation w.r.t. $\pi(\theta|\mathbf{X})$

$$\mathbb{E}_{\pi(\cdot|\mathbf{X})}[f(\theta)] = \int_{\Theta} f(\theta)\pi(\theta|\mathbf{X})\mathrm{d}\theta$$

- Statistical mechanics. Here, one needs to compute the partition function Z of a system with states s and Hamiltonian $E(s)$

$$Z = \sum_{s} \exp\left\{-\frac{E(s)}{kT}\right\},$$

where k is the Boltzmann's constant and T denotes the temperature of the system.

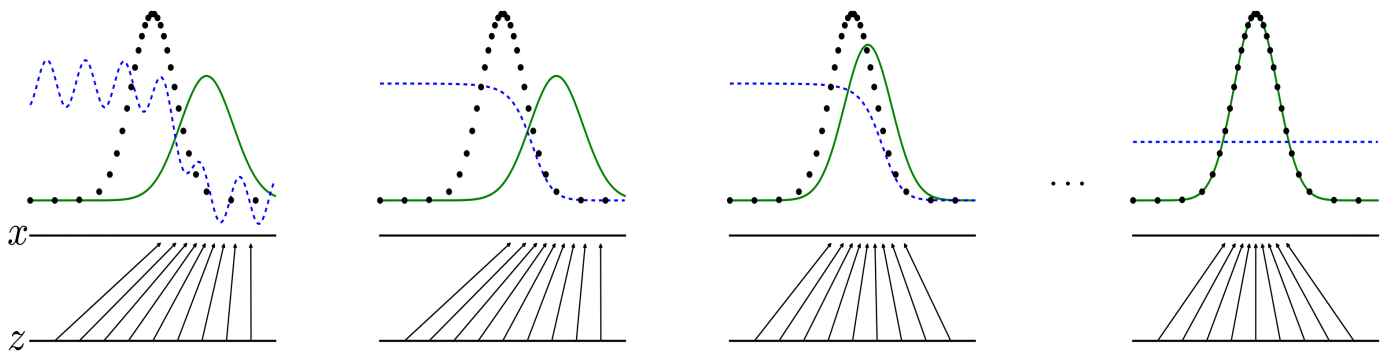
GANs framework

- Generator $G : \mathbb{R}^d \mapsto \mathbb{R}^D$: takes a latent variable z from a prior density $p_0(z)$, $z \in \mathbb{R}^d$, produces $G(z) \in \mathbb{R}^D$ in the observation space;
- Discriminator $D : \mathbb{R}^D \mapsto [0, 1]$: takes a sample in the observation space, distinguishes between real examples and fake ones;

GAN training objective

$$L(g, D) := \mathbb{E}_{X \sim p_{\text{data}}}[\log(D(X))] + \mathbb{E}_{Z \sim p_0}[\log(1 - D(g(Z)))] \to \min_{g \in \mathcal{G}} \max_{D \in \mathcal{D}}.$$

- Let $p_d(x)$ and $p_g(x)$ be the densities of real and fake observations;



-

$$\text{Optimal discriminator: } D^\star(x) = \frac{p_d(x)}{p_d(x) + p_g(x)} \tag{1}$$

GANs as an energy-based model

- Main drawback: information accumulated by discriminator is not used during the generation procedure;
- Let $d^\star(x) = \operatorname{logit} D^\star(x)$, therefore:

$$\frac{p_d(x)}{p_d(x) + p_g(x)} = \frac{1}{1 + \frac{p_g(x)}{p_d(x)}} = \frac{1}{1 + \exp(-d^\star(x))}$$

  Hence, we can express

$$p_d(x) = p_g(x) e^{d^\star(x)} .$$

- Let us introduce $d(x) = \operatorname{logit} D(x)$ and consider the corresponding energy-based model

$$\hat{p}_d(x) = p_g(x) e^{d(x)} / Z_0 ,$$

  where Z_0 is the normalizing constant. If $D(x) \approx D^\star(x)$, $\hat{p}_d(x)$ is close to $p_d(x)$;
- Sample from $\hat{p}_d(x)$ using MCMC.

GANs as an energy-based model

- Similar idea considered in Turner et al. [2019]; main issue: MCMC in pixel space is highly inefficient;
- Che et al. [2020] suggested latent-space sampling from the model

$$\hat{p}_d(x) = p_0(z) \exp\left\{\text{logit}\big(D(G(z))\right\}, z \in \mathbb{R}^d,$$

  where $p_0(z)$ is the generator's prior distribution in the latent space;
- Sampling using Langevin-based algorithms, as suggested in Che et al. [2020], can be inefficient, especially if d is large.

Introduction

This Course

We aim at sampling from π and computing expectation

$$\pi(f) := \mathbb{E}[f(X)] = \int_{\mathsf{X}} f(x)\pi(x)\,\mathrm{d}x, \quad f \in \mathsf{L}_2(\pi)$$

We discuss,

- Monte-Carlo method
- Rejection sampling
- Importance sampling
- MCMC
- Mixture of techniques

Monte-Carlo method

- Get an i.i.d. sample $(X_k)_{k=0}^{\infty}$ from π, estimate $\pi(f)$ by

$$\pi_n(f) := \frac{1}{n}\sum_{k=0}^{n-1} f(X_k),$$

- Kolmogorov's strong law of large numbers: with probability 1

$$\lim_{n\to\infty} \pi_n(f) = \mathbb{E}[f(X_0)] = \pi(f)$$

- Advantage over deterministic integration: MC positions the integration grid (samples) in regions of high probability.
- Disadvantage: when $\pi(x)$ has standard form, e.g. Gaussian, it is straightforward to sample from it using easily available routines. However, when this is not the case, we need to introduce more sophisticated techniques.

Monte-Carlo method

- Variance:

$$\mathrm{Var}[\pi_n(f)] = \frac{1}{n^2}\sum_{k=0}^{n-1}\mathrm{Var}[f(X_k)] = \frac{\sigma_\pi^2(f)}{n}$$

  where $\sigma_\pi^2(f) = \mathrm{Var}[f(X_0] = \pi(f^2) - \pi^2(f)$.

- Central limit theorem (CLT)

$$\sqrt{n}(\pi_n(f) - \pi(f)) \xrightarrow{Law} \mathrm{N}(0, \sigma_\pi^2(\mathrm{f})) \quad \mathrm{n} \to \infty$$

  Indeed,

$$\sqrt{n}(\pi_n(f) - \pi(f)) = \frac{\sum_{k=0}^{n-1}(f(X_k) - \mathbb{E}[f(X_k)])}{\sqrt{n}}$$

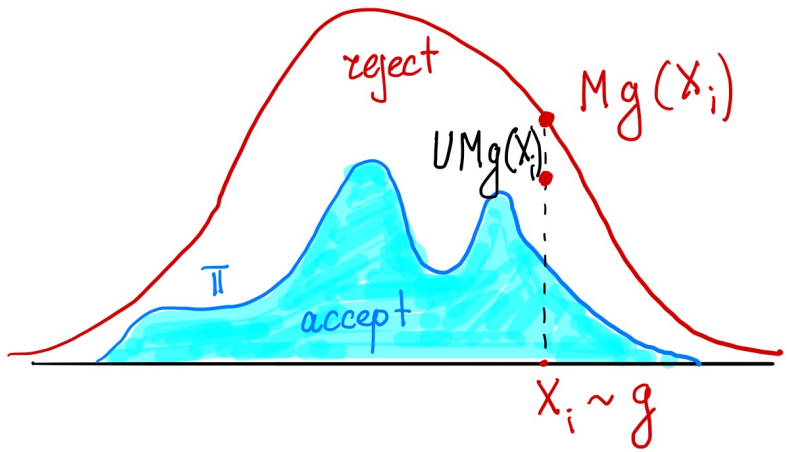
- Length of confidence interval for $\pi_n(f)$ proportional to $\frac{\sigma_\pi(f)}{\sqrt{n}}$

Rejection sampling

- Sample from a distribution π, which is known up to a proportionality constant, by sampling from another easy-to-sample proposal distribution g that satisfies $\pi(x) \leq Mg(x), M < \infty$.
- Algorithm:
  Set $k = 0$;
  Repeat until $k = n - 1$
  1. Sample $X_i \sim q$ and independent $U \sim \text{Uniform}[0, 1]$;
  2. Accept X_i and set $i := i + 1$, if

$$U < \frac{\pi(X_i)}{Mg(X_i)}.$$

  Otherwise, reject.



Rejection sampling

- Advantage: simple
- Disadvantage: impractical in high-dimensional scenarios.
  It is not always possible to bound $\pi(x)/g(x)$ with a reasonable constant M over the whole space X. If M is too large,

$$\begin{aligned}\mathrm{P}(X_i \text{ accepted}) &= \mathrm{P}\left(U < \frac{\pi(X_i)}{Mg(X_i)}\right) = \mathbb{E}\left[\mathrm{P}\left(U < \frac{\pi(X_i)}{Mg(X_i)}\right)\middle| X_i\right] \\ &= \mathbb{E}\left[\frac{\pi(X_i)}{Mg(X_i)}\right] = \int_{\mathrm{X}} \frac{\pi(x)}{Mg(x)} g(x)\mathrm{d}x = \frac{1}{M}\end{aligned}$$

  will be too small (here we also assume $g(x) > 0, x \in \mathrm{X}$)

Rejection sampling

We show that

$$\mathrm{P}\left(X_i \le x | U < \frac{\pi(X_i)}{Mg(X_i)}\right) = \pi\{(-\infty, x]\}$$

Indeed, let $A = \{X_i \le x\}, B = \left\{U < \frac{\pi(X_i)}{Mg(X_i)}\right\}$. Then

$$\mathrm{P}(A|B) = \mathrm{P}(B|A)\,\mathrm{P}(A)/\,\mathrm{P}(B).$$

We may check that

$$\begin{aligned}
\mathrm{P}(B|A) &= \frac{\mathrm{P}(A \cap B)}{G(x)} = \frac{1}{G(x)}\,\mathbb{E}[\mathbb{1}_{A\cap B}] \\
&= \frac{1}{G(x)}\,\mathbb{E}_{X_i}[\mathbb{1}_A]\,\mathbb{E}_U[\mathbb{1}_B] = \frac{1}{MG(x)}\,\mathbb{E}_{X_i}\left[\mathbb{1}_A \frac{\pi(X_i)}{g(X_i)}\right] \\
&= \frac{\pi\{(-\infty, x]\}}{MG(x)}.
\end{aligned}$$

Importance sampling

- Make change of measure: replace $\pi(x)$ by another easy-to-sample proposal distribution $\lambda(x)$:

$$\pi(f) = \int_{\mathsf{X}} f(x)\pi(x)\mathrm{d}x = \int_{\mathsf{X}} f(x)w(x)\lambda(x)\mathrm{d}x,$$

where $w(x)$ – importance weight (Radon-Nikodym derivative)

$$w(x) := \frac{\pi(x)}{\lambda(x)}$$

- Replace $\pi_n(f)$ by $\bar{\pi}_n(f)$,

$$\bar{\pi}_n(f) := \frac{1}{n}\sum_{k=0}^{n-1} f(X_i)w(X_i),$$

where $X_i \sim \lambda$.

Importance sampling

- Variance

$$\mathrm{Var}_\lambda[f(X_0)w(X_0)] = \mathbb{E}_\lambda[f^2(X_0)w^2(X_0)] - \pi^2(f)$$

- By Jensen's inequality

$$\mathbb{E}_\lambda[f^2(X_0)w^2(X_0)] \geq (\mathbb{E}_\lambda[|f(X_0)|w(X_0)])^2 = \left(\int_{\mathsf{X}} |f(x)|\pi(x)\mathrm{d}x\right)^2$$

- Lower bound is attained for

$$\lambda^*(x) = \frac{|f(x)|\pi(x)}{\int_{\mathsf{X}} |f(x)|\pi(x)\mathrm{d}x}$$

- High sampling efficiency is achieved when we focus on sampling from π in the importance regions where $|f(x)|\pi(x)$ is relatively large.

Self-Normalized Importance Sampling

- π is known up to a normalizing factor Z_π, $\pi(\mathrm{d}x) = \tilde{\pi}(\mathrm{d}x)/\,\mathsf{Z}_\Pi$;
- Define *importance weights* as $\tilde{w}(x) = \tilde{\pi}(x)/\lambda(x)$;
- Then

$$\begin{aligned}\pi(f) &= \int f(x)\pi(x)\mathrm{d}x = \mathsf{Z}_\pi^{-1}\int f(x)\tilde{w}(x)\lambda(x)\mathrm{d}x \\ &= \mathsf{Z}_\pi^{-1}\int f(x)\tilde{w}(x)\lambda(x)\mathrm{d}x \Big/ \left\{\mathsf{Z}_\pi^{-1}\int \tilde{w}(x)\lambda(x)\mathrm{d}x\right\}\end{aligned}$$

- The *self-normalized importance sampling* (SNIS) estimator of $\pi(f)$ is then given by

$$\widehat{\pi}_N(f) = \sum_{i=1}^N \omega_N^i f(X_i)\,,$$

where

$$X_i \sim \lambda\,, \omega_N^i = \frac{\tilde{w}(X_i)}{\sum_{j=1}^N \tilde{w}(X_j)}\,, i \in \{1,\ldots,N\}\,.$$

MCMC

- What can be done if drawing i.i.d. samples from π is not an option?
- If we run the (ergodic) Markov chain $(Z_k)_{k\geq 0}$ for a long time (started from anywhere), then for large N the distribution of Z_N will be approximately invariant: $\text{Law}(Z_N) \approx \pi$. We can then set $X_1 = Z_N$, and then restart and rerun the Markov chain to obtain $X_2, X_3, \ldots$, and then do estimates as in MC,

$$\pi_n(f) = \frac{1}{n} \sum_{k=0}^{n-1} f(X_k)$$

Important question

How to construct $\mathrm{P}(x, A)$ such that the distribution of X_n converges to invariant distribution π as quickly as possible for arbitrary initial distribution ξ?

Markov chains

What to read?

For more details see Douc et al. [2018]

Define a Markov chain (i.e., discrete time).

Ingredients of the definition:

- X – state space (e.g. $\mathsf{X} \subset \mathbb{R}^{\mathsf{d}}$), $\mathcal{X}$ – σ-algebra of X
- Initial distribution $X_0 \sim \xi$;
- Transition kernel $\mathsf{P}(x, A)$, where $x \in \mathsf{X}, \mathsf{A} \in \mathcal{X}$:

$$\mathsf{P}(X_{n+1} \in A | X_n = x) = \mathsf{P}(x, A)$$

- Markov property: X_{n+1} depends only on X_n;

Example: Model $X_0 \sim \xi$ and for $n \geq 1$

$$X_n = F(X_{n-1}, \varepsilon_n)$$

where $(\varepsilon_n)_{n \geq 1}$ is an i.i.d. sequence independent of $\sigma\{X_k, 0 \leq k \leq n-1\}$ and F is some function, $F : \mathsf{X} \times \mathbb{R}^{\mathsf{d}'} \to \mathsf{X}$

Markov chains: gym

- More about MK kernels
- Ergodicity (finite case)
- Ergodicity (not in this course:( )
- Ready for MCMC

Markov chains

Action on measures

Let μ be a probability measure on X

$$\mu P(A) = \int_{\mathrm{X}} \mu(\mathrm{d}x)\, \mathrm{P}(x, A)$$

Action on functions

$$\mathrm{P}\, f(x) = \int_{\mathrm{X}} f(y)\, \mathrm{P}(x, \mathrm{d}y)$$

Composition of kernels

$$\mathrm{P}^n(x, A) = \int_{\mathrm{X}} \mathrm{P}(x, \mathrm{d}y)\, \mathrm{P}^{n-1}(y, A)$$

(Kolmogorov-Chapman equation)

Markov chains

Tensor product (kernel $\otimes$ kernel)

$$\begin{aligned} \mathsf{P} \otimes \mathsf{P}\, f(x) &= \int_{\mathsf{X}} \mathsf{P}(x, \mathrm{d}y) \int_{\mathsf{X}} f(y, z)\, \mathsf{P}(y, \mathrm{d}z) \\ &= \int_{\mathsf{X}\times\mathsf{X}} f(y, z)\, \mathsf{P}(x, \mathrm{d}y)\, \mathsf{P}(y, \mathrm{d}z) \end{aligned}$$

Take $f(y, z) = 1(y \in A, z \in B)$. Then

$$\mathsf{P} \otimes \mathsf{P}\, f(x) = \mathsf{P}(X_1 \in A, X_2 \in B | X_0 = x) = \mathsf{P}^{\otimes 2}(x, A \times B)$$

Tensor product (measure $\otimes$ kernel)

$$\begin{aligned} \xi \otimes \mathsf{P}\, f &= \int_{\mathsf{X}} \xi(\mathrm{d}y) \int_{\mathsf{X}} f(y, z)\, \mathsf{P}(y, \mathrm{d}z) \\ &= \int_{\mathsf{X}\times\mathsf{X}} f(y, z) \xi(\mathrm{d}y)\, \mathsf{P}(y, \mathrm{d}z) \end{aligned}$$

Markov chains

Invariant distribution

Distrbution π is invariant w.r.t. P if

$$\pi P = \pi$$

Theorem

Let $(X_k)_{k=0}^{\infty}$ be a MC with initial distribution π and kernel P*. $(X_k)_{k=0}^{\infty}$ is stationary iff π is invariant.*

Proof.

Let $(X_k)_{k=0}^{\infty}$ be stationary. Then $\mathrm{Law}(X_1) = \mathrm{Law}(X_0)$. Hence,
$\pi\,\mathrm{P}(A) = \mathrm{P}_\pi(X_1 \in A) = \mathrm{P}(X_0 \in A) = \pi(A)$.
If π is invariant, then the distribution of $(X_n, \ldots, X_{n+k})$ is
$\pi\,\mathrm{P}^n \otimes \mathrm{P}^{\otimes k} = \pi \otimes \mathrm{P}^{\otimes k}$ is independent of n □

Markov chains

Reversibility

Distribution ξ is reversible w.r.t. P if

$$\xi \otimes \mathrm{P}(A \times B) = \xi \otimes \mathrm{P}(B \times A)$$

- If X is countable,

$$\xi(x)\,\mathrm{P}(x, x') = \xi(x')\,\mathrm{P}(x', x)$$

Detailed balance equation.

-

$$\begin{aligned}\mathbb{E}_\xi[f(X_0, X_1)] &= \int_{\mathrm{X}\times\mathrm{X}} \xi(\mathrm{d}x_0)\,\mathrm{P}(x_0, \mathrm{d}x_1) f(x_0, x_1) \\ &= \int_{\mathrm{X}\times\mathrm{X}} \xi(\mathrm{d}x_0)\,\mathrm{P}(x_0, \mathrm{d}x_1) f(x_1, x_0) = \mathbb{E}_\xi[f(X_1, X_0)]\end{aligned}$$

Hence, $\mathrm{Law}(X_0, X_1) = \mathrm{Law}(X_1, X_0)$

Markov chains

Theorem

Let P *be a MK. If* ξ *is reversible w.r.t.* P *then* ξ *is invariant.*

Proof.

$$\begin{aligned}\xi\,\mathsf{P}(A) &= \xi \otimes \mathsf{P}(\mathsf{X} \times \mathsf{A}) = \xi \otimes \mathsf{P}(\mathsf{A} \times \mathsf{X}) \\ &= \int_{\mathsf{X}} \xi(\mathrm{d}x)\,\mathsf{P}(x, \mathsf{X}) 1_{\mathsf{A}}(\mathsf{x}) = \xi(\mathsf{A})\end{aligned}$$

□

Ergodicity, finite case

Let X be finite, $X = [1, \dots, r]$

Total variation distance (finite case)

Let μ, ξ be probability measures on X. Define

$$d_{TV}(\xi, \mu) := \frac{1}{2} \sum_{i=1}^{r} |\mu(i) - \xi(i)| = \sum_{i:\mu(i)>\xi(i)} (\mu(i) - \xi(i))$$

Clearly, $d_{TV} \leq 1$.

- Denote $J := \{i : \mu Q(i) > \xi Q(i)\}$. Let Q be an arbitrary MK. Then for any μ, ξ

$$\begin{aligned} d_{TV}(\mu Q, \xi Q) &= \sum_{j \in J} (\mu Q(j) - \xi Q(j)) \\ &= \sum_{j \in J} \sum_{i \in X} (\mu(i) Q(i,j) - \xi(i) Q(i,j)) \\ &\leq \sum_{i:\mu(i)>\xi(i)} (\mu(i) - \xi(i)) \sum_{j \in J} Q(i,j) \leq d_{TV}(\mu, \xi) \end{aligned} \tag{2}$$

Ergodicity, finite case

- Let $Q(i,j) \geq a > 0$ for any $i, j \in X$. Then $\exists j' \notin J$ and hence for any $i \in X$

$$\sum_{j \in J} Q(i,j) < 1 - a$$

Eq. (2) my be improved:

$$d_{TV}(\mu Q, \xi Q) < (1-a) d_{TV}(\mu, \xi)$$

- Assume

$$\exists s : P^s(x, x') > 0 \text{ for any } x, x' \in X \tag{3}$$

- Let us fix arbitrary distribution μ_0 and denote $\mu_n = \mu_0 P^n$. Then

$$\begin{aligned} d_{TV}(\mu_n, \mu_{n+k}) &= d_{TV}(\mu_0 P^n, \mu_0 P^{n+k}) \\ &\leq (1-a) d_{TV}(\mu_0 P^{n-s}, \mu_0 P^{n+k-s}) \\ &\leq (1-a)^m d_{TV}(\mu_0 P^{n-ms}, \mu_0 P^{n+k-ms}), \end{aligned} \tag{4}$$

where $m : 0 < n - ms \leq s$. Take n large such that $(1-a)^m < \varepsilon$. Then $\{\mu_n\}_{n \geq 1}$ is a Cauchy sequence.

Ergodicity, finite case

- Set
$$\pi := \lim_{n\to\infty} \mu_n.$$
Then
$$\pi \mathsf{P} = \lim_{n\to\infty} \mu_n \mathsf{P} = \lim_{n\to\infty} \mu_0 \mathsf{P}^{n+1} = \pi$$
- Uniqueness: Assume $\pi_1 \neq \pi_2$ such that $\pi_1 \mathsf{P} = \pi_1, \pi_2 \mathsf{P} = \pi_2$. Then $\pi_i = \pi_i \mathsf{P}^s$, $i = 1, 2$ and
$$\mathsf{d}_{\mathsf{TV}}(\pi_1, \pi_2) \leq (1 - \mathsf{a})\mathsf{d}_{\mathsf{TV}}(\pi_1, \pi_2)$$
Hence, $\pi_1 = \pi_2$.
-
$$\begin{aligned}\mathsf{d}_{\mathsf{TV}}(\mu_0 \mathsf{P}^{\mathsf{n}}, \pi) = \mathsf{d}_{\mathsf{TV}}(\mu_0 \mathsf{P}^{\mathsf{n}}, \pi \mathsf{P}^{\mathsf{n}}) &\leq (1 - \mathsf{a})^{\mathsf{m}} \mathsf{d}_{\mathsf{TV}}(\mu_0 \mathsf{P}^{\mathsf{n}-\mathsf{ms}}, \pi \mathsf{P}^{\mathsf{n}-\mathsf{ms}}) \\ &\leq (1 - a)^m \leq (1 - a)^{n/s-1} = (1 - a)^{-1} \beta^n, \end{aligned} \tag{5}$$
where $\beta = (1 - a)^{1/s} < 1$.

Ergodicity, finite case

Theorem

Assume (3) *and let* π *be an invariant distribution. Then for any* $f : X \to \mathbb{R}$, *with probability* 1*:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) = \pi(f)$$

- Compare with SLLN for i.i.d. sequence.

MCMC

- What can be done if drawing i.i.d. samples from π is not an option?
- If we run the (ergodic) Markov chain $(Z_k)_{k\geq 0}$ for a long time (started from anywhere), then for large N the distribution of Z_N will be approximately invariant: $\text{Law}(Z_N) \approx \pi$. We can then set $X_1 = Z_N$, and then restart and rerun the Markov chain to obtain $X_2, X_3, \ldots$, and then do estimates as in MC,

$$\pi_n(f) = \frac{1}{n} \sum_{k=0}^{n-1} f(X_k)$$

Important question

How to construct $\mathrm{P}(x, A)$ such that the distribution of X_n converges to invariant distribution π as quickly as possible for arbitrary initial distribution ξ?

Example: Metropolis-Hastings algorithm

Let $Q(x, A) = \int_A q(x, y)\mathrm{d}y$ be some MK (e.g. Gaussian)

1. Choose X_0.
2. Given X_k, a candidate move Y_{k+1} is sampled from $Q(X_k, \cdot)$
3. $X_{k+1} = Y_{k+1}$ with probability $\alpha(X_k, Y_{k+1})$, otherwise $X_{k+1} = X_k$, where acceptance ratio

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\}$$

Example: Random walk MH
Take $q(x, y) = \overline{q}(y - x)$, where $\overline{q}(x) = \overline{q}(-x)$. Then

$$Y_{k+1} = X_k + Z_{k+1}, \quad Z_{k+1} \sim \overline{q}$$

In this case

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$$

Example: Langevin Dynamics

Langevin Dynamics Itô SDE:

$$\mathrm{d}\theta_t = -\nabla U(\theta_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t,$$

Invariant measure: $\pi(\theta) = e^{-U(\theta)}$ and $\mathrm{Law}(\theta_t) \to \pi$ as $t \to \infty$.

1. Take $\pi(\theta) = (2\pi)^{-1/2}\mathrm{e}^{-\theta^2/2}$.
2. SDE: $\mathrm{d}\theta_t = \theta_t\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$, θ_0 is independent of W. This is Ornstein–Uhlenbeck process
3. Apply Ito's formula to obtain

$$\theta_t = \theta_0\mathrm{e}^{-t} + \sqrt{2}\int_0^t e^{-(t-s)}\mathrm{d}W_s$$

4. Since the Itô integral of deterministic integrand is normally distributed, we readily have

$$\mathrm{Law}(\theta_t) = \mathcal{N}(\theta_0\mathrm{e}^{-t}, 1-\mathrm{e}^{-2t}) \to \mathcal{N}(0,1)$$

Example: Langevin Dynamics

Itô SDE:

$$\mathrm{d}\theta_t = -\nabla U(\theta_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t,$$

Invariant measure: $\pi(\theta) = e^{-U(\theta)}$

1. First-order discretization (Unadjusted Langevin Algorithm, ULA):

$$Y_{k+1} = Y_k - \gamma\nabla U(Y_k) + \sqrt{2\gamma}Z_{k+1}, \quad i.i.d.\ Z_k \sim \mathrm{N}(0, \mathrm{I}_d)$$

   Equivalently, $Y_{k+1} \sim \mathcal{N}(Y_k - \gamma\nabla U(Y_k), 2\gamma\,\mathrm{I})$
2. Metropolis-adjusted Langevin Algorithm (MALA):
   ULA + Metropolis-Hastings correction;
3. Demo: https://chi-feng.github.io/mcmc-demo
4. If we can't calculate ∇U replace it by its estimate over batch (SGLD, SGLD-FP, SAGA etc)

SGLD

1. Posterior distribution:

$$\pi(\theta|\mathbf{X}) = \frac{\prod_{i=1}^{N} p(X_i|\theta)\pi_0(\theta)}{\int_{\mathbb{R}^d} \prod_{i=1}^{N} p(X_i|\theta)\pi_0(\theta)\, d\theta} \propto \mathrm{e}^{-U(\theta)},$$

   where $U = \log \pi_0(\theta) + \sum_{i=1}^{N} \log p(X_i|\theta)$;
2. A computational bottleneck: calculating the full gradient ∇U scaling proportionally to N can be very time consuming in the "big data" limit;
3. Replace $\nabla U(\theta)$ by an unbiased estimate. This gives rise to the SGLD algorithm, where the parameters are updated according to

$$\begin{aligned} \theta_{k+1} &= \theta_k - \gamma G(\theta_k, S_{k+1}) + \sqrt{2\gamma}\, \xi_{k+1}, \\ G(\theta, S) &= \nabla U_0(\theta) + KM^{-1} \sum\nolimits_{i \in S} \nabla U_i(\theta), \end{aligned} \tag{6}$$

   where each S_{k+1} is a random batch taking values in S_M (here S_M is the set of all subsets S of $\{1, \ldots, N\}$ with $|S| = M$) which is sampled from a uniform distribution over S_M independently of $\mathfrak{F}_k$ (here $(\mathfrak{F}_k)_{k\geq 0}$ is the filtration generated by $\{(\theta_\ell, S_\ell)\}_{\ell \geq 0}$).
4. Note that $\mathbb{E}[G(\theta_k, S_{k+1})|\mathfrak{F}_k] = \nabla U(\theta_k)$ and therefore $G(\theta_k, S_{k+1})$ is an unbiased estimate of $\nabla U(\theta_k)$.

Transition kernel of MH algorithm

Let $Q(x, A) = \int_A q(x, y)\mathrm{d}y$ be some MK (e.g. Gaussian)

1. Choose X_0.
2. Given X_k, a candidate move Y_{k+1} is sampled from $Q(X_k, \cdot)$
3. $X_{k+1} = Y_{k+1}$ with probability $\alpha(X_k, Y_{k+1})$, otherwise $X_{k+1} = X_k$, where acceptance ratio

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\}$$

MH transition kernel

$$\mathrm{P}(x, A) = \int_A \alpha(x, y)q(x, y)\mathrm{d}y + \overline{\alpha}(x)\delta_x(A),$$

where

$$\overline{\alpha}(x) = \int_{\mathrm{X}} (1 - \alpha(x, y))q(x, y)\mathrm{d}y.$$

Invariance of π

Theorem

Distribution π is reversible w.r.t. P.

Proof.

We need to show that for any $C \in \mathcal{X} \times \mathcal{X}$

$$\int_{\mathsf{X}\times\mathsf{X}} \pi(x)\mathrm{d}x\,\mathsf{P}(x,\mathrm{d}y)1_C(x,y) = \int_{\mathsf{X}\times\mathsf{X}} \pi(y)\mathrm{d}y\,\mathsf{P}(y,\mathrm{d}x)1_C(x,y)$$

For any $x, y \in \mathsf{X}$

$$\pi(x)\alpha(x,y)q(x,y) = \{\pi(x)q(x,y)\} \vee \{\pi(y)q(y,x)\} = \pi(y)\alpha(y,x)q(y,x)$$

Moreover,

$$\int_{\mathsf{X}\times\mathsf{X}} \pi(x)\mathrm{d}x\delta_x(\mathrm{d}y)\overline{\alpha}(x)1_C(x,y) = \int_{\mathsf{X}} \pi(x)\mathrm{d}x\overline{\alpha}(x)1_C(x,x)$$
$$= \int_{\mathsf{X}} \pi(y)\mathrm{d}y\overline{\alpha}(y)1_C(y,y) = \int_{\mathsf{X}\times\mathsf{X}} \pi(y)\mathrm{d}y\delta_y(\mathrm{d}x)\overline{\alpha}(y)1_C(x,y)$$

Analysis of ULA

- Let $\pi(x) = Z_d^{-1} \mathrm{e}^{-U(x)}$;

L–smooth potential

U is L–smooth is $U \in C^2(\mathbb{R}^d)$ and there exists $L > 0$ such that

$$\|\nabla U(x) - \nabla U(y)\| \leq L\|x - y\|$$

for any $x, y \in \mathbb{R}^d$.

- Unadjusted Langevin Algorithm, ULA:

$$X_{k+1} = X_k - \gamma \nabla U(Y_k) + \sqrt{2\gamma} Z_{k+1}, \quad i.i.d.\ Z_k \sim \mathrm{N}(0, \mathrm{I}_d)$$

- Denote $\mathrm{P}_\gamma(x, \cdot) = \mathcal{N}(x - \gamma \nabla U(x), 2\gamma\, \mathrm{I})$.

Kantorovich–Wasserstein distance

Kantorovich–Wasserstein distance

For λ, ν, we denote their coupling set by $\Pi(\lambda, \nu)$, i.e. $\xi \in \Pi(\lambda, \nu)$ is the measure on $\mathsf{X} \times \mathsf{X}$ satisfying for all $A \in \mathcal{B}(\mathsf{X})$, $\xi(A, \mathsf{X}) = \lambda(\mathsf{A})$ and $\xi(\mathsf{X}, \mathsf{A}) = \nu(\mathsf{A})$. For $p \geq 1$ and λ, ν, let

$$W_{p,\mathsf{d}}(\lambda, \nu) := \inf_{\Pi(\lambda,\nu)} \left\{ \int_{\mathsf{X}\times\mathsf{X}} \mathsf{d}^p(x, y)\, \xi(\mathrm{d}x, \mathrm{d}y) \right\}^{1/p}$$

be the Kantorovich–Wasserstein distance of order p between λ and ν.

Analysis of ULA

A1

U is L–smooth and m–strongly convex:

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m \|x - y\|^2.$$

Theorem

For any $\gamma \in (0, m/L^2)$ there exists invariant distribution π_γ:

$$W_2^2(\delta_x \mathrm{P}_\gamma^k, \pi_\gamma) \leq (1 - m\gamma)^k \int \|x - y\|^2 \pi_\gamma(\mathrm{d}y)$$

Analysis of ULA

- Fix $x, \tilde{x} \in \mathbb{R}^d$. Synchronous coupling:

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1},$$
$$\tilde{X}_{k+1} = \tilde{X}_k - \gamma \nabla U(\tilde{X}_k) + \sqrt{2\gamma} Z_{k+1}$$

- Then

$$\begin{aligned}\|X_{k+1} - \tilde{X}_{k+1}\|^2 = \|X_k - \tilde{X}_k\|^2 \\ \gamma^2 \|\nabla U(X_k) - \nabla U(\tilde{X}_k)\|^2 \\ - 2\gamma \langle X_k - \tilde{X}_k, \nabla U(X_k) - \nabla U(\tilde{X}_k)[]\end{aligned}$$

- Use A1:

$$\begin{aligned}\|X_{k+1} - \tilde{X}_{k+1}\|^2 &\leq (1 + \gamma^2 L^2 - 2\gamma m)\|X_k - \tilde{X}_k\|^2 \\ &\leq (1 - \gamma m)\|X_k - \tilde{X}_k\|^2.\end{aligned}$$

- Hence

$$W_2^2(\delta_x \mathrm{P}_\gamma^k, \delta_{\tilde{x}} \mathrm{P}_\gamma^k) \leq (1 - m\gamma)^k W_2^2(\delta_x, \delta_{\tilde{x}})$$

- We may show that $(\lambda \mathrm{P}_\gamma^k)_{k \in \mathbb{N}}$ is a Cauchy sequence and there exists $\pi_\gamma^\lambda = \pi_\gamma$, moreover $\pi_\gamma \mathrm{P}_\gamma = \pi_\gamma$.

Variance of MCMC estimate

Let π be an invariant distribution. Assume $X_0 \sim \pi$, i.e. we start from the invariant distribution. Then

$$\mathrm{Var}_\pi\left[n^{-1}\sum_{k=0}^{n-1} f(X_k)\right] = \frac{\mathrm{Var}_\pi[f]}{n} + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}_\pi\left[(f(X_i)-\pi(f))(f(X_j)-\pi(f))\right] =$$

$$= \frac{\rho^{(f)}(0)}{n} + \frac{2}{n}\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\rho^{(f)}(k) \neq \frac{\mathrm{Var}_\pi[f]}{n}$$

where

$$\rho^{(f)}(k) = \mathbb{E}_\pi\left[(f(X_0)-\pi(f))(f(X_k)-\pi(f))\right]$$

Variance of MCMC estimate

- Under appropriate conditions (e.g. ϕ-irreducibility + apereodicity + existence of solution of Poisson eq.) CLT holds:

$$\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} [f(X_i) - \pi(f)] \xrightarrow{Law} \mathcal{N}(0, V_\infty(f)),$$

  where $V_\infty(f) := \lim_{n\to\infty} \mathrm{Var}_\pi \left[\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} (f(X_i) - \pi(f)) \right]$

- Length of confidence interval for $\pi_n(f)$ proportional to $\frac{\sqrt{V_\infty(f)}}{\sqrt{n}}$

Ex^{2}MCMC: Sampling through Exploration Exploitation

Importance Sampling procedure

- Aim: sample from π and estimate $\pi(f) = \int_{\mathbb{R}^D} f(x)\pi(\mathrm{d}x)$;
- π is known up to a normalizing factor Z_Π, $\pi(\mathrm{d}x) = \tilde{\pi}(\mathrm{d}x)/Z_\Pi$;
- Importance Sampling (IS) consists of re-weighting samples from a proposal distribution λ.
- Define *importance weights* as $\tilde{w}(x) = \tilde{\pi}(x)/\lambda(x)$;
- The *self-normalized importance sampling* (SNIS) estimator of $\pi(f)$ is then given by

$$\widehat{\pi}_N(f) = \sum_{i=1}^{N} \omega_N^i f(X^i)\,,$$

where

$$X^{1:N} \sim \lambda\,, \omega_N^i = \frac{\tilde{w}(X^i)}{\sum_{j=1}^{N} \tilde{w}(X^j)}\,, i \in \{1, \ldots, N\}\,.$$

From IS to SIR

- Sampling counterpart of the IS procedure is known as Sampling Importance Resampling (SIR; Rubin [1987]);
- Sample $X^1, \ldots, X^N$ - i.i.d. from λ and compute the importance weights $\omega_N^1, \ldots, \omega_N^N$;
- Sample $Y^1, \ldots, Y^M$ from $X^1, \ldots, X^N$ with replacement, and with probabilities proportional to the weights $\omega_N^1, \ldots, \omega_N^N$. That is, we sample from the empirical distribution

$$\hat{\pi}(\mathrm{d}x) = \sum_{i=1}^{N} \omega_N^i \delta_{X^i}(\mathrm{d}x)\,,$$

  where $\delta_y(\mathrm{d}x)$ denotes the Dirac mass at y.
- As $N \to \infty$, $Y^1, \ldots, Y^M \sim \hat{\Pi}$ will be distributed according to π.
- Main drawback: the described procedure is only asymptotically valid.

Iterated SIR (i-SIR) algorithm

Iterating samples from λ, we arrive at iterated SIR algorithm (i-SIR, Andrieu et al. [2010], and Andrieu et al. [2018]).

Algorithm 1: Single stage of i-SIR algorithm

Input : Sample Y_j from previous iteration
Output: New sample Y_{j+1}

1 Set $X_{j+1}^1 = Y_j$ and draw $X_{j+1}^{2:N} \sim \lambda$.
2 **for** $i \in [N]$ **do**
3 compute the normalized weights $\omega_{i,j+1} = \tilde{w}(X_{j+1}^i) / \sum_{k=1}^N \tilde{w}(X_{j+1}^k)$.
4 Set $I_{j+1} = \mathrm{Cat}(\omega_{1,j+1}, \ldots, \omega_{N,j+1})$.
5 Draw $Y_{j+1} = X_{j+1}^{I_{j+1}}$.

The Markov chain $\{Y_k,\, k \in \mathbb{N}\}$ generated by i-SIR has the following Markov kernel

$$\mathrm{P}_N(x, \mathrm{A}) = \int \delta_x(\mathrm{d}x^1) \sum_{i=1}^N \frac{\tilde{w}(x^i)}{\sum_{j=1}^N \tilde{w}(x^j)} \mathbb{1}_{\mathrm{A}}(x^i) \prod_{j=2}^N \lambda(\mathrm{d}x^j). \tag{7}$$

i-SIR algorithm

- Provided also that $|\tilde{w}|_\infty < \infty$, it was shown in Andrieu et al. [2018] that the Markov kernel P_N is uniformly geometrically ergodic. Namely, for any initial distribution ξ on $(\mathrm{X}, \mathcal{X})$ and $k \in \mathbb{N}$,

$$\|\xi \mathrm{P}_N^k - \pi\|_{\mathrm{TV}} \leq \kappa_N^k, \tag{8}$$

  with $\epsilon_N = \frac{N-1}{2\mathrm{L}+N-2}, \mathrm{L} = |\tilde{w}|_\infty / \lambda(\tilde{w})$ and $\kappa_N = 1 - \epsilon_N$.
- Note that the bound (8) relies significantly on the restrictive condition that weights are uniformly bounded $|\tilde{w}|_\infty < \infty$.
- Moreover, even when this condition is satisfied, the rate κ_N can be close to 1 when the dimension d is large.
- Indeed, consider a simple scenario $\pi(x) = \prod_{i=1}^d p(x_i)$ and $\lambda(x) = \prod_{i=1}^d q(x_i)$ for some densities $p(\cdot)$ and $q(\cdot)$ on $\mathbb{R}$. Then it is easy to see that $\mathrm{L} = (\sup_{y \in \mathbb{R}} p(y)/q(y))^d$ grows exponentially with d.

i-SIR algorithm

To illustrate this phenomenon, we consider a simple problem of sampling from the standard normal distribution $\mathcal{N}(0, \mathrm{I}_d)$ with the proposal $\mathcal{N}(0, 2\,\mathrm{I}_d)$ in increasing dimensions d up to 300.

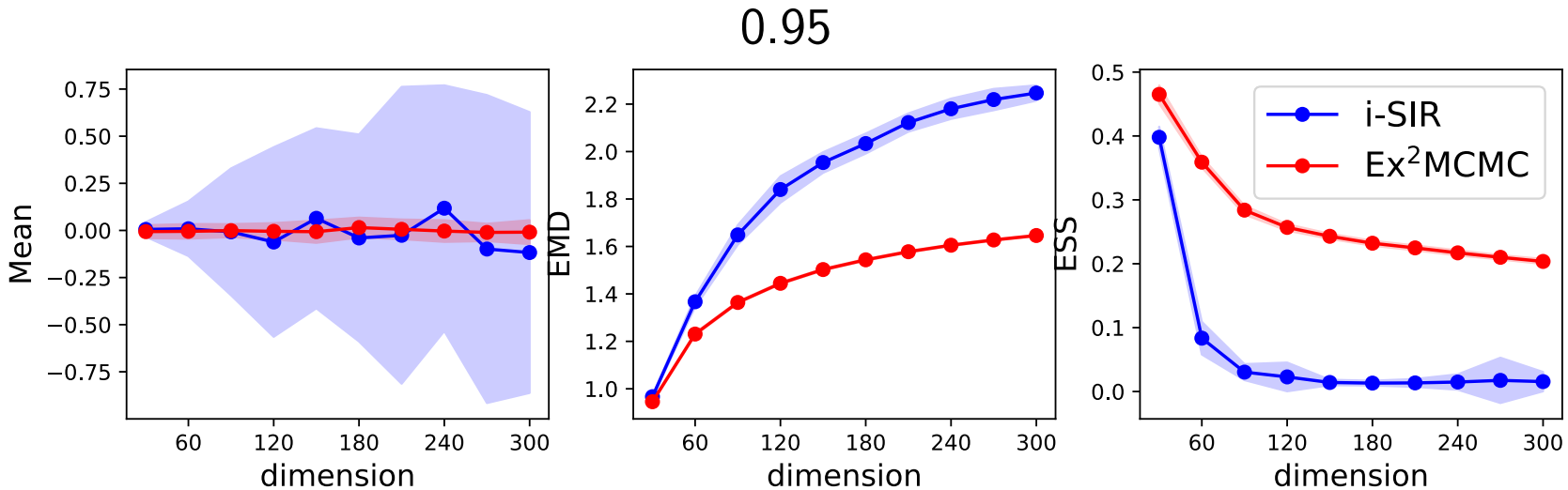


Figure: Sampling from $\mathcal{N}(0, \mathrm{I}_d)$ with the proposal $\mathcal{N}(0, 2\,\mathrm{I}_d)$. We display confidence intervals for i-SIR and Ex^{2}MCMC obtained from 100 independent runs as blue and red regions, respectively. Ex^{2}MCMC helps to achieve efficient sampling even in high dimensions.

Ex^{2}MCMC algorithm

- Main i-SIR drawback: absence of local exploration moves;
- Idea: apply a local MCMC kernel R (*rejuvenation kernel*) after each i-SIR step;
- R has π as invariant distribution;
- Here comes Ex^{2}MCMC : Exploration steps through i-SIR , Exploitation steps through R$(x, \cdot)$;
- As our default choice we consider MALA as rejuvenation, but other ones (HMC, NUTS) are also possible.

Ex^{2}MCMC algorithm

Algorithm 2: Single stage of Ex^{2}MCMC algorithm with independent proposals

Procedure Ex^{2}MCMC $(Y_j, \Lambda, \mathrm{R})$**:**

  Input : Previous sample Y_j;
  proposal distribution Λ;
  rejuvenation kernel R;

  Output: New sample Y_{j+1};

  Set $X_{j+1}^1 = Y_j$, draw $X_{j+1}^{2:N} \sim \lambda$;

  for $i \in [N]$ **do**

    compute the normalized weights $\omega_{i,j+1} = \tilde{w}(X_{j+1}^i) / \sum_{k=1}^N \tilde{w}(X_{j+1}^k)$;

  Set $I_{j+1} = \mathrm{Cat}(\omega_{1,j+1}, \ldots, \omega_{N,j+1})$;

  Draw $Y_{j+1} \sim \mathrm{R}(X_{j+1}^{I_{j+1}}, \cdot)$.

Ex^{2}MCMC algorithm

V-geometric ergodicity

A Markov kernel Q with invariant probability measure π is V-geometrically ergodic if there exist constants $\rho \in (0, 1)$ and $M < \infty$ such that, for all $x \in \mathsf{X}$ and $k \in \mathbb{N}$,

$$\|\mathrm{Q}^k(x, \cdot) - \pi\|_V \leq M \{V(x) + \pi(V)\} \rho^k.$$

Assumptions

A1

(i) R has π as its unique invariant distribution;
(ii) There exists a function $V\colon \mathsf{X} \to [1, \infty)$, such that for all $r \geq r_{\mathsf{R}} > 1$ there exist $\lambda_{\mathsf{R},r} \in [0, 1)$, $\mathsf{b}_{\mathsf{R},r} < \infty$, such that
$\mathsf{R}V(x) \leq \lambda_{\mathsf{R},r} V(x) + \mathsf{b}_{\mathsf{R},r} \mathbb{1}_{\mathsf{V}_r}$, where $\mathsf{V}_r = \{x\colon V(x) \leq r\}$;

A2

(i) For all $r \geq r_{\mathsf{R}}$, $\tilde{w}_{\infty,r} := \sup_{x \in \mathsf{V}_r} \{\tilde{w}(x)/\lambda(\tilde{w})\} < \infty$;
(ii) $\mathrm{Var}_\lambda[\tilde{w}]/\{\lambda(\tilde{w})\}^2 < \infty$.

Ex^{2}MCMC algorithm

Theorem

Let A1 and A2 hold. Then, for all $x \in \mathsf{X}$ and $k \in \mathbb{N}$,

$$\|\mathsf{K}_N^k(x, \cdot) - \pi\|_V \leq c_{\mathsf{K}_N}\{\pi(V) + V(x)\}\tilde{\kappa}_{\mathsf{K}_N}^k \,, \tag{9}$$

where c_{K_N}, $\tilde{\kappa}_{\mathsf{K}_N} \in [0, 1)$ are some constants. In addition, $c_{\mathsf{K}_N} = c_{\mathsf{K}_\infty} + O(N^{-1})$ and $\tilde{\kappa}_{\mathsf{K}_N} = \tilde{\kappa}_{\mathsf{K}_\infty} + O(N^{-1})$.

Toy example

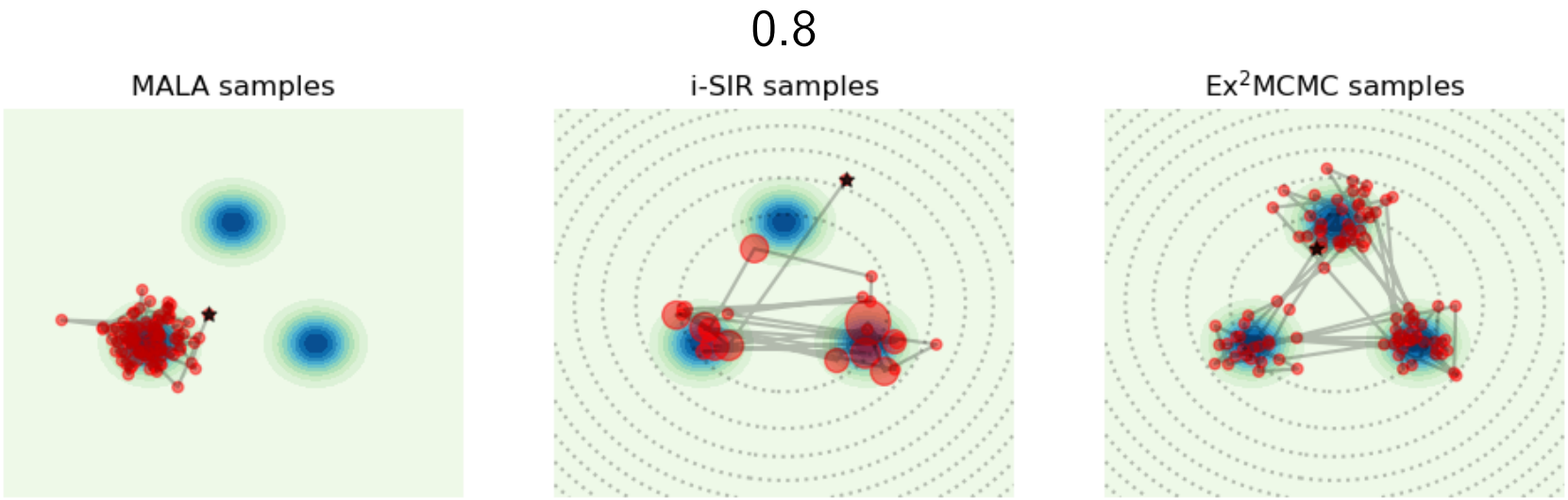


Figure: Single chain mixing visualization. – Blue color levels represent the target 2d density. Random chain initialization is noted in black, 100 steps are plotted per sampler: the size of each red dot corresponds to the number of consecutive steps the walkers remains at a given location. For MALA, we generate 300 samples and choose each 3-rd one for comparability. Note that the variance of the global proposal (dotted countour lines) should be relatively large to cover well all the modes. The step size of MALA also can not be increased much to keep reasonable acceptance ratio.

Adaptive proposals

- Consider family of proposals $\{\lambda_\theta\}, \theta \in \mathbb{R}^D$, chosen to match the target distribution $\tilde{\pi}$;
- Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be smooth and invertible. Denote by $T\#\Lambda$ the distribution of $Y = T(X)$ with $X \sim \lambda$;
- The corresponding density is given by $\lambda_T(y) = \lambda\big(T^{-1}(y)\big)\, \mathrm{J}_{T^{-1}}(y)$, where J_T denotes the Jacobian determinant of T;

Adaptive proposals: learning procedure

- Disperancy measure: linear combination of forward and backward KL divergence (generalizations to [Papamakarios et al., 2021] possible);
- Forward and backward KL:

$$\mathcal{L}^f(\theta) = \int \log \frac{\pi(x)}{\lambda_\theta(x)} \pi(x) \mathrm{d}x \,,$$

$$\mathcal{L}^b(\theta) = \int \log \frac{\lambda(x)}{\pi\big(T_\theta(x)\big)\, \mathrm{J}_{T_\theta}(x)} \lambda(x) \mathrm{d}x \,.$$

- Given a sample $Y_k \sim \pi$ and $Z_k \sim \lambda$ for $k \in [K]$, by

$$\widehat{\nabla \mathcal{L}^f}(Y^{1:K}, \theta) = -\frac{1}{K} \sum_{k=1}^{K} \nabla \log \lambda_\theta(Y_k) \,,$$

$$\widehat{\nabla \mathcal{L}^b}(Z^{1:K}, \theta) = -\frac{1}{K} \sum_{k=1}^{K} \nabla \log\big(\tilde{\pi}(T_\theta(Z_k)\, \mathrm{J}_{T_\theta}(Z_k)\big) .$$

- Following Gabrié et al. [2021], we consider

$$\widehat{\mathcal{L}}(Y^{1:K}, Z^{1:K}, \theta) = \alpha \widehat{\mathcal{L}^f}(Y^{1:K}, \theta) + \beta \widehat{\mathcal{L}^b}(Z^{1:K}, \theta) \,.$$

FIEx^{2}MCMC algorithm with adaptive proposals

Algorithm 3: Single stage of FIEx^{2}MCMC. Steps of Ex^{2}MCMC are done in parallel with common values of proposal parameters θ_j. Step 4 updates the parameters using the gradient estimate obtained from all the chains.

Input : weights θ_j, batch $Y_j^{1:K}$

Output: new weights θ_{j+1}, batch $Y_{j+1}^{1:K}$

1 **for** $k \in [K]$ **do**

2 $\quad Y_{j+1,k} = \text{Ex}^2\text{MCMC}\,(Y_{j,k}, T_{\theta_j}\#\Lambda, \text{R})$

3 Draw $\bar{Z}^{1:K} \sim \lambda$.

4 Update $\theta_{j+1} = \theta_j - \gamma\widehat{\nabla\mathcal{L}}(Y_{j+1}, \bar{Z}, \theta_j)$.

Practical note

In our experiments: T_θ is modelled as a normalizing flow based on RealNVP architecture (Dinh et al. [2017]).

Take-home Messages & Future Works

- We know basics of MC, rejection sampling, importance samling, MCMC, normalizing flows
- To become world expert in Markov chains read Douc et al. [2018]
- We are ready for 'real' projects (join HDI Lab team)

Thank you!

References

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 72(3):269–342, 2010.

Christophe Andrieu, Anthony Lee, Matti Vihola, et al. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2):842–872, 2018.

Tong Che, Ruixiang ZHANG, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, and Yoshua Bengio. Your GAN is Secretly an Energy-based Model and You Should Use Discriminator Driven Latent Sampling. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12275–12287. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/90525e70b7842930586545c6f1c9310c-Paper.pdf.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018. ISBN 978-3-319-97703-4; 978-3-319-97704-1. doi: 10.1007/978-3-319-97704-1. URL https://doi.org/10.1007/978-3-319-97704-1.

Marylou Gabrié, Grant M. Rotskoff, and Eric Vanden-Eijnden. Adaptive Monte Carlo augmented with normalizing flows. *arXiv preprint arXiv:2105.12603*, 2021.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Donald B Rubin. Comment: A noniterative Sampling/Importance Resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82 (398):542–543, 1987.

Ryan Turner, Jane Hung, Eric Frank, Yunus Saatchi, and Jason Yosinski. Metropolis-Hastings generative adversarial networks. In *International Conference on Machine Learning*, pages 6345–6353. PMLR, 2019.