Sampling Basics

References:

① Blog posts by Gregory Gunderson (several posts on rejection sampling, Metropolis-Hastings algo, Gibbs sampling — see gregorygunderson.com/blog/)

② A. Brown & G. Jones, Convergence rates of Metropolis-Hastings algorithms, Advanced Review, Wiley 2024

③ P. Bremaud, Markov Chains, Springer 2008.

④ S. Geman and C. Hwang, Diffusions for Global Optimization, SIAM J. of Control and Optimization, Sep. 86.

⑤ S. Meyn and R. L. Tweedie, Markov Chains and Stochastic Stability, Springer 2005.

⑥ N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, Equation of state calculations by fast computing machines, The journal of Chemical Physics, 1953.

⑦ W. Hastings, Monte Carlo methods using Markov Chains and their applications, Biometrika, 1970.

One of the original motivations for sampling was to solve (non-convex) optimization problems.

$\min_x U(x)$ $\longleftrightarrow$ sample from $\pi_T(x) \propto e^{-U(x)/T}$

optimization over $x \in \mathbb{R}^d$

T = temperature.

$\pi_T(x)$ a dist. over $\mathbb{R}^d$

as $T \to 0$, the distribution concentrates on the set $\{x^* : x^* \in \operatorname{argmin}_x U(x)\}$.

Thus, for small enough T, samples $y \sim \pi_T(\cdot)$ will very likely satisfy $U(y) \approx U(x^*)$.

1. Rejection sampling (late 1940s)
2. Inverse CDF sampling
3. Metropolis-Hasting algo (1950s+)
4. Gibbs sampling (1990s+)

General setting: We want to sample from some distribution $\pi^*(x)$ ($\pi^*(x)$ could be discrete or continuous; in these notes I'll generally stick to continuous but ignore technical details involving changing order of integrals, switching limits/integrals, dominated conv., etc.).

In this note, we assume we know $\pi^*(x)$, either exactly (rejection sampling), or up to a scaling

factor (Metropolis-Hastings) or the score function $\nabla \ln \pi^*(x)$ (Langevin sampling). By "know", we mean query access $\longrightarrow$ given x, blackbox returns one of above. In future notes, we will study learning based approaches where we only have access to samples $x_1, \ldots, x_n \sim \pi^*(x)$.

Rejection sampling Assume $x \in \mathbb{R}$, $\pi^*(x)$ a density function. Further assume that we can generate samples from some other known distribution $g(x)$ such that $\exists M < \infty$ with

$$0 \le \pi^*(x) \le M g(x) \quad \forall \ x \in \mathbb{R}.$$

(for instance, $g(x)$ could be Gaussian, or if $\pi^*(x)$ has bounded support, then a Uniform sample over the domain).

candidate generating function

Algorithm:
- (a) Sample $X \sim g(x)$.
- (b) Sample $U \sim \text{Uniform}[0,1]$ ($X \perp\!\!\!\perp U$) (indep.)
- (c) Accept x if $U \le \dfrac{\pi^*(x)}{M g(x)}$.

else reject the sample X.

Denote $A = \chi\left\{U \le \frac{\pi^*(X)}{M g(X)}\right\}$ (acceptance r.v.)

→ indicator function

Claim: $P(X \in B \mid A=1) = \int_B \pi^*(x)\,dx$.

"Pf:" (with densities)

$$\underbrace{P(X \in [x, x+dx] \mid A=1)}_{P(X \approx x \mid A=1)\,dx} = \frac{P(A=1 \mid \underbrace{X \in [x, x+dx]}_{X \approx x})\; \underbrace{P(X \in [x, x+dx])}_{P(X \approx x)\,dx}}{P(A=1)}$$

$P(X \approx x \mid A=1)\,dx$: (conditional density function)

$$= \frac{P\left(U \le \frac{\pi^*(x)}{M g(x)}\right) g(x)\,dx}{P(A=1)}$$

$$= \frac{\pi^*(x)}{M \cancel{g(x)} \cdot P(A=1)} \cancel{g(x)}\,dx$$

$$= \underbrace{\left(\frac{1}{M P(A=1)}\right)}_{\text{does not depend on } x} \pi^*(x)\,dx$$

i.e., $P(X \approx x \mid A=1) \propto \pi^*(x)$.

Since both sides are non-negative and integrate to 1, the proportionality constant must be '1', i.e.,

$$P(A=1) = \frac{1}{M} \quad \text{and} \quad P(X \in B \mid A=1) = \int_B \pi^*(x)\,dx$$

∎

(we will defer the vector case to after M-H algorithm).

Inverse CDF Sampling: Let $F^*(x)$ be the CDF of $\pi^*(x)$. Let $x \in \mathbb{R}$.

Algorithm:

(a) $Y \sim g(y)$

(b) $Z = G(Y)$ $\quad (Z \sim \text{Unif}[0,1])$

(c) $X = F^{*-1}(Z)$

$$G(z) = \int_{-\infty}^{z} g(w)\,dw.$$

Then, $X \sim \pi^*(x)$.

Pf: $P(Z \le z) = P(G(Y) \le z)$

$$= P\left(Y \le G^{-1}(z)\right) = \int_{-\infty}^{G^{-1}(z)} g(x)\,dx$$

$$\therefore \frac{d}{dz} P(Z \le z) = g\left(G^{-1}(z)\right) \cdot \left(G^{-1}(z)\right)'$$

$$= g\left(G^{-1}(z)\right) \cdot \frac{1}{G'\left(G^{-1}(z)\right)}$$

$x = G^{-1}(z)$
i.e., $z = G(x)$
$\frac{dz}{dx} = G'(x)$
$\frac{dx}{dz} = \frac{1}{G'(G^{-1}(z))}$

$$= \frac{g\left(G^{-1}(z)\right)}{g\left(G^{-1}(z)\right)} \cdot \frac{1}{} = 1$$

i.e., $Z \sim \text{Unif}[0,1]$.

$$P(X \le x) = P\left(F^{*-1}(Z) \le x\right) = P\left(Z \le F^{*}(x)\right)$$

$$= F^{*}(x),$$

i.e., $X \sim \pi^{*}(x)$. ☑

Vector version: Let $Y \sim g(y)$, $Y \in \mathbb{R}^d$

$h(\cdot)$ an invertible function, $X = h(Y)$.
$(\mathbb{R}^d \longmapsto \mathbb{R}^d)$

Then, $f_x(x) = g\left(h^{-1}(x)\right) \cdot \left|J\left(h^{-1}(y)\right)\right|$ (determinant)

where $J_r(x) = \begin{pmatrix} \frac{\partial r_1}{\partial x_1} & \cdots & \frac{\partial r_1}{\partial x_d} \\ \vdots & \ddots & \\ \frac{\partial r_d}{\partial x_1} & \cdots & \frac{\partial r_d}{\partial x_d} \end{pmatrix}$

Jacobean matrix.

(This is useful when studying Normalizing Flows; we will not cover it in this class).

Metropolis-Hastings Algorithm [6, 7]: Sampling in the setting where we do not have access to $\pi^*(x)$. Instead, we only have access to $C\pi^*(x)$ for some unknown C.

High level: We construct a discrete-time, and (discrete or continuous) state-space Markov process.

Below, we write the equations for continuous

state-space, but the analogous equations (typically replace $x \to i$, $y \to j$) hold for discrete state-spaces.

The Markov process evolves over time, and the stationary distribution of the M.P. is $\pi^*(x)$.

State-space S (typically $\mathbb{R}^d$).

Transition density: $p(y|x)$. (analog of the trans. prob p_{ij})

$$p(y|x)\,dy \approx P\left(X_{t+1} \in [y, y+dy] \,\middle|\, X_t = x\right).$$

More formally, $P(x, B) = \int_{y \in B} p(y|x)\,dy$.

($x \in S$; $P(x,B)$: transition kernel; $B \subseteq S$)

---

e.g., $p(y|x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\|x-y\|^2}$, $x, y \in \mathbb{R}^d$.

(Gaussian trans. density).

Reversibility, A M.P. (π, p) is reversible if

$$\pi(x)\, p(y|x) = \pi(y)\, p(x|y) \quad \forall\, x, y \in \mathbb{R}^d$$

In this case, $\pi(\cdot)$ is the stat. dist. of the M.P. (analogous to disc. stat space).

check: $$\int_B P(x, B)\, \pi(x)\, dx = \int_B \pi(y)\, dy$$

Aside: When does a unique stat. dist exist for cont. stat-space, discrete time M.P.?

The M.P. is aperiodic, irreducible, positive Harris recurrent (see Meyn & Tweedie book); sufficient condition using Poisson's equation to construct a Lyapunov function. $V(\cdot)$ satisfying $E\left[V(X_{t+1}) \mid X_t = x\right]$

$V(\cdot) \geq 0$, $\lambda \in (0, 1)$ $\leq \lambda V(x) + c$,

$\forall\, x \notin K$ ↳ compact set

Back to M-H Algorithm: Construct a reversible D.T. M.P., whose stat. dist. is $\pi^*(x)$, $x \in S$ (e.g. $\mathbb{R}^d$).

Thus, we need to construct a $p(y|x)$ s.t.

$$\pi^*(x)p(y|x) = \pi^*(y)p(x|y).$$

Approach: $q(\cdot|x) \geq 0$, $\int_{y \in S} q(y|x)dy = 1$

Choose $q(y|x)$: generating density function

We have access to $\left(\frac{\pi^*(x)}{\pi^*(y)}\right)$ for any (x, y)

If $\frac{q(y|x)}{q(x|y)} = \frac{\pi^*(y)}{\pi^*(x)}$, then we are done. (we would simply choose $p(y|x) = q(y|x)$).

In general, this is not likely to be true.

Construct an acceptance function $\alpha(x, y)$

s.t. $$\frac{q(y|x)\,\alpha(x,y)}{q(x|y)\,\alpha(y,x)} = \frac{\pi^*(y)}{\pi^*(x)}$$ → (*)

and $0 \le \alpha(x,y) \le 1 \quad \forall\, x, y.$

In this case, $p(y|x) = q(y|x)\,\alpha(x,y)$. → (**)

Notice that for any $q(\cdot|\cdot)$, if we choose

$$\alpha(x,y) = \min\left\{1, \frac{q(x|y)}{q(y|x)} \cdot \frac{\pi^*(y)}{\pi^*(x)}\right\},$$ → (***)

then (*) is satisfied. Thus, the D.T.M.P. with $p(y|x)$ as (**) will have a stat. dist. of $\pi^*(\cdot)$.

Algorithm: (a) generate $Y \sim q(\cdot|x)$

(b) generate $U \sim \text{Unif}[0,1]$.

(c) Compute $\alpha(x,y)$ as above.

If $U \leq \alpha(x,Y)$ accept Y,

and $x \longrightarrow Y$

else reject Y, and $x \longrightarrow x$.

Let $A = \mathcal{X}_{\{U \leq \alpha(x,Y)\}}$. (acceptance r.v.)

Note that the above algo. results in a M.P. For any $y \neq x$, we have:

(Notation: $\{Y \approx y\} = \{Y \in [y, y+dy]\}$)

$$p(y|x)\,dy = P\left(X_{t+1} \in [y, y+dy] \,\middle|\, X_t = x\right)$$

$$= \underbrace{P\left(X_{t+1} \in [y, y+dy] \,\middle|\, X_t = x, A=1\right)}_{= q(y|x)\,dy} \cdot P\left(A=1 \,\middle|\, X_t = x\right)$$

$$+ \cancel{P\left(X_{t+1} \in [y, y+dy] \,\middle|\, X_t = x, A=0\right)} \; P\left(A=0 \,\middle|\, X_t = x\right)$$

$$= 0$$

$$= P\left(A=1 \,\middle|\, X_t = x\right) \cdot q(y|x)\,dy$$

$$= \underbrace{P\left(U \leq \alpha(x,y) \,\middle|\, X_t = x\right)}_{\alpha(x,y)} q(y|x)\,dy$$

$$= q(y|x)\,\alpha(x,y)\,dy. \qquad \text{(F)}$$

Rate of convergence: This D.T.M.P converges geometrically fast in time if we can construct a Lyapunov function $V(\cdot)$, as discussed above.

(See [3] for more details; Meyn & Tweedie's book [5]).

We can then establish:

$$d_{TV}(p_t, \pi^*) \leq \rho^t \quad \text{for some } \rho \in (0,1)$$

↳ density at time t.

This look great, but not clear what the dimension dependence of ρ is. In general, it is very bad (ρ can be very close to '1', with the dependence being exponential in d, if $x \in \mathbb{R}^d$)

Gibbs Sampler $x \in \mathbb{R}^d$, sample $x \sim \pi^*(x)$

(can be viewed as a special choice of $q(\cdot|\cdot)$ in the Metropolis-Hastings sampler such that $\alpha(\cdot,\cdot)=1$, i.e., **no rejections**.

$$\pi^*(x_1, x_2, \cdots, x_d) = \pi^*(x_i \mid \underbrace{x_1, \cdots x_{i-1}, x_{i+1}, \cdots, x_d}_{x_{-i}})$$
$$\cdot \pi^*(x_1, \cdots x_{i-1}, x_{i+1}, \cdots x_d)$$

for each $i = 1, 2, \cdots, d$

Algorithm: Suppose $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \cdots, x_d^{(t)})$

(a) Pick a random coordinate: $v \in \{1, 2, \cdots, d\}$, $V \sim r(v)$, s.t. $r(v) > 0 \ \forall v = 1, 2, \cdots, d$.

(b) Update the random coordinate:

$$x^{(t+1)} = (x_1^{(t+1)}, x_2^{(t+1)}, \cdots, x_d^{(t+1)}), \text{ where}$$

$$x_v^{(t+1)} \sim \pi^*(x_v \mid x_{-v}^{(t)}), \quad x_{-v}^{(t)} = (x_1^{(t)}, \cdots, x_{v-1}^{(t)}, x_{v+1}^{(t)}, \cdots x_d^{(t)}).$$

For all other coordinates, $x_w^{(t+1)} = x_w^{(t)}$.

Observe that for $x = (x_1, x_2, \cdots, x_d)$, the next step y is of the form $y = (x_1, x_2, \cdots, x_{v-1}, y_v, x_{v+1}, \cdots, x_d)$

Depending on y, $q(y|x)$ is one of the following:

For y s.t. the v^{th} coordinate changes,

$$q(y|x) = r(v) \cdot \pi^*\left(y_v \,|\, x_1, x_2, \cdots, x_{v-1}, x_{v+1}, \cdots, x_d\right) \cdot \chi_{\{y_{-v} = x_{-v}\}}.$$

Now, observe for any v,

$$\pi^*(x)\, q(y|x) = \pi^*(x) \cdot r(v) \cdot \frac{\pi^*(x_1, x_2, \cdots \overbrace{y_v, \cdots x_d}^{y})}{\int \pi^*(x_1, \cdots, z, x_{v+1}, \cdots, x_d)\, dz}$$

$$= \frac{\pi^*(x) \cdot \pi^*(y) \cdot r(v)}{\tilde{\pi}^*(x_1, \cdots x_{v-1}, x_{v+1}, \cdots x_d)}$$

marginalize out x_v.

Similarly $\pi^*(y) q(x|y) = \pi^*(y) r(v) \cdot \dfrac{\pi^*(x)}{\tilde{\pi}^*(y_1, \cdots y_{v-1}, y_{v+1}, \cdots y_d)}$

$= (x_1, x_2 \cdots x_{v-1}, x_{v+1}, \cdots x_d)$

because $x_i = y_i \quad \forall i \neq v.$

$\therefore \pi^*(x) q(y|x) = \pi^*(y) q(x|y)$

$\Rightarrow \alpha(x, y) = 1$ ∎