Monte-Carlo simulation

Sampling and estimation

É. Savin[1,2]
eric.savin@centralesupelec.fr

[1]Information Processing and Systems Dept.
ONERA, France

[2]Mechanical and Environmental Engineering Dept.
CentraleSupélec, France

Outline

When does sampling occur?

- The i.i.d. random stiffnesses of a beam (lecture #1);
- Classical r.v.: exponential, Gamma, β, β', log-normal... (lecture #2);
- Polynomial chaos expansion for the representation of second-order r.v. (lecture #2);
- $\boldsymbol{X} \sim p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathrm{e}^{-\lambda_0 - \boldsymbol{\lambda}\cdot\boldsymbol{C}(\boldsymbol{x})}$ derived from the MaxEnt principle (lecture #2);
- Karhunen-Loève expansion of second-order random processes (lecture #3);
- Spectral expansion of stationary second-order random processes (lecture #4).

Outline

Inverse transform method

Principle

Definition

Uniform probability law $\mathcal{U}(0,1) = \mathbf{1}_{[0,1]}(u)\mathrm{d}u$ *and:*

$$m_U = \underline{U} = \frac{1}{2}\,, \quad \sigma_U = \sqrt{\mathbb{E}\{(U - m_U)^2\}} = \frac{1}{2\sqrt{3}}\,.$$

Theorem

Let F *be the distribution function of a continuous r.v.* X*. Then if* $U \sim \mathcal{U}(0,1)$ *the r.v.* $F^{-1}(U)$ *has distribution function* F *and the r.v.* $F(X)$ *has a uniform probability law.*

- $P(F^{-1}(U) \leq x) = P(U \leq F(x)) = \int_0^{F(x)} \mathrm{d}u = F(x)\,.$
- $P(F(X) \leq u) = P(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$ with the generalized inverse (or quantile) function:

$$F^{-1}(u) = \inf\{x | F(x) \geq u\}\,, \quad 0 < u < 1\,.$$

Inverse transform method
Examples

PDF	F	$X = F^{-1}(U)$	Equivalent form
Exponential ($\sigma > 0$)			
$\frac{1}{\sigma} \mathrm{e}^{-\frac{x}{\sigma}} \mathbf{1}_{\mathbb{R}_+}(x)$	$1 - \mathrm{e}^{-\frac{x}{\sigma}}$	$-\sigma \ln(1-U)$	$-\sigma \ln(U)$
Weibull ($\sigma, k > 0$)			
$\frac{k}{\sigma}(\frac{x}{\sigma})^{k-1} \mathrm{e}^{-(\frac{x}{\sigma})^k} \mathbf{1}_{\mathbb{R}_+}(x)$	$1 - \mathrm{e}^{-(\frac{x}{\sigma})^k}$	$\sigma(-\ln(1-U))^{\frac{1}{k}}$	$\sigma(-\ln(U))^{\frac{1}{k}}$
Cauchy			
$\frac{\sigma}{\pi[\sigma^2+(x-\mu)^2]}$	$\frac{1}{2} + \frac{1}{\pi}\arctan(\frac{x-\mu}{\sigma})$	$\mu + \sigma \tan \pi(U - \frac{1}{2})$	$\mu + \sigma \tan(\pi U)$
Rayleigh			
$\frac{x}{\sigma} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \mathbf{1}_{\mathbb{R}_+}(x)$	$1 - \mathrm{e}^{-\frac{x^2}{2\sigma^2}}$	$\sigma\sqrt{-\ln(1-U)}$	$\sigma\sqrt{-\ln(U)}$
Gamma ($k \in \mathbb{N}^*$)			
$\frac{1}{\sigma\Gamma(k)}(\frac{x}{\sigma})^{k-1} \mathrm{e}^{-\frac{x}{\sigma}} \mathbf{1}_{\mathbb{R}_+}(x)$	$\int_0^{\frac{x}{\sigma}} \frac{t^{k-1}}{\Gamma(k)} \mathrm{e}^{-t}\,\mathrm{d}t$	$-\sigma \sum_{j=1}^{k} \ln U_j$	
Triangular			
$\frac{2}{\sigma}(1 - \frac{x}{\sigma})\mathbf{1}_{[0,\sigma]}(x)$	$\frac{2}{\sigma}(x - \frac{x^2}{2\sigma})\mathbf{1}_{[0,\sigma]}(x)$	$\sigma(1 - \sqrt{1-U})$	$\sigma(1 - \sqrt{U})$

The case of Gaussian r.v.

- Let $G \sim \mathcal{N}(0,1)$, the distribution function is the error function (erf), which is difficult to inverse:

$$F(g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{g} \mathrm{e}^{-\frac{x^2}{2}} \, \mathrm{d}x \, .$$

- *Box-Muller algorithm*: Let $U_1, U_2 \sim \mathcal{U}(0,1)$ independent and

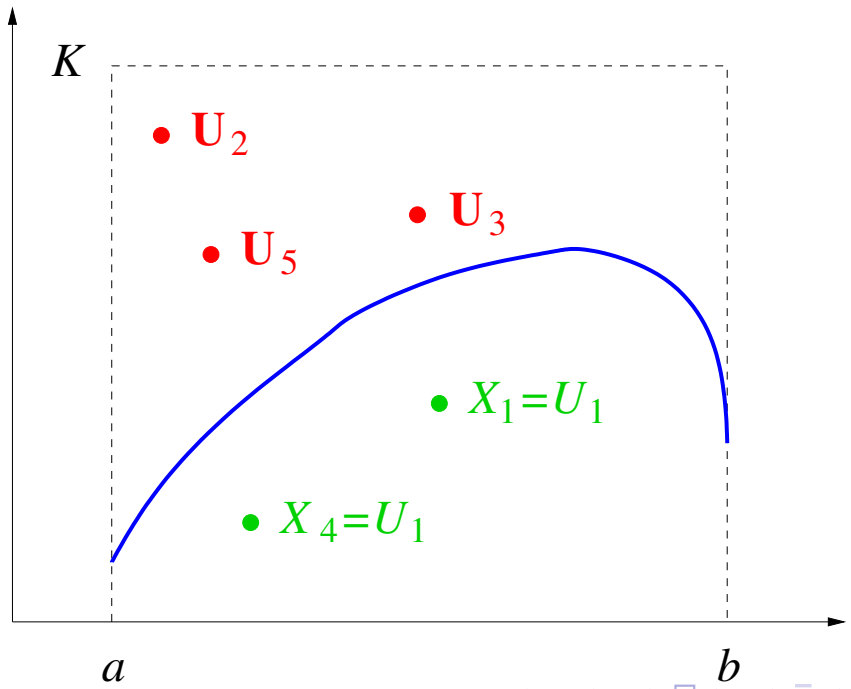
$$G_1 = \sqrt{-2 \ln U_1} \cos(2\pi U_2) \, ,$$
$$G_2 = \sqrt{-2 \ln U_1} \sin(2\pi U_2) \, ,$$

then G_1 and G_2 are i.i.d. normal r.v. $G_1, G_2 \sim \mathcal{N}(0,1)$.
By causality principle $p_G(g) = p_U(h^{-1}(g)) \det(\boldsymbol{\nabla}_g h^{-1}(g))$, and the inverse of $u \mapsto h(u)$ is $h^{-1}(g) = (\exp(-\|g\|^2/2), \frac{1}{2\pi} \arctan(g_2/g_1))$.

Outline

Rejection method

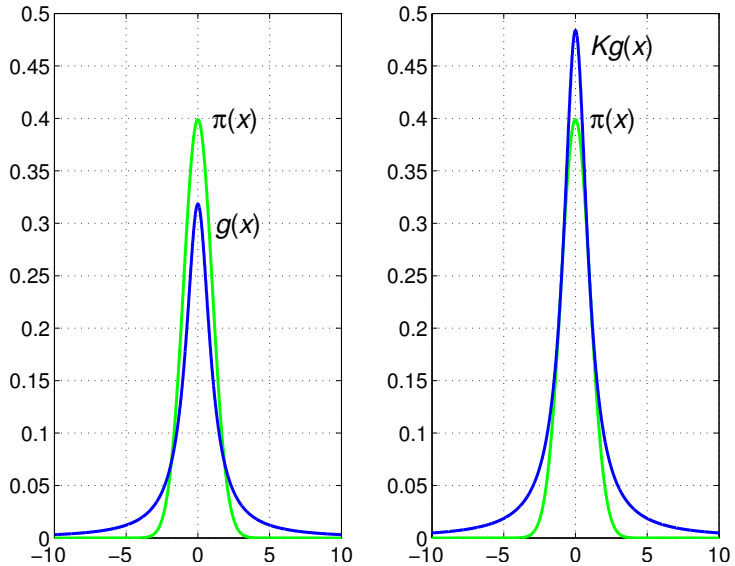
- Objective: simulate $X \sim \pi(x)\mathrm{d}x$ where $\pi(x)$ has compact support within $[a, b]$.
- Basic algorithm for $0 < \sup_{x\in[a,b]} \pi(x) \leq K$:
  1. $\boldsymbol{U} \sim \mathcal{U}([a, b] \times [0, K])$ a uniform r.v. on $[a, b] \times [0, K]$;
  2. `if` $U_2 \leq \pi(U_1)$ `then` $X = U_1$, `else goto` 1.



Rejection method

- Generalization for $0 < \pi(x) \leq Kg(x)$ where g is a known PDF s.t. $g \approx \pi$ and $G \sim g(x)\mathrm{d}x$ is easy to simulate:
  1. $U \sim \mathcal{U}(0,1)$ and $G \sim g(x)$ independently;
  2. `if` $K \times U \times g(G) < \pi(G)$ `then` $X = G$, `else goto` 1.
- The optimum choice $K = \sup_x \frac{\pi(x)}{g(x)}$, and K^{-1} is the probability of acceptance.

Example: simulate a normal r.v. from a Cauchy r.v.



Target: $\pi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$
Proposal: $g(x) = [\pi(1+x^2)]^{-1}$,
then $K = \sqrt{2\pi/\mathrm{e}}$ and use
$G = \tan(\pi U)$ with $U \sim \mathcal{U}(0,1)$.

Outline

Pseudo-random number generators PRNGs

Objectives

Generate a sequence of pseudo-random numbers of which distribution is close to $\mathcal{U}(0,1)$ with:

- good statistical properties and uniformness;
- long periods;
- efficiency vs. computational time;
- repeatability (in order to test programs);
- ease of implementation for all programming languages;
- unpredictability: it should be impossible to infer U_i from the knowledge of U_{i-1} (applications in cryptography).

Pseudo-random number generators PRNGs

Linear Congruential Generator

- Pseudo-random numbers $\{U_i\}_{0 \leq i \leq m-1}$ following a uniform law may be obtained by a *Linear Congruential Generator* $\mathrm{LCG}(a, b, m)$:

$$I_i = (aI_{i-1} + b) \bmod(m)\,, \quad I_0 = \texttt{SEED}\,,$$
$$U_i = \frac{I_i}{m}\,,$$

  with $0 < a < m$ (multiplier), $0 \leq b < m$ (increment), $0 < m$ (modulus), $0 \leq I_0 < m$ (SEED).

  Example: $a = 7^5$, $b = 0$ (*Multiplicative* Congruential Generator), $m = 2^{31} - 1$ is frequently used.

- For $\boldsymbol{U} \in [0, 1]^n$ with mutually independent components use n independent versions of U.

Pseudo-random number generators PRNGs

Multiple Recursive Generator

- Pseudo-random numbers $\{U_i\}_{0 \leq i \leq m-1}$ following a uniform law may be obtained by *Multiple Recursive Generator* $\mathrm{MRG}(k, m)$:

$$I_i = (a_1 I_{i-1} + \cdots + a_k I_{i-k}) \bmod(m)\,,$$
$$U_i = \frac{I_i}{m}\,,$$

  where $a_k \in \{-(m-1), \ldots (m+1)\}$.
- **Properties**:
  - The period is greater than m and the maximum period is $m^k - 1$ whenever m is a prime number;
  - Generalization to the multi-dimensional case.

Pseudo-random number generators PRNGs

Other examples

- *Combined Multiple Recursive Generator* `MRG32k3a` (L'Écuyer 1999):

$$I_i = (1403580 I_{i-2} - 810728 I_{i-3}) \bmod (2^{32} - 209)\,,$$
$$J_i = (527612 J_{i-1} - 1370589 J_{i-3}) \bmod (2^{32} - 22853)\,,$$
$$U_i = \frac{(I_i - J_i) \bmod (2^{32} - 209)}{2^{32} - 209}\,.$$

- *Mersenne Twister algorithm* (because its period is the Mersenne number $2^{19937} - 1$, see Matsumoto-Nishimura 1997): http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html

Pseudo-random number generators PRNGs

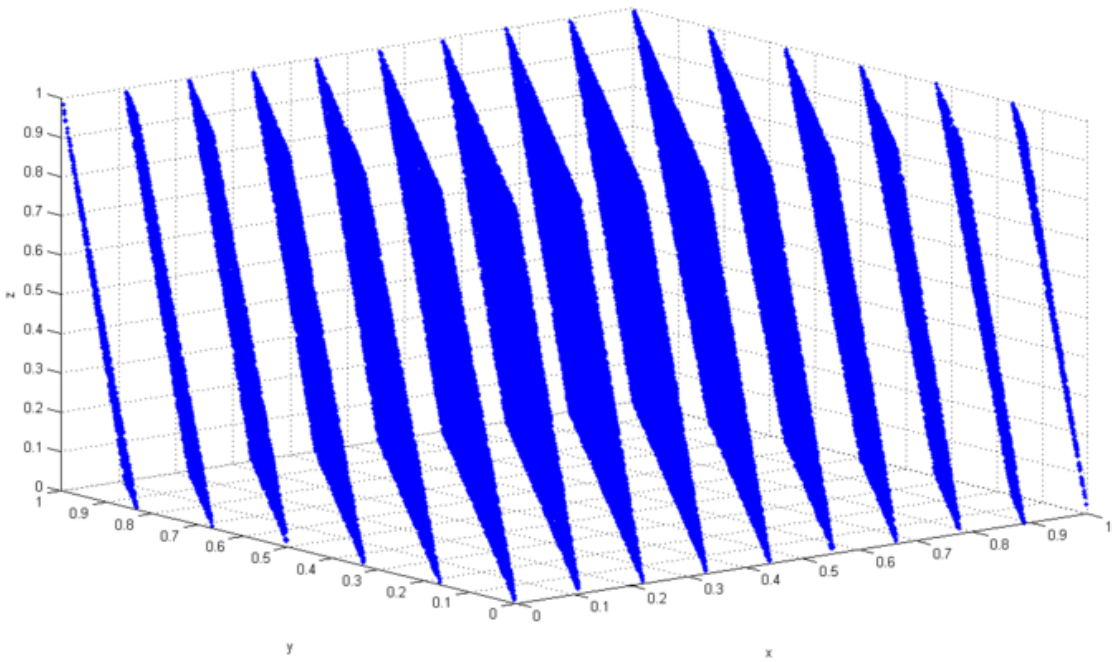
Other examples

- `RANDU` (IBM 1967):

$$I_i = 65539 I_{i-1} \mod(2^{31}) ,$$

$$U_i = \frac{I_i}{2^{31}} ,$$

with I_0 odd.



Outline

Definition

- *A Markov*[1] *chain* $(\boldsymbol{U}_m, m \in \mathbb{N})$ *is a sequence of r.v. with values in a finite or countable set* E *defined by:*
  - $E = \{\boldsymbol{u}_0, \boldsymbol{u}_1, \dots\}$,

$$P(\boldsymbol{U}_{m+1} = \boldsymbol{u}_{m+1} | \boldsymbol{U}_0 = \boldsymbol{u}_0, \boldsymbol{U}_1 = \boldsymbol{u}_1, \dots \boldsymbol{U}_m = \boldsymbol{u}_m) = P(\boldsymbol{U}_{m+1} = \boldsymbol{u}_{m+1} | \boldsymbol{U}_m = \boldsymbol{u}_m), \quad \forall m \in \mathbb{N};$$

  - *its initial probability law* $\pi_{0,j} = P(\boldsymbol{U}_0 = \boldsymbol{u}_j)$.
- *A homogeneous Markov chain:*

$$P(\boldsymbol{U}_{m+1} = \boldsymbol{u}_j | \boldsymbol{U}_m = \boldsymbol{u}_i) = P(\boldsymbol{U}_m = \boldsymbol{u}_j | \boldsymbol{U}_{m-1} = \boldsymbol{u}_i), \quad \forall m \in \mathbb{N}.$$

[1] Andreï Markov (1856–1922): Russian mathematician.

Markov chains
Examples

- Let $(\boldsymbol{V}_m,\, m \in \mathbb{N}^*)$ be a sequence of i.i.d. r.v. with values in F, and let $f : E \times F \to E$ be measurable. Then the sequence $(\boldsymbol{U}_m,\, m \in \mathbb{N})$ defined by:

$$\boldsymbol{U}_{m+1} = f(\boldsymbol{U}_m, \boldsymbol{V}_{m+1})\,, \quad \forall m \in \mathbb{N}\,,$$

  is an homogeneous Markov chain as soon as $\boldsymbol{V}_m$ is independent of $\boldsymbol{U}_0$.
- Random walk in $E = \mathbb{Z}^2$:

$$\boldsymbol{U}_{m+1} = \boldsymbol{U}_m + \boldsymbol{V}_{m+1}\,, \quad \forall m \in \mathbb{N}\,,$$

  with $\boldsymbol{V}_m \in F = \{(-1,0),(0,-1),(+1,0),(0,+1)\}$.

MCMC methods
Examples

- Sampling a Markov chain:
  1. $\boldsymbol{U}_0 \sim \boldsymbol{\pi}_0$;
  2. `repeat`
     - $\boldsymbol{U}_{m-1} = \boldsymbol{u}_i$
     - $U \sim \mathcal{U}(0,1)$ and find j s.t.:

$$\sum_{k=1}^{j-1} P(\boldsymbol{U}_m = \boldsymbol{u}_k | \boldsymbol{U}_{m-1} = \boldsymbol{u}_i) \leq U < \sum_{k=1}^{j} P(\boldsymbol{U}_m = \boldsymbol{u}_k | \boldsymbol{U}_{m-1} = \boldsymbol{u}_i)$$

     - $\boldsymbol{U}_m = \boldsymbol{u}_j$

     `until` $m = m_{\text{final}}$.

Markov chains

Transition kernel and invariant measure

- A Markov chain is fully characterized by its initial distribution $\boldsymbol{\pi}_0$ and its *transition kernel* $\boldsymbol{\Pi}(m)$ s.t.:

$$\boldsymbol{\Pi}(m) = [\pi_{ij}(m)]_{i,j\in E}\,,\ \pi_{ij}(m) = P(\boldsymbol{U}_m = \boldsymbol{u}_j | \boldsymbol{U}_{m-1} = \boldsymbol{u}_i)\,.$$

Remarks:

  - $\boldsymbol{\Pi}$ is independent of m if the Markov chain is homogeneous.
  - π_{ii} is not necessarily zero.
  - $\sum_{j\in E} \pi_{ij}(m) = 1,\ \forall i \in E.$

- An *invariant measure* (or *stationary distribution*) $\boldsymbol{\pi}^* = (\pi_j^*)_{j\in E}$ is s.t.:

$$\pi_j^* \geq 0, \quad \pi_j^* = \sum_{i\in E} \pi_i^* \pi_{ij}(m), \quad \forall j \in E\,,\ \forall m \in \mathbb{N}\,,$$

with the normalization $\sum_{j\in E} \pi_j^* = 1$.

Definition

- *A Markov chain is irreducible if it is possible to reach any state from any other state.*
- *The state $j \in E$ is transient if, starting from it, the probability it is never reached again is non-zero; otherwise it is recurrent.*
- *A recurrent state is positive recurrent if the mean return time is finite:* $\mathbb{E}\{\inf\{k \geq 1 | \boldsymbol{U}_k = \boldsymbol{u}_j\} | \boldsymbol{U}_0 = \boldsymbol{u}_j\} < +\infty$.
- *The state $j \in E$ is aperiodic if, starting from it, it can be reached again at any subsequent time with a non-zero probability.*

If $\pi_{ij} > 0 \; \forall i, j \in E$, then the MC is irreducible and aperiodic.

Outline

Markov chains

Existence of an invariant measure for homogeneous MC

Theorem

- *If a MC has at least one positive recurrent state, it has an invariant measure. It is unique if in addition the MC is irreducible.*
- *If a MC is irreducible, positive recurrent and aperiodic, then there exists a unique invariant measure* $\boldsymbol{\pi}^*$ *such that* $(\boldsymbol{U}_m) \xrightarrow[m\to+\infty]{\mathscr{L}} \boldsymbol{\pi}^*$ *independently of* $\boldsymbol{\pi}_0$.

- The latter result is the convergence of the k-stage transition $\boldsymbol{\Pi}^k$ through the kernel $\boldsymbol{\Pi}$ to the invariant measure:

$$\boldsymbol{\Pi}^k = [\pi_{ij}^{(k)}]_{i,j\in E}\,, \quad \pi_{ij}^{(k)} = P(\boldsymbol{U}_{m+k} = \boldsymbol{u}_j | \boldsymbol{U}_m = \boldsymbol{u}_i)\,,$$

  such that if $\boldsymbol{\pi}_k = P(\boldsymbol{U}_k = \boldsymbol{u}_j)_{j\in E}$ then $\boldsymbol{\pi}_k = \boldsymbol{\pi}_0\boldsymbol{\Pi}^k$.
- If $\boldsymbol{\pi}_0$ is an invariant measure, then $\boldsymbol{\pi}_k = \boldsymbol{\pi}_0\ \forall k \in \mathbb{N}$.

Markov chains
The ergodic theorem

Theorem

Let $(\boldsymbol{U}_m, m \in \mathbb{N})$ be an irreducible, positive recurrent homogeneous MC. Then there exists an invariant measure $\boldsymbol{\pi}^*$ such that for all regular functions f:

$$\lim_{m\to+\infty} \mathbb{E}\{f(\boldsymbol{U}_m)\} = \lim_{m\to+\infty} \frac{1}{m} \sum_{k=0}^{m-1} f(\boldsymbol{U}_k) = \sum_{j\in E} \pi_j^* f(\boldsymbol{u}_j) \ a.s.$$

- The ergodic theorem is an extension of the *Law of Large Numbers* to homogeneous MCs: let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_m$ be m i.i.d. r.v. with the same law $P_{\boldsymbol{X}}$, then for all regular functions f:

$$\lim_{m\to+\infty} \frac{1}{m} \sum_{k=1}^{m} f(\boldsymbol{X}_k) = \mathbb{E}\{f(\boldsymbol{X})\} = \int f(\boldsymbol{x}) P_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x}) \quad \text{a.s.}$$

- It allows to estimate the invariant measure from one sample chain.

Outline

MCMC methods

Reversible transition kernels

Proposition

Let $\pi(\boldsymbol{u}_i, \boldsymbol{u}_j) = p(\boldsymbol{u}_i, \boldsymbol{u}_j) + r(\boldsymbol{u}_i)\delta(\boldsymbol{u}_j - \boldsymbol{u}_i)$ *where:*

- $p(\boldsymbol{u}_i, \boldsymbol{u}_i) = 0$ *and a reversibility condition holds:*

$$\boldsymbol{\pi}^*(\boldsymbol{u}_i)p(\boldsymbol{u}_i, \boldsymbol{u}_j) = \boldsymbol{\pi}^*(\boldsymbol{u}_j)p(\boldsymbol{u}_j, \boldsymbol{u}_i)\,, \quad \forall i, j \in E\,;$$

- $r(\boldsymbol{u}_i) = 1 - \sum_{j \in E} p(\boldsymbol{u}_i, \boldsymbol{u}_j)$.

Then $\pi^*(\boldsymbol{u})$ *is the invariant density of the MC of which transition kernel is* $\boldsymbol{\Pi} = [\pi(\boldsymbol{u}_i, \boldsymbol{u}_j)]_{i,j \in E}$.

$$\begin{aligned}
\sum_{i \in E} \pi^*(\boldsymbol{u}_i)\pi(\boldsymbol{u}_i, \boldsymbol{u}_j) &= \sum_{i \in E} \pi^*(\boldsymbol{u}_i)p(\boldsymbol{u}_i, \boldsymbol{u}_j) + \sum_{i \in E} \pi^*(\boldsymbol{u}_i)r(\boldsymbol{u}_i)\delta(\boldsymbol{u}_j - \boldsymbol{u}_i) \\
&= \sum_{i \in E} \pi^*(\boldsymbol{u}_j)p(\boldsymbol{u}_j, \boldsymbol{u}_i) + \pi^*(\boldsymbol{u}_j)r(\boldsymbol{u}_j) \\
&= \pi^*(\boldsymbol{u}_j)(1 - r(\boldsymbol{u}_j)) + \pi^*(\boldsymbol{u}_j)r(\boldsymbol{u}_j) \\
&= \pi^*(\boldsymbol{u}_j)\,.
\end{aligned}$$

MCMC methods

Metropolis-Hastings algorithm (1953, 1970)

- Objective: simulate $\boldsymbol{X} \sim C^* \pi^*(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$, $E \equiv \Omega \subseteq \mathbb{R}^n$.
- Let $q(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ be a *candidate* (or *instrumental* or *proposal*) density s.t. $\operatorname{supp} \pi^*(\cdot) \subset \operatorname{supp} q(\boldsymbol{x}, \cdot)$ and define the *probability of move*:

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

1. $\boldsymbol{x}_0 \sim \boldsymbol{\pi}_0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$
2. `repeat`
   - $\boldsymbol{X}_m = \boldsymbol{x}_m$
   - $\boldsymbol{Y} \sim q(\boldsymbol{x}_m, \boldsymbol{y}) \mathrm{d}\boldsymbol{y}$ and $U \sim \mathcal{U}(0, 1)$
   - `if` $U \leq \alpha(\boldsymbol{x}_m, \boldsymbol{Y})$ `then` $\boldsymbol{X}_{m+1} = \boldsymbol{Y}$, `else` $\boldsymbol{X}_{m+1} = \boldsymbol{X}_m$

   `until` $m = m_{\text{final}}$.

Assume $\pi^*(\boldsymbol{x}) q(\boldsymbol{x}, \boldsymbol{y}) > \pi^*(\boldsymbol{y}) q(\boldsymbol{y}, \boldsymbol{x})$, then $\alpha(\boldsymbol{x}, \boldsymbol{y})$ is tuned so that $\alpha(\boldsymbol{x}, \boldsymbol{y}) q(\boldsymbol{x}, \boldsymbol{y})$ satisfies the reversibility condition.

MCMC methods
Application to the MaxEnt

- The MaxEnt principle leads to the non-linear optimization problem of finding $\boldsymbol{\lambda} \in \mathbb{R}^\ell$ s.t.:

$$\boldsymbol{g}(\boldsymbol{\lambda}) := \int_{\mathbb{R}^n} (\boldsymbol{C}(\boldsymbol{x}) - \underline{\boldsymbol{C}})\, \mathrm{e}^{-\boldsymbol{\lambda}\cdot(\boldsymbol{C}(\boldsymbol{x})-\underline{\boldsymbol{C}})}\, \mathrm{d}\boldsymbol{x} = \boldsymbol{0}\,.$$

- By the ergodic theorem:

$$\boldsymbol{g}(\boldsymbol{\lambda}) \simeq \lim_{m\to+\infty} \frac{1}{m} \sum_{k=0}^{m-1} (\boldsymbol{C}(\boldsymbol{X}_k(\theta)) - \underline{\boldsymbol{C}})\ ,$$

and by the law of large number for $M \in \mathbb{N}$ large enough:

$$\boldsymbol{g}(\boldsymbol{\lambda}) \simeq \lim_{n_s\to+\infty} \frac{1}{n_s} \sum_{k=1}^{n_s} (\boldsymbol{C}(\boldsymbol{X}_M(\theta_k)) - \underline{\boldsymbol{C}})\ ,$$

where $(\boldsymbol{X}_m,\, m \in \mathbb{N})$ is the MC of which invariant density is $\pi^*(\boldsymbol{x}) = \mathrm{e}^{-\boldsymbol{\lambda}\cdot(\boldsymbol{C}(\boldsymbol{x})-\underline{\boldsymbol{C}})}$.

When does estimation occur?

- Parametrization of *e.g.* Gamma or β distributions (Lamé's moduli, Poisson's coefficient):
  - If $X \sim \Gamma(k, \frac{1}{\sigma})$ then

$$\underline{X} = k\sigma \,,\; \sigma_X^2 = k\sigma^2 \,,\; s_X = \frac{2}{\sqrt{k}} \,,\; \kappa_X = \frac{6}{k} \,,\; etc.$$

  - If $X \sim \beta(k, \frac{1}{\sigma})$ then

$$\underline{X} = \frac{k\sigma}{1+k\sigma} \,,\; \sigma_X^2 = \frac{k\sigma^2}{(1+k\sigma)^2(1+\sigma+k\sigma)} \,,\; etc.$$

- $p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathrm{e}^{-\lambda_0 - \boldsymbol{\lambda}\cdot\boldsymbol{C}(\boldsymbol{x})}$ where $(\lambda_0, \boldsymbol{\lambda})$ depend on $\underline{\boldsymbol{C}}$.

Estimator

Punctual & sequential estimation

- Let $\boldsymbol{X}(\theta)$ be a second order r.v. defined on $(\Omega_\theta, \mathcal{E}, P)$ with values in $\Omega_{\boldsymbol{X}} \subseteq \mathbb{R}^n$ and probability distribution $P_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x}; \boldsymbol{\nu})$ depending on (deterministic) parameters $\boldsymbol{\nu} \in \Theta \subseteq \mathbb{R}^p$.
  - Sequential or continuous estimation: estimate $\boldsymbol{\nu}$ from M independent realizations $\boldsymbol{x}^{(m)} = \boldsymbol{X}(\theta_m)$, $\theta_m \in \Omega_\theta$, $1 \leq m \leq M$;
  - Punctual estimation: $M = 1$.
- An estimation is a (measurable) rule $\boldsymbol{x}_M \mapsto \hat{\boldsymbol{\nu}}_M(\boldsymbol{x}_M)$, $\boldsymbol{x}_M = (\boldsymbol{x}^{(1)}, \ldots \boldsymbol{x}^{(M)})$, for calculating the *estimate* $\hat{\boldsymbol{\nu}}_M(\boldsymbol{x}_M)$. The corresponding *estimator* is the r.v. $\hat{\boldsymbol{\nu}}_M(\boldsymbol{X}_M)$.
- The *estimation error* is $\boldsymbol{\epsilon}_M(\boldsymbol{X}_M) = \hat{\boldsymbol{\nu}}_M(\boldsymbol{X}_M) - \boldsymbol{\nu}$.
- Two classes of methods for building estimators: (i) method of moments and (ii) maximum likelihood.

Estimator

Characterizing an estimator

Definition

- *The bias of the estimator* $\boldsymbol{b}_M(\boldsymbol{\nu}) = \mathbb{E}\{\boldsymbol{\epsilon}_M(\boldsymbol{X}_M)\}$. *It is unbiased if* $\boldsymbol{b}_M = \mathbf{0}\ \forall M$, *or asymptotically unbiased if* $\lim_{M\to+\infty} \boldsymbol{b}_M = \mathbf{0}$.
- *The dispersion of the estimator:*

$$\begin{aligned}\boldsymbol{V}_M(\boldsymbol{\nu}) &= \mathbb{E}\{\boldsymbol{\epsilon}_M(\boldsymbol{X}_M) \otimes \boldsymbol{\epsilon}_M(\boldsymbol{X}_M)\} \\ &= \mathbf{C}_{\hat{\boldsymbol{\nu}}_M} + \boldsymbol{b}_M(\boldsymbol{\nu}) \otimes \boldsymbol{b}_M(\boldsymbol{\nu})\,.\end{aligned}$$

  The lower $\operatorname{Tr} \mathbf{C}_{\hat{\boldsymbol{\nu}}_M}$ *is, the more efficient it is.*
- *The symmetric, positive Fisher information matrix:*

$$[\mathcal{I}_M(\boldsymbol{\nu})] = \mathbb{E}\{\boldsymbol{\nabla}_{\boldsymbol{\nu}} \ln \mathcal{L}(\boldsymbol{\nu}|\boldsymbol{X}_M) \otimes \boldsymbol{\nabla}_{\boldsymbol{\nu}} \ln \mathcal{L}(\boldsymbol{\nu}|\boldsymbol{X}_M)\}\,,$$

for $\mathcal{L}(\boldsymbol{\nu}|\boldsymbol{x}_M) := \prod_{m=1}^{M} p_{\boldsymbol{X}}(\boldsymbol{x}^{(m)};\boldsymbol{\nu})$, $p_{\boldsymbol{X}} = \frac{\mathrm{d}P_{\boldsymbol{X}}}{\mathrm{d}\boldsymbol{x}}$.

Cramér-Rao inequality and efficiency

Theorem

The Cramér-Rao inequality for the j-th parameter:

$$[\boldsymbol{V}_M(\boldsymbol{\nu})]_{jj} = \mathbb{E}\{\epsilon^2_{M,j}(\boldsymbol{X}_M)\} \geq \frac{\left(1+\partial_{\nu_j} b_{M,j}(\boldsymbol{\nu})\right)^2}{[\mathcal{I}_M(\boldsymbol{\nu})]_{jj}}\,,$$

with an equality iff $\partial_{\nu_j} \ln \mathcal{L}(\boldsymbol{\nu}|\boldsymbol{x}_M) = \varphi_j(\boldsymbol{\nu})\epsilon_{M,j}(\boldsymbol{x}_M)$.

- An estimator is *efficient* if it minimizes its dispersion, thus iff the above condition holds, or:

$$\mathbb{E}\{\partial_{\nu_j} \ln \mathcal{L}(\boldsymbol{\nu}|\boldsymbol{X}_M)\} = \varphi_j(\boldsymbol{\nu}) b_{M,j}(\boldsymbol{\nu})\,.$$

- If it is unbiased $[\boldsymbol{V}_M(\boldsymbol{\nu})]_{jj} \geq [\mathcal{I}_M(\boldsymbol{\nu})]^{-1}_{jj}$ (Cramér-Rao lower bound) and

$$\mathbb{E}\{\partial_{\nu_j} \ln \mathcal{L}(\boldsymbol{\nu}|\boldsymbol{X}_M)\} = 0\,.$$

Maximum likelihood method

ML estimator

Definition

- *Consider the following estimate* $\hat{\ell}(\boldsymbol{\nu})$ *of the expected log-likelihood* $\ell(\boldsymbol{\nu}) = \mathbb{E}\{\ln p_{\boldsymbol{X}}(\boldsymbol{X};\boldsymbol{\nu})\}$ *of a single observation:*

$$\hat{\ell}(\boldsymbol{\nu}) = \frac{1}{M} \ln \mathcal{L}(\boldsymbol{\nu}|\boldsymbol{x}_M)\,.$$

- *The maximum likelihood (ML) estimator of* $\boldsymbol{\nu}$ *is:*

$$\hat{\boldsymbol{\nu}}_{\mathrm{ML}}(\boldsymbol{X}_M) = \arg\max_{\boldsymbol{\nu}\in\Theta} \hat{\ell}(\boldsymbol{\nu})\,.$$

The likelihood equations relative to the realizations $\boldsymbol{x}_M$:

$$\partial_{\nu_j} \ln \mathcal{L}(\boldsymbol{\nu}|\boldsymbol{x}_M) = \sum_{m=1}^{M} \partial_{\nu_j} \ln p_{\boldsymbol{X}}(\boldsymbol{x}^{(m)};\boldsymbol{\nu}) = 0\,, \quad 1 \le j \le p\,.$$

Maximum likelihood method

Properties of the MLE

- Consistency: assume $p_{\boldsymbol{X}}(\cdot;\boldsymbol{\nu}_1) \neq p_{\boldsymbol{X}}(\cdot;\boldsymbol{\nu}_2)$ iff $\boldsymbol{\nu}_1 \neq \boldsymbol{\nu}_2$, the MLE is asymptotically unbiased and:

$$\hat{\boldsymbol{\nu}}_{\mathrm{ML}}(\boldsymbol{X}_M) \underset{M\to+\infty}{\xrightarrow{\mathscr{P}}} \boldsymbol{\nu}\,.$$

- Asymptotic normality:

$$\sqrt{M}\left(\hat{\boldsymbol{\nu}}_{\mathrm{ML}}(\boldsymbol{X}_M) - \boldsymbol{\nu}\right) \underset{M\to+\infty}{\xrightarrow{\mathscr{L}}} \mathcal{N}\left(0, [\mathcal{I}_1(\boldsymbol{\nu})]^{-1}\right)\,.$$

- Efficiency: it achieves the Cramér-Rao lower bound asymptotically. Conversely, if an efficient unbiased estimator exists, then it is the MLE and it is unique.
- Functional invariance: $\varphi(\hat{\boldsymbol{\nu}}_{\mathrm{ML}})$ is the MLE of $\varphi(\boldsymbol{\nu})$.

Maximum likelihood method

Example #1

Consider $X \sim \mathcal{N}(\mu, \sigma^2)$ in $\mathbb{R}$, *i.e.* $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$, with $\nu_1 = \mu$ and $\nu_2 = \sigma^2$.

- The likelihood function is:

$$\mathcal{L}(\boldsymbol{\nu}|x_M) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{M}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{m=1}^{M}\left(x^{(m)} - \mu\right)^2\right) .$$

- The ML estimator is:

$$\hat{\mu}_{\mathrm{ML}}(X_M) = \frac{1}{M}\sum_{m=1}^{M} X^{(m)} ,$$

$$\hat{\sigma}^2_{\mathrm{ML}}(X_M) = \frac{1}{M}\sum_{m=1}^{M}\left(X^{(m)} - \frac{1}{M}\sum_{n=1}^{M} X^{(n)}\right)^2 .$$

- Then $b_{M,1} = 0$, $b_{M,2} = -\frac{\sigma^2}{M}$, but is asymptotically unbiased.

Maximum likelihood method

Example #2

Consider $X \sim \Gamma(k, \frac{1}{\sigma})$ on $\mathbb{R}_+$ with $\nu = \sigma$:

$$p_X(x) = \frac{1}{\sigma\Gamma(k)} \left(\frac{x}{\sigma}\right)^{k-1} \mathrm{e}^{-\frac{x}{\sigma}} \mathbf{1}_{\mathbb{R}_+}(x)\,,$$

then $\underline{X} = k\sigma = k\nu$.

- The log-likelihood function is:

$$\frac{\ln \mathcal{L}(\nu|x_M)}{M} = \frac{k-1}{M} \sum_{m=1}^{M} \ln x^{(m)} - \frac{1}{\sigma M} \sum_{m=1}^{M} x^{(m)} - k \ln \sigma - \ln \Gamma(k)\,.$$

- The (unbiased) ML estimator is:

$$k\hat{\sigma}_{\mathrm{ML}}(X_M) = \frac{1}{M} \sum_{m=1}^{M} X^{(m)}\,.$$

Maximum likelihood method
Example #2

Consider the same law but now $\nu_1 = \sigma$ and $\nu_2 = k$, thus $\underline{X} = k\sigma = \nu_1\nu_2$ and $\sigma_X^2 = k\sigma^2 = \nu_1^2\nu_2$.

- The log-likelihood function is unchanged.
- But $\hat{\sigma}_{\mathrm{ML}}$ and $\hat{k}_{\mathrm{ML}}$ are given by a system of *non-linear* equations:

$$\hat{k}_{\mathrm{ML}}(X_M)\hat{\sigma}_{\mathrm{ML}}(X_M) = \frac{1}{M}\sum_{m=1}^{M} X^{(m)},$$

$$\psi_0(\hat{k}_{\mathrm{ML}}(X_M)) = \frac{1}{M}\sum_{m=1}^{M} \ln\left(\frac{X^{(m)}}{\hat{\sigma}_{\mathrm{ML}}(X_M)}\right),$$

where $z \mapsto \psi_0(z) = (\ln\Gamma(z))'$ is the diGamma function.

Exhaustive estimator

Definition

- The probability law of the estimator $\hat{\boldsymbol{\nu}}_M(\boldsymbol{X}_M)$ should depend on $\boldsymbol{\nu}$ if it is used to retrieve some information on the parameters.
- It is called *exhaustive* if it conserves the information on $\boldsymbol{\nu}$ contained in the sample $\boldsymbol{x}_M$, *i.e.* $P(\boldsymbol{X}_M|\hat{\boldsymbol{\nu}}_M(\boldsymbol{X}_M) = \hat{\boldsymbol{\nu}})$ does not depend on $\boldsymbol{\nu}$.

Definition

$\hat{\boldsymbol{\nu}}_M(\boldsymbol{X}_M)$ is exhaustive iff $\mathcal{L}(\boldsymbol{\nu}|\boldsymbol{x}_M) = \rho_t(\boldsymbol{x}_M|\hat{\boldsymbol{\nu}})g(\hat{\boldsymbol{\nu}};\boldsymbol{\nu})$, where ρ_t is the conditional PDF of $\boldsymbol{X}_M$ provided that $\hat{\boldsymbol{\nu}}_M(\boldsymbol{X}_M) = \hat{\boldsymbol{\nu}}$, and g is its PDF.

Exhaustive estimator

Characterization

Theorem

A sample $\boldsymbol{X}_M$ *s.t.* $\operatorname{supp} p_{\boldsymbol{X}}$ *does not depend on* $\boldsymbol{\nu}$ *admits an exhaustive estimator iff:*

$$p_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\nu}) = \mathrm{e}^{a(\boldsymbol{x})\alpha(\boldsymbol{\nu})+b(\boldsymbol{x})+\beta(\boldsymbol{\nu})} \ .$$

Further reading...

- P. Brémaud: *Markov Chains: Gibbs Fields, Monte-Carlo Simulation, and Queues*, Springer (2008);
- L. Devroye: *Non Uniform Random Variate Generation*, Springer-Verlag (1986); Available at http://luc.devroye.org/rnbookindex.html;
- O. Häggström: *Finite Markov Chains and Algorithmic Applications*, London Mathematical Society Student Texts (2002);
- P.K. McKeown: *Stochastic Simulation in Physics*, Springer (1997);
- S.P. Meyn & R.L. Tweedie: *Markov Chains and Stochastic Stability*, Springer-Verlag (1993); Re-compiled version as of 2005 available at http://probability.ca/MT.