Introduction to Computational Statistics

First Concepts in Data Simulation – juliette.chevallier@insa-toulouse.fr

Spring School on Statistics and Machine Learning

plmlab.math.cnrs.fr/chevallier-teaching/hcmus-springschool-computationalstatistics

All materials for the course are available at

plmlab.math.cnrs.fr/chevallier-teaching/hcmus-springschool-computationalstatistics

Remarks and Outlines

- **Aim**: Given a probability distribution π, more or less well known
  $\rightsquigarrow$ How sampling according to π?

- **Remark**: Here, we assume that we are able to sample according to $\mathcal{U}([0,1])$
  $\rightsquigarrow$ pseudo-random number generator (pRNG)

Program

1. Basic methods
2. Monte Carlo Markov Chain (MCMC) methods

Basic Simulation Methods

Definition (Cumulative distribution function)

$$F : \mathbb{R} \to [0,1]$$
$$x \mapsto F(x) = \mathbb{P}(X \leqslant x)$$

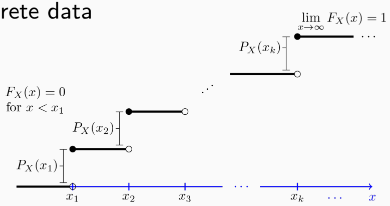
Theorem (Properties of F)

(i) *F is growing*

(ii) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$

(iii) *F is càdlàg: "right continuous with left limits"*

⚠ F not necessarily bijective ⇝ See discrete data



Definition (Quantiles)

Generalized inversion of F

$$F^{-1} :]0,1[\to \mathbb{R}$$
$$p \mapsto \inf\{x \in \mathbb{R} \text{ s.t. } F(x) \geqslant p\}$$

Theorem (Properties of F^{-1})

(i) *F^{-1} is growing, continuous on the left*

(ii) $F(x) \geqslant p \iff x \geqslant F^{-1}(p)$

(iii) $F\left(F^{-1}(p)\right) \geqslant p$ *with equality if F continuous in p*

Inversion Method: For real-valued data only $\equiv$ Values in $\mathbb{R}$

Theorem (Inversion Method)

- *Let* $U \sim \mathcal{U}([0,1])$. *Then* $\boxed{X = F^{-1}(U) \sim F}$
- *If* $X \sim F$ *and* F *is continuous. Then* $\boxed{F(X) \sim \mathcal{U}([0,1])}$

Proof:

- $\mathbb{P}\left(F^{-1}(U) \leqslant x\right) \overset{\text{(ii)}}{=} \mathbb{P}\left(F(x) \geqslant U\right) \overset{U \text{ uniform}}{=} F(x)$
- $$\mathbb{P}\left(F(X) < u\right) = 1 - \mathbb{P}\left(F(X) \geqslant u\right) \overset{\text{(ii)}}{=} 1 - \mathbb{P}\left(X \geqslant F^{-1}(u)\right)$$
  $$= \mathbb{P}\left(X < F^{-1}(u)\right) \overset{F\, \mathcal{C}^0}{=} F\left(F^{-1}(u)\right) \overset{F\, \mathcal{C}^0}{=} u$$

Inversion Algorithm

To sample $X \sim F$

1. Sample $U \sim \mathcal{U}([0,1])$
2. Set $X = F^{-1}(U)$

Expl: $X \sim \mathcal{E}xp(\lambda)$

$F(x) = 1 - \mathrm{e}^{-\lambda x}$ continuous and invertible

$F^{-1}(u) = -\dfrac{1}{\lambda}\log(1-u)$

Note: $U \sim \mathcal{U}([0,1]) \implies 1 - U \sim \mathcal{U}([0,1])$

$\rightsquigarrow F^{-1}(u) = x = -\frac{1}{\lambda}\log(u)$

How to compute F^{-1}?

1. **Explicit formula**

2. **Numerical resolution** of $F(x) = u$
   $\rightsquigarrow$ Bisection algorithm
   - If $x \in [a, b]$, we compare $F\left(\frac{a+b}{2}\right)$ to u
   - We get into the correct half-interval
   - We iterate

3. **Newton-Raphson** algorithm
   $\rightsquigarrow$ ⚠ Stop conditions

4. **Approximation** of F^{-1}
   $\rightsquigarrow$ Gaussian variable: Let A and B be 2 polynomial function

$$g(u) = \sqrt{-2\ln u} + \frac{A\left(\sqrt{-2\ln u}\right)}{B\left(\sqrt{-2\ln u}\right)}$$

Basic Simulation Methods

Uniform Distribution on $\mathcal{C}$ 1/2

Theorem (Uniform Distribution on $\mathcal{C}$ – Thm 1)

(a) *Denote* $\mathcal{C} = \Big\{(x,y) \text{ s.t. } 0 \leqslant y \leqslant cf(x)\Big\}$.

Assume:

- $X \sim f$, *with* f *density on* $\mathbb{R}^d$
- $U \sim \mathcal{U}([0,1])$ *such that* $U \perp\!\!\!\perp X$
- $c > 0$

Then: $\boxed{\big(X, c\,Uf(X)\big) \sim \mathcal{U}(\mathcal{C})}$

(b) *Conversely if* $(X,Y) \sim \mathcal{U}(\mathcal{C})$ *then* X *admits* f *for density*

Proof: Since $X \perp\!\!\!\perp U$, (X,U) admits for density $f(x)\mathbb{1}_{[0,1]}(u)\,\mathrm{d}x\,\mathrm{d}u$

(a)
$$\mathbb{E}\Big[h(X, c\,Uf(X))\Big] = \iint h\Big(x, c\,u f(x)\Big) f(x)\mathbb{1}_{[0,1]}(u)\,\mathrm{d}x\,\mathrm{d}u$$
$$= \iint h(x,y) f(x)\mathbb{1}_{\mathcal{C}}(x,y)\frac{\mathrm{d}y}{c\,f(x)}\,\mathrm{d}x \qquad \text{setting } y = c\,u f(x)$$

Note that $\displaystyle\int_{\mathcal{C}} \mathrm{d}x\,\mathrm{d}y = \iint_0^{cf(x)} \mathrm{d}y\,\mathrm{d}y = c\int f(x)\,\mathrm{d}x = c$

$\rightsquigarrow$ c normalization constant and $(X,Y) \sim \mathcal{U}(\mathcal{C})$

(b) Conversely $\displaystyle f_X(x) = \int f_{(X,Y)}(x,y)\,\mathrm{d}y = \frac{1}{c}\int_0^{cf(x)} \mathrm{d}y = f(x)$

Theorem (Uniform Distribution on $\mathcal{C}$ – Thm 1)

(a) *Denote* $\mathcal{C} = \Big\{(x, y)$ *s.t.* $0 \leqslant y \leqslant cf(x)\Big\}$.

Assume:

- $X \sim f$, *with* f *density on* $\mathbb{R}^d$
- $U \sim \mathcal{U}([0,1])$ *such that* $U \perp\!\!\!\perp X$
- $c > 0$

Then: $\boxed{\big(X, cUf(X)\big) \sim \mathcal{U}(\mathcal{C})}$

(b) *Conversely if* $(X, Y) \sim \mathcal{U}(\mathcal{C})$ *then* X *admits* f *for density*

Algorithm: Uniform distribution on $\mathcal{C}$, *i.e.* below the curve g

1. Sample $X \sim \dfrac{g(x)}{\int g(x)\,\mathrm{d}x}$
2. Sample $U \sim \mathcal{U}\big([0,1]\big)$ and set $Y = g(x) \times u$
3. Return (X, Y)

Requirements for the Accept-Reject Method

Theorem (Requirements for the accept-reject method – Thm 2)

- *Let* $(X_k)_{l\in\mathbb{N}}$ *a seq. of random variables valued in* $\mathbb{R}^d$, *same law than* X
- *Let* $A \in \mathcal{X}$ *s.t.* $\mathbb{P}(X \in A) = p > 0$
- *Let* $Y = X_{\tau_A}$, *where* $\tau_A = \inf\{k \geqslant 1 \text{ s.t. } X_k \in A\}$ *"First instant in* A*"*

Then: $\forall B \in \mathcal{X}$, $\boxed{\mathbb{P}(Y \in B) = \mathbb{P}(X \in B \,|X \in A)}$, *conditional proba knowing* A

Proof: Since τ_A is the first time in A,

$$\mathbb{E}\Big[h(Y)\Big] = \sum_{n\geqslant 1} \mathbb{E}\Big[h(X_n)\mathbb{1}_{\{\tau=n\}}\Big] = \sum_{n\geqslant 1} \mathbb{E}\Big[h(X_n)\mathbb{1}_{\{X_1\notin A\}} \dots \mathbb{1}_{\{X_{n-1}\notin A\}}\,\mathbb{1}_{\{X_n\in A\}}\Big]$$

$$\overset{i.i.d}{=} \sum_{n\geqslant 1} \mathbb{P}(X_1 \notin A) \dots \mathbb{P}(X_{n-1} \notin A)\,\mathbb{E}\Big[h(X_n)\mathbb{1}_{\{X_n\in A\}}\Big]$$

$$\overset{X_i\sim X}{=} \sum_{n\geqslant 1} \Big(\mathbb{P}(X \notin A)\Big)^{n-1} \mathbb{E}\Big[h(X_n)\mathbb{1}_{\{X_n\in A\}}\Big] = \left(\sum_{n\geqslant 1}(1-p)^{n-1}\right) \mathbb{E}\Big[h(X_n)\mathbb{1}_{\{X_n\in A$$

And the result follows with $h = \mathbb{1}_{\{X\in B\}}$

$$\rightsquigarrow \quad \mathbb{P}(Y \in B) = \frac{1}{1-(1-p)}\mathbb{P}(X \in A \cup B) = \frac{\mathbb{P}(X \in A \cap B)}{\mathbb{P}(X \in A)}$$

Accept-Reject Method

Corollary

Let $X \sim \mathcal{U}(\mathcal{C})$ *where* $\mathcal{C} \subset \mathbb{R}^d$. *Let* $\mathcal{B} \subset \mathcal{C}$. ***Then:*** $\boxed{Y = X_{\tau_{\mathcal{B}}} \sim \mathcal{U}(\mathcal{B})}$

Idea Accept-Reject Method:

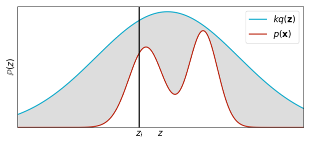
- Consider a density g s.t. $\exists c > 0$ s.t. $\forall x \in \mathbb{R}^d$, $f(x) \leqslant c\, g(x)$
- [Thm 1] Sample $U \sim \mathcal{U}(\mathcal{C})$, where $\mathcal{C} = \{(x, y) \text{ s.t. } 0 \leqslant y \leqslant c\, g(x)\}$
- [Thm 2] As long as U is not in $\mathcal{B} = \{(x, y) \text{ s.t. } 0 \leqslant y \leqslant f(x)\}$, draw new U
- [Cor] As soon as U falls into $\mathcal{B}$, we know that its abscissa $X \sim f$

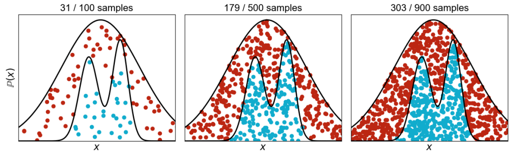
Algorithm: Accept-Reject

While $Y \geqslant f(x)$

1. Sample $X \sim g \quad (\implies x)$
2. Sample $U \sim \mathcal{U}\big([0,1]\big)$
   Set $Y = c\, g(x)\, U$

Return X





Images credit: Gregory Gundersen

⚠ *Easy simulation according to g is essential, and the choice of g is crucial!*

Question (Waiting time): What is the proba distribution of τ_A?

- $\mathbb{P}(\tau_A = n) \overset{\perp\!\!\!\perp}{=} \mathbb{P}(X_1 \notin A) \dots \mathbb{P}(X_{n-1} \notin A)\mathbb{P}(X_n \in A) \overset{i.i.d.}{=} (1-p)^{n-1}p$

  $\rightsquigarrow$ $\tau_A \sim \mathcal{G}eom(p)$ with $p = \mathbb{P}(X \in A)$

  In particular, $\mathbb{E}[\tau_A] = \dfrac{1}{p}$ and $\mathcal{V}ar(\tau_A) = \dfrac{1-p}{p^2}$

- [Reject Algo] "While $U > \dfrac{f(x)}{c\,g(x)}$ sample $X \sim g$ and $U \sim \mathcal{U}([0,1])$"

$$p = \mathbb{P}\left(U \leqslant \frac{f(x)}{c\,g(x)}\right) = \int \mathbb{1}_{\{ucg(x) \leqslant f(x)\}} \mathbb{1}_{[0,1]}(u) g(x)\,\mathrm{d}x\,\mathrm{d}u = \int \frac{f(x)}{c\,g(x)} g(x)\,\mathrm{d}x = \frac{1}{c}$$

$\rightsquigarrow$ The greater the c, the smaller the p and therefore the greater the average number of draws before success $\mathbb{E}[\tau_A] = \frac{1}{p} = c$

$\Longrightarrow$ *Less efficient algorithm*

Calibration of the Method 2/2

Choice of (c, g)

- $f > 0 \implies g > 0$
- $c = \sup_{x \text{ s.t. } f(x)>0} \frac{f(x)}{g(x)}$
- g belongs to a parametric family, and we optimize the parameter

---

Evaluating f can be costly

$\rightsquigarrow$ Major and minor f with easily evaluable functions to make comparisons

Exercise: Box-Muller Algorithm

- Let R a random variable with Rayleigh distribution with parameter 1

$$\forall r \in \mathcal{R}, \qquad f_R(r) = r \exp\left(-\frac{r^2}{2}\right) \mathbb{1}_{\mathbb{R}+}(r)$$

- Let Θ with uniform distribution on $[0, 2\pi]$
- Assume that R and Θ are independent

1. Let X and Y such that

$$X = R\cos(\Theta) \qquad \text{and} \qquad Y = R\sin(\Theta)\,.$$

   Prove that both X and Y have $\mathcal{N}(0, 1)$ distribution and are independent.

2. Write an algorithm for sampling independent Gaussian distribution $\mathcal{N}(0, 1)$.

The Metropolis-Hastings Algorithm

Motivation

- The Metropolis-Hastings (MH) algorithm manage to sample from "almost any" probability distribution (up to the burn-in period) in a very simple way
- It only requires to know the target π up to a multiplicative constant (usually the normalization constant)

Aim and setting:

- Denote π our *target distribution*, which we assume to have a pdf w.r.t the Lebesgue measure
- We know π up to a multiplicative constant
- Assume that we are able to sample from a proposal density $q(x, \cdot)$

The Metropolis-Hastings Algorithm

1. Start with an initial value X_0
2. Given X_n, we propose a candidate $Y_n \sim q(X_n, \cdot)$ from the *proposal distribution* $q(x, \cdot)$
3. Compute the *acceptation ratio*

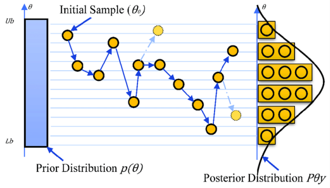
$$\boxed{\alpha(X_n, Y_{n+1}) = \begin{cases} \min\left(1\,;\ \dfrac{\pi(Y_{n+1})\, q(Y_{n+1}, X_n)}{\pi(X_n)\, q(X_n, Y_{n+1})}\right) & \text{if } denom > 0 \\ 1 & \text{otherwise} \end{cases}}$$

4. If we accept Y_{n+1}, set $X_{n+1} = Y_{n+1}$. Otherwise $X_{n+1} = X_n$

Algorithm: Metropolis Hastings

For $n = 1 \ldots N$

- $U_{n+1} \sim \mathcal{U}([0,1])$
- $Y_{n+1} \sim q(X_n, \cdot)$
- If $U_{n+1} < \alpha(X_n, Y_n)$
  Then $X_{n+1} = Y_{n+1}$
  Otherwise Then $X_{n+1} = X_n$



Interpretation: We accept move $X_n \mapsto Y_{n+1}$ when the "likelihood ratio" $\frac{\pi(Y_{n+1})}{q(X_n, Y_{n+1})}$ is greater than the one for the opposite move $Y_{n+1} \mapsto X_n$

Transition Kernel, Ergodicity, etc.

The MH algorithm produces a Markov chain of *transition kernel*

$$\boxed{P(x,A) = \int_A q(x,y)\,\mathrm{d}y + \delta_x(A)\int_{\mathbb{X}} (1-\alpha(x,y))q(x,y)\,\mathrm{d}y}$$

Intuition

- Il we accept y which has been proposed $p(x,y) = q(x,y)\,\alpha(x,y)$
- If we have rejected, we stay at x and $p(x,\{x\}) = \int q(x,y)\big(1-\alpha(x,y)\big)\,\mathrm{d}y$

Invariant measure $\boxed{\pi P = \pi}$

To ensure that the MCMC chain converges to the correct minima, we need the chain's ergodicity $\rightsquigarrow$ Depends on the choice of the proposal!

Remark: $\lim\limits_{n\to\infty} \sup\limits_{A\in\mathcal{X}} |\mu P^n(A) - \pi(A)| = 0$ and π is independant of μ

The Metropolis-Hastings Algorithm

Independence sampling – IS-MH

$$\boxed{q(x,y) = q(y)}$$

- The proposal does not depend on the current state of the chain
- **Remark**
  - The dependence is present through the acceptation ratio $\alpha(x,y) = \frac{\pi(y)\,q(x)}{q(y)\,\pi(x)} \wedge 1$
  - Looks like a generalization of the accept-jerect method
  - Convergence properties of the generated chain obviously depends on Under certain conditions on q

---

Definition (Total Variation norm)

Use notation P for botch the kernel and its corresponding probability measure

$$\boxed{\|P(x,\cdot) - \pi\|_{TV} \;=\; \sup_{A\in\mathcal{X}} \left| \int_A \big(P(x,y) - f(y)\big)\,\mathrm{d}y \right|}$$

Convergence of the Independence Sampling

Following theorem provides simple conditions to prove strong convergence property *i.e.* the geometric ergodicity

Theorem

- *The MCMC algorithm given by the Independence Sampling (IS) MH algorithm produces a uniformaly ergodic Markov Chain if:*

  (M) *There exists a constant* $M > 0$ *s.t. that* $\forall x \in \mathcal{S}upp(\pi)$, $\pi(x) \leqslant Mq(x)$

  In this case $\boxed{\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \leqslant \left(1 - \frac{1}{M}\right)^n}$

- *Moreover if* $\forall M > 0\ \exists A \in \mathcal{X}$ *s.t.* $\pi(A) \neq 0$ *and (M) is not satisfied, then* (X_n) *is not geometrically ergodic*

Efficiency of the Independence Sampling ?

Measures through its mean acceptance ratio $\mathbb{E}[\alpha(X,Y)]$

Lemma

If the chain is stationnary and q *satisfies (M) then* $\boxed{\mathbb{E}[\alpha(X,Y)] \geqslant \frac{1}{M}}$

Example: $\pi = \mathcal{N}(0,1)$ from $q = \mathcal{N}(0,\sigma^2)$

Example Simulation of $\pi = \mathcal{N}(0,1)$ using $q = \mathcal{N}(0,\sigma^2)$

$$\begin{aligned}\alpha(x,y) &= \inf\left\{1\,,\,\exp\left(-\frac{1}{2}y^2+\frac{1}{2}x^2+\frac{1}{2\sigma^2}y^2-\frac{1}{2\sigma^2}x^2\right)\right\}\\ &= \inf\left\{1\,,\,\exp\left(-\frac{1}{2}\left(y^2-x^2\right)\left(1-\sigma^{-2}\right)\right)\right\}\end{aligned}$$

- If $\sigma^{-1} > 1$:
  - The proposal have heavier tail than the target distribution
  - We will propose many elements far from 0, *i.e.* such that $|Y_{n+1}| > |X_n|$
  - But $1 - \frac{1}{\sigma^2} > 0$. Hence:
    - If $Y_{n+1}^2 >> X_n^2$, $\exp\left(-\frac{1}{2}\left(y^2-x^2\right)\left(1-\sigma^{-2}\right)\right) < 1$
      (and the more $Y_{n+1}^2 >> X_n^2$, the more $\exp(\cdot) < 1$)
      $\rightsquigarrow$ These elements are very likely to be reject
    - The elements s.t. $Y_{n+1}^2 \leqslant X_n^2$ will be accepted with probability 1

  To compensate the heavy tail proposal, many elements will be rejected

- If $\sigma^{-1} < 1$:
  - All the candidates s.t. $|Y_{n+1}| > |X_n|$ will be accepted with probability 1
  - But: not all elements in the complementary set

  This compensate that many elements will be proposed close to 0

The Metropolis-Hastings Algorithm

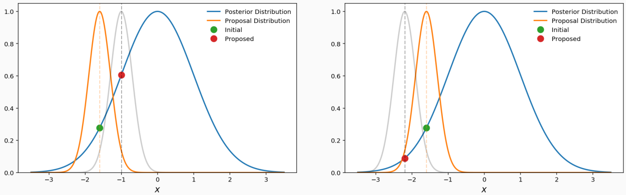
Symmetric Random Walk – SRW-MH

- Assume that $\mathbb{X}$ is e vector space (typically $\mathbb{R}^d$) on which we can define a random walk with transition density $q(x,y) = q(|x-y|)$, q even function
- In this case, the acceptation ratio reduce to $\boxed{\alpha(x,y) = \inf\left(1, \frac{\pi(y)}{\pi(x)}\right)}$

Interpretation:

- In the *usual MH*, with accept the Y_{n+1} s.t. that the "likelihood ratio" $\frac{\pi(\cdot)}{q(X_n,\cdot)}$ is greater than the opposite one
- **Here**: Condition becomes π increases $\Longrightarrow$ Accept with probability 1

  In other words, *the SRW-MH biases the Random Walk by promoting samples which move toward the modes of* π



Considerations around the SRW-MH

- Large range of possible $q \rightsquigarrow$ Very versatile algorithm
- But some drawbacks

Theorem

If π has non-compact support and q symmetric, the Markov chain $(X_n)_n$ is not uniformly ergodig

- Additional assumptions on π and q allow geometric ergodicity

---

Example: Simulation of $\pi(x) \propto \exp(-\gamma|x|^{\beta})$, $\gamma > 0$ and $\beta > 0$,
From $q(x,y) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)$

- Given X_n, the SRW samples $Y_{n+1} \sim \mathcal{N}(X_n, \sigma^2)$

$$\rightsquigarrow \quad \alpha(X_n, Y_{n+1}) \;=\; \inf\left\{1\,,\, \exp\Big(\gamma\left(|X_n|^{\beta} - |Y_{n+1}^{\beta}|\right)\Big)\right\}$$

- If $|Y_{n+1}| \leqslant |X_n|$, automatically accept
- If $|Y_{n+1}| \leqslant |X_n|$, it may be rejected

Remark: It seems that σ^2 does not have a real impact on the results. In fact, il have but in the ability of the chain to "forget its past"

Auto-Correlation (as a “measure” of stationarity)

Definition (Auto-Correlation)

Consider a chain extended to the N-th iteration and denote $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

$$\boxed{C_i = \frac{\sum_{j=1}^{N-i}(X_j - \bar{X})(X_{j+i} - \bar{X})}{N-i} \times \frac{N}{\sum_{j=1}^{N}(X_j - \bar{X})^2}}$$

- The auto-correlation measures how “independent” are the first samplers w.r.t. the last ones
- **Expectation**: It decreases quickly
  - $\rightsquigarrow$ The chain does not depends too long on its initial value

To help the chain to “forget its past”, one may also rely on a burn-in period

The Metropolis-Hastings Algorithm

Intuition behind MALA

- Recall that the SRW-MH can be seen as a Random-Walk biased to promote samples around the modes of π

  Can we go further, *i.e.* better use the target π into the proposal?

- Writes π, known up to a constant, in the form $\boxed{\pi(x) = \dfrac{\mathrm{e}^{-U(x)}}{\int_{\mathbb{R}^d} \mathrm{e}^{-U(y)}\,\mathrm{d}y}}$

- Assume that $U : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^2$ and $\exists L > 0$ s.t.

  $\forall x, y \in \mathbb{R}^d,\ \|\nabla U(x) - \nabla U(y)\| \leqslant L\|x - y\|$

(Overdamped) Langevin diffusion SDE:

Denote $(B_t)t \geqslant 0$ a d-dimensional Brownian Motion

$$\mathrm{d}Y_t = \nabla U(Y_t)\,\mathrm{d}t + \sqrt{2d}B_t = -\nabla \log \pi(Y_t)\,\mathrm{d}t + \sqrt{2d}B_t$$

Properties of the Langevin SDE (admitted)

Theorem

1. $\pi(x) \propto \exp(-U(x))$ *is the unique invariant probability measure*
2. *For all* $x \in \mathbb{R}^d$, $\lim_{t \to +\infty} \|\delta_x P_t - \pi\|_{TV} = 0$
3. *For "nice" (i.e. sufficiently regular) functions* f

$$\frac{1}{T}\int_0^T f(X_t)\,\mathrm{d}t \xrightarrow{\mathbb{P}_x - a.s.} \pi(f) = \int f(x)\pi(\mathrm{d}x)$$

$$\frac{1}{\sqrt{T}}\int_0^T (f(X_t) - \pi(f))\,\mathrm{d}t \xrightarrow{\mathbb{P}_x} \mathcal{N}\big(0, \sigma^2(\pi, f)\big)$$

In other words, Langevin diffusion provides a way to sample any smooth distribution
Of course, this is a highly theoretical solution!

$\rightsquigarrow$ *Discretized Langevin diffusion*

Metropolis-Adjusted Langevin Algorithm – MALA

Sample the diffusion paths using the Euler–Maruyama (EM) scheme:

$$X_{k+1} = X_k - \gamma_{k+1}\nabla U(X_k) + \sqrt{2\gamma_{k+1}}Z_{k+1} = X_k + \gamma_{k+1}\nabla \log \pi(X_t) + \sqrt{2\gamma_{k+1}}Z_{k+1}$$

where

- $(Z_k)_{k\geqslant 1}$ is a sequence of $i.i.d.$ standard Gaussian random vectors
- $(\gamma_k)_{k\geqslant 1}$ is a sequence of stepsizes, which can either be held constant or be chosen to decrease to 0 at a certain rate

Remark: Can be viewed as a noisy (stochastic) gradient descent in some sense

---

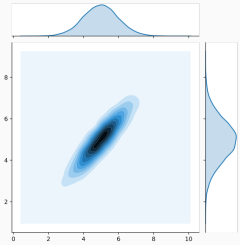
Metropolis-Adjusted Langevin Algorithm:

$\rightsquigarrow$ Use this discretization scheme as a proposal in the MH algorithm

1. Propose $\boxed{Y_{n+1} \sim \mathcal{N}\big(X_n + \gamma\nabla \log \pi(X_n), 2\gamma\big)}$
2. Compute the acceptance ratio $\alpha(X_n, Y_{n+1})$
3. Accept/Reject the proposal

The Gibbs Sampler

3.1 Principle



Idea: Substitute the sampling of a r.v. in $\mathbb{R}^d$ by the sampling of d r.d. in $\mathbb{R}$ ()or at least smaller

Example (cf figure):

- Narrow shape in $\mathbb{R}^2$ that induce many rejection with a MH algorithm
- Larger possibilities in $\mathbb{R}$

Algorithm – Assumptions:

- $\mathbb{X}$ can be decomposed into $\mathbb{X}_1 \times \dots \mathbb{X}_k$
  We denote x^i the element of the i-th bloc, and $x^{(-i)}$ the vector x where i have been removed
- We know how to sample from the conditional distribution of the bloc i given the other while targeting π: $X_i|X_{(-i)} \sim \pi_i(x^i|x^{(-i)}) = \frac{\pi(x)}{\pi_{(-i)}(x)}$

Algorithm – Pseudo-Code: Given X_0, iterate:

For $j = 1 \to k$,
For $n = 1 \to N$

$$\boxed{X_{n+1}^j \sim \pi_j(\cdot|X_{n+1}^1, \dots, X_{n+1}^{j-1}, X_n^{j+1}, \dots, X_n^k)}$$