Markov chain Monte Carlo

We assume that the random variables (X, Y) have an unknown distribution and we usually propose a parametric family of joint probability distributions $(x, y) \mapsto p_\theta(x, y)$, $\theta \in \Theta$, with respect to a dominating measure λ, in order to estimate this unknown distribution.

In machine/statistical learning most of the problems we aim to solve require then to sample random variables from a complex distribution. Depending on the context, this distribution can be given for instance by a marginal density $p_\theta(y) = \int p_\theta(x, y)\lambda(\mathrm{d}x)$ or by a conditional/posterior distribution $p_\theta(x|y) = p_\theta(x)p_\theta(y|x)/\int p_\theta(x, y)\lambda(\mathrm{d}x)$.

Example: Bayesian inference. In Bayesian inference, prior belief is combined with data to obtain posterior distributions on which statistical inference is based. Except for some simple cases, Bayesian inference can be computationally intensive and may rely on computational techniques.

The basic idea in Bayesian analysis is that a parameter vector $\theta \in \Theta$ is unknown, so it is endowed with a *prior* distribution with probability density $\pi_0(\theta)$. We also introduce a model or likelihood, $\mathrm{p}(y_{1:n} \mid \theta)$, that is a probability density function for the data $Y = y_{1:n}$ which depends on the parameter vector. Inference about θ is then based on the *posterior* distribution, which is obtained via Bayes's theorem,

$$\theta \mapsto \pi(\theta) = \frac{\pi_0(\theta)\,\mathrm{p}(y_{1:n} \mid \theta)}{\int \pi_0(\theta)\,\mathrm{p}(y_{1:n} \mid \theta)\lambda(\mathrm{d}\theta)} \propto \pi_0(\theta)\,\mathrm{p}(y_{1:n} \mid \theta).$$

In some simple cases, the prior and the likelihood are *conjugate* distributions that may be combined easily. For example, in n fixed repeated (i.i.d.) Bernoulli experiments with probability of success θ, a *Beta-Binomial* conjugate pair is taken. In this case the prior is $\mathrm{Beta}(a, b)$: $\pi_0(\theta) \propto \theta^a(1-\theta)^b\mathbb{1}_{(0,1)}(\theta)$; the values $a, b > -1$ are called hyperparameters. The likelihood in this example is $\mathrm{Binomial}(n, \theta)$: $\mathrm{p}(y \mid \theta) \propto \theta^y(1-\theta)^{n-y}$, from which we easily deduce that the posterior is also Beta, $\pi(\theta) \propto \theta^{y+a}(1-\theta)^{n+b-y}\mathbb{1}_{(0,1)}(\theta)$, and from which inference may easily be achieved.

In more complex experiments, the posterior distribution is often difficult to obtain by direct calculation, so alternatives have to be deployed to obtain samples approximately distributed as the posterior distribution.

Definitions

Notations. Let $(\mathsf{X}, \mathcal{X})$ be a measurable space, i.e. $\mathcal{X}$ is a σ-algebra on X, and consider the following notations.

- $\mathsf{M}_+(\mathsf{X})$ is the set of non-negative measures on $(\mathsf{X}, \mathcal{X})$.
- $\mathsf{M}_1(\mathsf{X})$ is the set of probability measures on $(\mathsf{X}, \mathcal{X})$.
- $\mathsf{F}(\mathsf{X})$ is the set of real-valued measurable functions f on X and $\mathsf{F}_+(\mathsf{X})$ the set of non-negative measurable functions on X.
- If $k \leq \ell$, $u_{k:\ell}$ means $(u_k, \ldots, u_\ell)$ and $u_{k:\infty}$ means $(u_{k+\ell})_{\ell\in\mathbb{N}}$.

Markov kernel. We say that $P : \mathsf{X} \times \mathcal{X} \to \mathbb{R}^+$ is a Markov kernel, if for all $(x, A) \in \mathsf{X} \times \mathcal{X}$,

- $\mathsf{X} \ni y \mapsto P(y, A)$ is $\mathcal{X}/\mathcal{B}(\mathbb{R}^+)$ measurable,
- $\mathcal{X} \ni B \mapsto P(x, B)$ is a probability measure on $(\mathsf{X}, \mathcal{X})$.

For all $(x, A) \in \mathsf{X} \times \mathcal{X}$, as a function of the first component only, $P(\cdot, A)$ is measurable and as a function of the second component only, $P(x, \cdot)$ is a probability measure. In particular, $P(x, \mathsf{X}) = 1$ for all $x \in \mathsf{X}$. Since $P(x, \cdot)$ is a measure, we also use the infinitesimal notation: $P(x, \mathrm{d}y)$. For example,

$$P(x, A) = \int_{\mathsf{X}} \mathbf{1}_A(y) P(x, \mathrm{d}y) = \int_A P(x, \mathrm{d}y) \,.$$

For all $\mu \in \mathsf{M}_+(\mathsf{X})$, all Markov kernels P, Q on $\mathsf{X} \times \mathcal{X}$, and all measurable non-negative or bounded functions h on X, we use the following convention and notation.

- μP is the (positive) measure: $\mathcal{X} \ni A \mapsto \mu P(A) = \int \mu(\mathrm{d}x) P(x, A)$,
- PQ is the Markov kernel: $(x, A) \mapsto \int_{\mathsf{X}} P(x, \mathrm{d}y) Q(y, A)$,
- Ph is the measurable function $x \mapsto \int_{\mathsf{X}} P(x, \mathrm{d}y) h(y)$.

It is easy to check that if μ is a probability measure, then μP is also a probability measure (since $\mu P(\mathsf{X}) = \int_{\mathsf{X}} \mu(\mathrm{d}x) P(x, \mathsf{X}) = \int_{\mathsf{X}} \mu(\mathrm{d}x) = 1$). With this notation, using Fubini's theorem,

$$\begin{aligned} \mu(P(Qh)) &= (\mu P)(Qh) = (\mu(PQ))h \\ &= \mu((PQ)h) = \int_{\mathsf{X}^3} \mu(\mathrm{d}x) P(x, \mathrm{d}y) Q(y, \mathrm{d}z) h(z). \end{aligned}$$

To finish up with notation, we now define the iterates of a Markov kernel P, which will come in very handy thereafter: for a given Markov kernel P on $\mathsf{X} \times \mathcal{X}$, define $P^0 = I$ where I is the identity kernel: $(x, A) \mapsto \mathbf{1}_A(x)$, and set for $k \geq 0$, $P^{k+1} = P^k P$.

Markov chain. Let$\{X_k : k \in \mathbb{N}\}$ be a sequence of random variables on the same probability space $(\Omega, \mathcal{G}, \mathbb{P})$ and taking values on X, we say that $\{X_k : k \in \mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution $\nu \in \mathsf{M}_1(\mathsf{X})$ if and only if

1. for all $(k, A) \in \mathbb{N} \times \mathcal{X}$, $\mathbb{P}(X_{k+1} \in A | X_{0:k}) = P(X_k, A)$, $\mathbb{P}$-a.s.
2. $\mathbb{P}(X_0 \in A) = \nu(A)$.

Note that in the definition we consider $\mathbb{P}(X_{k+1} \in A | X_{0:k})$, that is, the conditional probability is with respect to the sigma-field $\sigma(X_{0:k})$. We can actually replace $\sigma(X_{0:k})$ by $\mathcal{F}_k$ as soon as we know that $(X_k)_{k \geq 0}$ is $(\mathcal{F}_k)_{k \geq 0}$-adapted.

Invariant probability measure. We say that $\pi \in \mathsf{M}_1(\mathsf{X})$ is an invariant probability measure for the Markov kernel P on $\mathsf{X} \times \mathcal{X}$ if $\pi P = \pi$.

If (X_k) is a Markov chain with Markov kernel P and assuming that $X_0 \sim \pi$, then for all $k \geq 1$, we have $X_k \sim \pi$ since applying P^k on both sides of $\pi P = \pi$ shows that $\pi P^{k+1} = \pi P^k$ and therefore, for all $k \in \mathbb{N}$, $\pi P^k = \pi$.

Metropolis-Hastings algorithm

In this section, we are given a probability measure $\pi \in \mathsf{M}_1(\mathsf{X})$ and the idea now is to construct a Markov chain $\{X_k : k \in \mathbb{N}\}$ admitting π as invariant probability measure, in which case we say that π is a target distribution. In other words, we try to find a Markov kernel P on $\mathsf{X} \times \mathcal{X}$ such that P is π-invariant.

For simplicity we now assume that π has a density with respect to some dominating σ-finite measure λ and by abuse of notation, we also denote by π this density, and we assume that this density π is **positive**.

Moreover, let Q be Markov kernel on $\mathsf{X} \times \mathcal{X}$ such that $Q(x, \mathrm{d}y) = q(x, y)\lambda(\mathrm{d}y)$, that is, for any $x \in \mathsf{X}$, $Q(x, \cdot)$ is also dominated by λ and denoting by $q(x, \cdot)$ this density, we assume for simplicity that $q(x, y)$ is **positive** for all $x, y \in \mathsf{X}$.

Define

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

Then, with this choice of acceptance rate, the Markov chain produced by the Metropolis-Hastings algorithm is π-reversible.

```
output: $X_0, \ldots, X_n$
At $k = 0$, draw $X_0$ according to some arbitrary distribution
for $k \leftarrow 0$ to $n - 1$ do
    • Draw independently $Y_{k+1} \sim Q(X_k, \cdot)$ and $U_{k+1} \sim \mathrm{Unif}(0, 1)$
    • Set $X_{k+1} = \begin{cases} Y_{k+1} & \text{if } U_{k+1} \leq \alpha(X_k, Y_{k+1}) \,, \\ X_k & \text{otherwise.} \end{cases}$
end
```

Algorithm 1: The Metropolis-Hastings Algorithm

In words, Q allows to propose a candidate for the next value of the Markov chain $(X_k)_{k\in\mathbb{N}}$ and this candidate is accepted or refused according to a probability given by the function α.

The random walk MH sampler. If $\mathsf{X} = \mathbb{R}^p$ and if the proposal kernel is $Q(x, \mathrm{d}y) = q(y - x)\lambda(\mathrm{d}y)$ where q is a symmetric density with respect to λ on X, (by symmetric, we mean that $q(u) = q(-u)$ for all $u \in \mathsf{X}$) then at each time step in the MH algorithm, we draw a candidate $Y_{k+1} \sim q\,(y\ - X_k)\,\lambda(\mathrm{d}y)$. In such a case, the acceptance probability is $\alpha(x, y) = \min\,(\pi(y)/\pi(x), 1)$ and the associated algorithm is called the *(symmetric) Random Walk Metropolis-Hasting.* Another way of writing the proposal update is $Y_{k+1} = X_k + \eta_k$ where $\eta_k \sim q(\cdot)$.

Gibbs sampler

The Gibbs sampler is a specific version of Metropolis-Hastings algorithm in cases where the state $\mathsf{X} = \mathbb{R}^d$ and where for all $x \in \mathbb{R}^d$, x can be decomposed into $x = (x^{(1)}, \ldots, x^{(q)})$ so that for all $1 \leq j \leq d$, **we know how to sample from** $\pi(\cdot|x^{(-j)})$ **where** $x^{(-j)} = (x^{(\ell)})_{\ell\neq j}$. It is easy to write the proposal distribution associated with the Gibbs sampler and to show that the acceptance rate is 1.

```
input  : $X_k = (X_k^{(1)}, \ldots, X_k^{(q)})$
output: $X_{k+1}$
Sample uniformly $J$ in $\{1, \ldots, q\}$.
Sample $X_{k+1}^{(J)} \sim \pi(\cdot|X_k^{(-J)})$.
For all $1 \leq j \leq q$, $j \neq J$, set $X_{k+1}^{(-j)} = X_k^{(-j)}$.
```

Algorithm 2: One iteration of the Gibbs sampler