Markov Chain Monte Carlo

Mladen Victor WICKERHAUSER

Washington University in St. Louis, Missouri
victor@wustl.edu
http://www.math.wustl.edu/~victor

Dimensionality Reduction and Manifold Estimation
PMF — University of Zagreb
Winter, 2022

Goals

Find a global maximum for $f : \mathbf{R}^d \to \mathbf{R}$.

- Expect d to be large.
- Only evaluate ratios $f(\mathbf{x})/f(\mathbf{y})$.
- Assume f is regular (continuous, maybe smooth).

Method: Markov chain Monte Carlo (MCMC).

Probability Spaces

These are triples $(\Omega, \mathcal{F}, \Pr)$, with

- Set Ω, called the probability space,
- Measurable subsets $A \subset \Omega$ forming a σ-algebra $\mathcal{F}$ of *events*, satisfying
  - $\Omega, \emptyset \in \mathcal{F}$, and $A \in \mathcal{F} \implies \Omega \setminus A \in \mathcal{F}$,
  - $\{A_i : i \in \mathbf{N}\} \subset \mathcal{F} \implies \cup_i A_i \in \mathcal{F}$ and $\cap_i A_i \in \mathcal{F}$,
- Probability function $\Pr : \mathcal{F} \to \mathbf{R}$ satisfying
  - $(\forall A \in \mathcal{F})\, 0 \leq \Pr(A) \leq 1$, with $\Pr(\Omega) = 1$ and $\Pr(\emptyset) = 0$,
  - $\Pr(\Omega \setminus A) = 1 - \Pr(A)$,
  - If $A, B \in \mathcal{F}$ with $A \subset B$, then $\Pr(A) \leq \Pr(B)$.
  - If $\{A_i : i \in \mathbf{N}\} \subset \mathcal{F}$ is a countable collection of disjoint measurable sets, then $\Pr(\cup_i A_i) = \sum_i \Pr(A_i)$.

Bayes' Rule for Events

Conditional probability: for $A, B \in \mathcal{F}$ with $\mathrm{Pr}(B) \neq 0$,

$$\mathrm{Pr}(A|B) \stackrel{\text{def}}{=} \frac{\mathrm{Pr}(A \cap B)}{\mathrm{Pr}(B)}.$$

Bayes' Rule: for $A, B \in \mathcal{F}$ with $\mathrm{Pr}(A) \neq 0$ and $\mathrm{Pr}(B) \neq 0$,

$$\mathrm{Pr}(A|B)\mathrm{Pr}(B) = \mathrm{Pr}(B|A)\mathrm{Pr}(A).$$

The proof is obvious from the definition, since

$$\frac{\mathrm{Pr}(A \cap B)}{\mathrm{Pr}(B)}\mathrm{Pr}(B) = \mathrm{Pr}(A \cap B) = \frac{\mathrm{Pr}(B \cap A)}{\mathrm{Pr}(A)}\mathrm{Pr}(A).$$

The nonvanishing of $\mathrm{Pr}(A)$ and $\mathrm{Pr}(B)$ is only needed to define the conditional probabilities.

Interpretation

Consider

- $E \in \mathcal{F}$ is an experiment
- $H \in \mathcal{F}$ is a hypothesis

Then

- $\Pr(H|E)$ is a test of H (by p-value!)
- $\Pr(E|H)$ is a model predicting E from H

Use Bayes' rule to test the hypothesis from the result:

$$\Pr(H|E) = \Pr(E|H)\Pr(H)/\Pr(E) \propto \Pr(E|H)\Pr(H).$$

This requires a *prior* probability $\Pr(H)$

After the experiment, $\Pr(H|E)$ is the *posterior* probability of H.

Random Variables

This is a function $X : \Omega \to \mathbf{R}$ that is *measurable*:

$$(\forall t \in \mathbf{R})\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F}$$

Then X has a *cumulative distribution function (cdf)*:

$$F(t) \stackrel{\text{def}}{=} \Pr(\{\omega \in \Omega : X(\omega) \leq t\}), \qquad t \in \mathbf{R},$$

with $0 \leq F(t) \leq 1$, $F(t) \to 0$ as $t \to -\infty$, and $F(t) \to 1$ as $t \to +\infty$

If F is differentiable, then $F'(t)$ is called the *density* of X.

Probability Spaces from Distributions

Canonical choices for $(\Omega, \mathcal{F}, \Pr)$, given an r.v. X:

- $\Omega = \mathbf{R}$
- $\mathcal{F}$ is the smallest σ-algebra that contains $\emptyset$, Ω, and $\{\omega \in \mathbf{R} : X(\omega) \leq t\}$ for all $t \in \mathbf{R}$. (This is sometimes denoted by $\mathcal{F}_X$, the σ-algebra "generated" by the r.v. X.)
- $\Pr$ is defined on intervals by $\Pr((a, b]) \stackrel{\text{def}}{=} F(b) - F(a)$, then extended to $\mathcal{F}$ by countable additivity.

Remark. $\mathcal{F} \subset \mathcal{B}$, the Borel subsets, which form the smallest σ-algebra that contains all open subsets of $\mathbf{R}$. Also:

Lemma

Every open set in $\mathbf{R}$ *is a countable disjoint union of open intervals.* $\square$

Random Vectors

Random vectors, for $k > 1$, are $\mathbf{R}^k$-valued random variables:

- Write $X = (X_1, \ldots, X_k)$, where X_i is a random variable.
- Joint distribution (jcdf) is

$$F(t_1, \ldots, t_k) \stackrel{\text{def}}{=} \Pr(\{X_1 \le t_1\} \cap \cdots \cap \{X_k \le t_k\})$$

- Joint density (jpdf), if it exists, is a function $f(t_1, \ldots, t_k)$ satisfying

$$\Pr(X_1 \in A_1 \wedge \cdots \wedge X_k \in A_k) = \int_{A_1} \cdots \int_{A_k} f(t_1, \ldots, t_k)\, dt_1 \cdots dt_k$$

(For $i = 1, \ldots, k$, $A_i \subset \mathbf{R}$ is a measurable subset.)

Remark. R.v. is either "random vector" or "random variable."

Independence

Say that the coordinates of random vector X are *independent* iff, for all values of $t_1, \ldots, t_k$,

$$F(t_1, \ldots, t_k) = F_1(t_1) \cdots F_k(t_k)$$

for some cumulative distribution functions $F_1, \ldots, F_k$.

Equivalently, if there is a density, say that the coordinates of X are independent iff

$$f(t_1, \ldots, t_k) = f_1(t_1) \cdots f_k(t_k)$$

for some functions $f_1, \ldots, f_k$.

Marginal Densities

For joint density $f = f(\mathbf{x}, \mathbf{y})$ obtain *marginal densities* by partial integration:

$$f_X(\mathbf{x}) \stackrel{\text{def}}{=} \int_Y f(\mathbf{x}, \mathbf{y})\, d\mathbf{y}.$$

Here Y denotes all values of $\mathbf{y}$.

Theorem

Random vectors X, Y with joint pdf f are independent iff

$$f(\mathbf{x}, \mathbf{y}) = f_X(\mathbf{x}) f_Y(\mathbf{y}), \qquad a.e. (\mathbf{x}, \mathbf{y}),$$

where f_X, f_Y are the X and Y marginals, respectively. $\square$

Bayes' Rule for Densities

Conditional density:

$$f_{X|Y}(\mathbf{x}|\mathbf{y}) \stackrel{\text{def}}{=} \frac{f(\mathbf{x},\mathbf{y})}{f_Y(\mathbf{y})},$$

where f is the joint density for the r.v. (X, Y), and f_Y is the Y-marginal density

$$f_Y(\mathbf{y}) = \int_X f(\mathbf{x},\mathbf{y})\,d\mathbf{x}.$$

Then, with corresponding definitions for f_Y and $f_{Y|X}$,

$$f_{X|Y}(\mathbf{x}|\mathbf{y})f_Y(\mathbf{y}) = f(\mathbf{x},\mathbf{y}) = f_{Y|X}(\mathbf{y}|\mathbf{x})f_X(\mathbf{x}).$$

Interpretation

Consider two random vectors X, Y:

- X is a vector of fixed, but unobservable parameters;
- Y is a vector of observable variables.

Then

- f_X is a prior density describing the uncertainty in X,
- $f_{Y|X}$ is the density model for Y at each value of X,
- $f_{X|Y}$ is a posterior density describing the uncertainty in X after an experiment produces a particular result Y.

Bayes' rule: $f_{X|Y} \propto f_{Y|X} f_X$ gives the posterior density for X.

Problem: need to use pdfs without normalization.

Techniques and Examples

- Interchange parameters and variables in joint densities.
  - Conjugate densities
  - Simple algebraic relationships
- Use densities with means, modes, variances, etc., that can be determined from parameters without normalization.
  - Special functions
  - Not suitable for empirical densities.
- Use mode rather than expectation
  - Search with ratios of the posterior
  - Seek global maximum likelihood probabilistically

Multinomial Random Vectors

Fix $k \in \mathbf{Z}^+$, fix $p_1, \ldots, p_k \in [0,1]$ with $\sum_i p_i = 1$, and say that r.v. X is multinomial with parameters n and $\{p_i\}$ iff

- X takes values in k-tuples of nonnegative integers $\mathbf{n} = (n_1, \ldots, n_k)$, with fixed $n = n_1 + \cdots + n_k$.
- The probability of the event $A = \{\omega \in \Omega : X = \mathbf{n}\}$ is

$$\mathrm{Pr}(A) = \binom{n_1 + \cdots + n_k}{n_1, \ldots, n_k} p_1^{n_1} \cdots p_k^{n_k}.$$

- The *multinomial coefficient*

$$\binom{n}{n_1, \ldots, n_k} = \binom{n_1 + \cdots + n_k}{n_1, \ldots, n_k} \stackrel{\text{def}}{=} \frac{(n_1 + \cdots + n_k)!}{n_1! \cdots n_k!}$$

is the number of ways to choose n_i objects in category i, for $i = 1, \ldots, k$, if there are n total choices made.

Dirichlet Random Vectors

Fix $k \in \mathbf{Z}^+$, fix $\alpha_1, \ldots, \alpha_k \in [1, +\infty)$, and say that r.v. X is Dirichlet with parameters $\{\alpha_i\}$ if

- X takes values in k-tuples of nonnegative real numbers $\mathbf{p} = (p_1, \ldots, p_k)$, satisfying $p_1 + \cdots + p_k = 1$.
- The probability density function at $\mathbf{p}$ is

$$f(\mathbf{p}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1},$$

- Function Γ, analytic on $\mathbf{R}^+$ (and also on $\mathbf{C} \setminus \{0\}$), is

$$\Gamma(z) \stackrel{\text{def}}{=} \int_0^\infty t^{z-1} e^{-t}\, dt.$$

  It satisfies $\Gamma(1) = \Gamma(2) = 1$ and $(\forall z > 0)\Gamma(z+1) = z\Gamma(z)$, so that $\Gamma(n) = (n-1)!$ for all integers $n \geq 1$.

Canonical Simplexes

Fix $k \in \mathbf{Z}$ with $k \geq 2$. Put $\mathbf{p} = (p_1, \ldots, p_k)$.

If $\mathbf{p}$ the parameter vector of a multinomial random vector, or in the domain of a Dirichlet random vector, then $\mathbf{p}$ is confined to a simplex:

$$(\forall i)\, p_i \geq 0; \quad \sum_{i=1}^{k} p_i = 1.$$

This is a compact convex set in $\mathbf{R}^{k-1}$ parameterized by

$$p_1 \in [0,1], p_2 \in [0, 1-p_1], \ldots, p_{k-1} \in [0, 1-(p_1 + \cdots + p_{k-2})],$$

with $p_k = 1 - (p_1 + \cdots p_{k-1})$ determined by $k-1$ previous choices.

Conjugate Densities

Let $F = F(\mathbf{a}; \mathbf{x})$ be a joint probability density function for a random vector taking values $\mathbf{x} \in \mathbf{R}^n$, with *shape parameters* $\mathbf{a} = (a_1, \ldots, \mathbf{a}_m)$.

Example: Dirichlet density, shape parameters $\{\alpha_i\}$, variables $\{p_i\}$.

A *conjugate density* is another density (or probability function, for discrete r.v.s) with the roles of $\mathbf{a}$ and $\mathbf{x}$ interchanged.

Example: multinomial probability function versus Dirichlet density.

Dirichlet Properties

Describe uncertainty in Dirichlet r.v. $X = (X_1, \ldots, X_k)$ with p.d.f. f, using shape parameters $\mathbf{a} = (\alpha_1, \ldots, \alpha_k)$:

- Write $A = \sum_j \alpha_j$.
- Mean $\mathrm{E}(X_i) = \alpha_i / A$, so $\mathrm{E}(X) = A^{-1}\mathbf{a}$
- Mode $\arg\max f = (A - k)^{-1}(\mathbf{a} - \mathbf{1})$
- Variances $\mathrm{Var}(X_i) = \alpha_i (A - \alpha_i)/(A^2(A+1))$.
- Covariance $\mathrm{cov}\,(X_i, X_j) = \alpha_i(\delta_{ij}A - \alpha_j)/(A^2(A+1))$.

Application

Setup: unknown success parameters $\mathbf{p}$ in sampling k categories.

Prior density: Dirichlet with shape $\mathbf{a} = (\alpha_1, \ldots, \alpha_k)$.

Experiment: collect n samples with n_i in category i. $i = 1, \ldots, k$.

Bayes' rule gives the posterior density:

$$f_{\text{posterior}}(\mathbf{p}|\mathbf{n}) \propto f_{\text{prior}}(\mathbf{p}) \mathbf{f}_{\text{experiment}}(\mathbf{n}|\mathbf{p}),$$

with $\mathbf{n} = (n_1, \ldots, n_k)$.

$$\left(p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1}\right) \left(p_1^{n_1} \cdots p_k^{n_k}\right) \propto p_1^{n_1 + \alpha_1 - 1} \cdots p_k^{n_k \alpha_k - 1},$$

Recognize Dirichlet density with new shape parameters $\mathbf{a} + \mathbf{n}$.

Initial Choices

Uninformative Dirichlet prior: Shape parameters $\mathbf{a} = \mathbf{1}$ gives the uniform density on $\mathbf{p}$.

Repeated experiments: outcomes of experiments $\mathbf{n}$ and $\mathbf{n}'$ add to give new shape parameters $\mathbf{a} + \mathbf{n} + \mathbf{n}'$.

Exercise: Fix i with $1 \leq i \leq k$. Under what conditions $\mathrm{Var}(p_i) \to 0$ as the number of experiments tends to infinity?

Monte Carlo Integration

Theorem

Suppose that $\{X_k : k = 1, 2, \ldots\}$ *is an ergodic Markov chain on a finite state space* $S = \{1, \ldots, m\}$. *Let* $\pi \in \mathbf{R}^m$ *be its stationary distribution. Then for any bounded function* F *on* S*,*

$$\frac{1}{N}\sum_{k=1}^{N} F(X_k) \to \sum_{i=1}^{m} F(i)\pi(i),$$

almost surely as $N \to \infty$.

Proof.

Exercise. $\square$

Remark. The Birkhoff ergodic theorem is a more general version of this result. Its proof may be found in `MetropHastingsEtc.pdf` on the class website.

Metropolis Algorithm

Goal: given $g = C\pi$ with pdf $\pi \in \mathbf{R}^m$ and unknown constant C, construct an ergodic Markov chain with stationary pdf π.

Idea: from an initial Markov chain with transition function M,

- If $X_n = i$, sample random j from distribution $M(i, \cdot)$.
- Define an acceptance function $0 \leq a(i,j) \leq 1$.
- Let $X_{n+1} = j$ with probability $a(i,j)$, else keep $X_{n+1} = i$.
- Increment $n \leftarrow n+1$ and repeat.

Theorem (Metropolis-Hastings)

To get the desired stationary distribution for X, choose

$$a(i,j) = \min\left\{1, \frac{\pi(j)M(j,i)}{\pi(i)M(i,j)}\right\} = \min\left\{1, \frac{g(j)M(j,i)}{g(i)M(i,j)}\right\}.$$

Simplifications

- If $M(i,j) = M(j,i)$ is symmetric, then

$$a(i,j) = \min\left\{1, \frac{g(j)M(j,i)}{g(i)M(i,j)}\right\} = \min\left\{1, \frac{g(j)}{g(i)}\right\}.$$

- If M is ergodic with stationary pdf p, then $\lim_{k\to\infty} M^k(i,j) = p(j)$ for all j. Use this limit to get

$$(\forall i)a(i,j) = \min\left\{1, \frac{g(j)p(i)}{g(i)p(j)}\right\}.$$

- If $g(j) = L(j)p(j)$ is a Bayesian posterior with prior p and likelihood L, then

$$(\forall i)a(i,j) = \min\left\{1, \frac{g(j)p(i)}{g(i)p(j)}\right\} = \min\left\{1, \frac{L(j)}{L(i)}\right\}.$$

Programming Issues

- See the example R code in `07metro.txt`
- Reduce the influence of intial state X_0 with a "burn-in" period.
- Estimate convergence with multiple chains X^l, $l = 1, 2, \ldots,$ and for large N let

$$\Phi_l(N) = \frac{1}{N} \sum_{k=1}^{N} F(X_k^l),$$

  Then $\mathrm{Var}(\Phi(N))$ is an estimate for mean squared error in $\langle F, \pi \rangle \approx \mathrm{E}(\Phi(N))$.
- Extend to random vectors componentwise by "Gibbs sampling".

Gibbs Sampling

Suppose X, Y are random vectors with joint pdf $f(\mathbf{x}, \mathbf{y})$.

To generate samples (X, Y) from f, find:

- marginal pdfs $f_X(\mathbf{x})$ and $f_Y(\mathbf{y})$
- conditional pdfs $f_{X|Y}(\mathbf{x}, \mathbf{y})$ and $f_{Y|X}(\mathbf{x}, \mathbf{y})$

Then iterate for $n = 1, 2, \dots$ from initial $X = \mathbf{x}_0$ and $Y = \mathbf{y}_0$:

- get sample $X = \mathbf{x}_{n+1}$ using $f_{X|Y}(\cdot, \mathbf{y}_n)$,
- get sample $Y = \mathbf{y}_{n+1}$ using $f_{Y|X}(\mathbf{x}_n, \cdot)$,

Remark. If X, Y are independent, then $f_{X|Y}(\cdot, \mathbf{y}) = f_Y(\mathbf{y})$ and $f_{Y|X}(\mathbf{x}, \cdot) = f_X(\mathbf{x})$, so only the marginal pdfs are needed.

Example: Dirichlet Gibbs Sampling

Let $\alpha_0 \stackrel{\text{def}}{=} \alpha_1 + \cdots + \alpha_k$ in the Dirichlet pdf on $\mathbf{p} = (p_1, \ldots, p_k)$:

$$f(\mathbf{p}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1}.$$

Lemma

If $X = (X_1, \ldots, X_k)$ has the Dirchlet pdf, namely $X \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, then the marginal pdf for X_1, removing p_j for all $j = 2, \ldots k$, is:

$$f_1(p_1) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0 - \alpha_1)} p_1^{\alpha_1 - 1}(1 - p_1)^{\alpha_0 - \alpha_1 - 1}$$

In other words, $X_1 \sim B(\alpha_1, \alpha_0 - \alpha_1)$. $\square$

Dirichlet Conditional PDF

The conditional pdf $f_{2,\ldots,k|1}(\mathbf{p})$, given $X_1 = p_1$, is therefore

$$
\begin{aligned}
\frac{f(\mathbf{p})}{f_1(p_1)} &= \frac{\Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(\alpha_0-\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} \frac{p_1^{\alpha_1-1}p_2^{\alpha_2-1}\cdots p_k^{\alpha_k-1}}{p_1^{\alpha_1-1}(1-p_1)^{\alpha_0-\alpha_1-1}} \\
&= \frac{\Gamma(\alpha_2+\cdots+\alpha_k)}{\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} \frac{p_2^{\alpha_2-1}\cdots p_k^{\alpha_k-1}}{(1-p_1)^{\alpha_2+\cdots+\alpha_k-1}} \\
&= \frac{\Gamma(\alpha_2+\cdots+\alpha_k)}{\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} \bar{p}_2^{\alpha_2-1}\cdots\bar{p}_k^{\alpha_k-1}(1-p_1)^{k-1},
\end{aligned}
$$

where $\bar{p}_i = p_i/(1-p_1) = p_i/(p_2+\cdots+p_k)$, for $i = 2,\ldots,k$, so that $\bar{p}_2+\cdots+\bar{p}_k = 1$.