Course Overview

Scott C. Schmidler

Stat 863: Advanced Statistical Computing
Duke University
Fall 2018

Instructor Info

Professor

- Prof Scott Schmidler
- scott.schmidler@duke.edu *(put "Stat 863" in subject line.)*
- Office: 112A Old Chemistry
- OH: Fri 2-3:30pm

TA

- Kyle Burris
- kyle.burris@duke.edu
- Statistical Education & Consulting Center (SECC)
  - stat.duke.edu/courses/resources-students
  - TA hours, help center
- OH: TBA

Course overview

This course covers advanced statistical computing with an emphasis on techniques for Bayesian analysis. Lecture topics can be divided into two parts: **algorithms** for various problems in statistical inference and model selection, and **theoretical foundations**. We will spend significant time on *approximate integration*, especially *Monte Carlo sampling* algorithms, and will also cover *optimization* algorithms used for minimization, maximization, and approximation of functions. Along the way we will consider several classes of advanced statistical models as motivating applications. Additional topics in numerical analysis, computational complexity, and specialized Bayesian approximation will be included as time permits.

Previous years' course topics have included

- Numerical integration
- Advanced MCMC
  - Adaptive MCMC, Hamiltonian MC, annealing & tempering, slice sampling, Wang-Landau, ...
  - convergence diagnostics; convergence theory
- Sequential Monte Carlo
- Variational Approximation
- Optimization
  - gradient and conjugate gradient; Newton-style methods; derivative-free methods; convexity
  - convergence analysis
- Belief propagation

And methods that combine more than one of these.

Class prerequisites

Formal prerequisites:

- Stat 831
- Stat 721

You'll need familiarity with:

- multivariate conjugate manipulations
- basic Markov chain theory in finite and general state spaces
- basic MCMC: Metropolis and Gibbs samplers
- numerical programming

I will do my best to accomodate everyone.

Grading

Homeworks	35%
Paper presentation	15%
Final project	50%

HWs will be approximately 1/week for the first 4-5 weeks.

¡Details of paper presentations, final project will be handed out next week.

Statistical computing: the two paradigms

- **Optimization**: minimize loss function
  - key task in non-computation
  - Often (penalized) negative log-likelihood, but not always
  - May be subject to constraints
- **Integration**: compute posterior expectations
  - key task in Bayesian computation

Unified by *decision theory*: solution to optimization problem may be an *integral*.

General problem statement

A general problem in (esp Bayesian) statistics and statistical mechanics is calculation of integrals of the form:

$$\langle h\rangle_\pi = E_\pi\left(h(x)\right) = \int_{\mathcal{X}} h(x)\pi(dx)$$

Bayesian statistics: for prior π_0 and sampling distn L

$$\pi(x) \propto \pi_0(x)L(y;x)$$

Statistical mechanics: for potential U and $\beta = \frac{1}{k_B T}$

$$\pi(x) \propto \exp -\beta U(x)$$

In both cases, $Z = \int_{\mathcal{X}} \pi(dx)$ typically unknown.

MCMC

A common, powerful approach is *Monte Carlo* integration:

$$\langle h\rangle \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i) \qquad \text{for} \quad X_1, X_2, \ldots, X_n \sim \pi$$

When sampling π is difficult, can construct a *Markov chain* with limiting dist π and simulate to convergence.

Metropolis:

- Draw y from proposal kernel $Q(x, dy)$
- Accept with prob $\alpha(x,y) = \min(1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)})$

π-invariant, so if irreducibile, aperiodic, chain has limiting distn π.

Bayesian Analysis

Statistics as probability.

Example: (Generalized) Linear Regression

$$g(\mu_y) = \alpha + \textstyle\sum_{i=1}^{p} \beta_i x_i$$

g is *link function*, e.g. $g(x) = x$ or $g(x) = \operatorname{logit}(x)$
x_i's are covariates (or *predictors*, or *features*)

Denote parameters $\theta = (\alpha, \beta_1, \ldots, \beta_p, \sigma^2)$.
Specify *prior* distribution $\pi_0(\theta)$, and base conclusions on *posterior*:

$$\pi(\theta|Y) = \frac{L(Y;\theta)\pi_0(\theta)}{m(Y)}$$

Monte Carlo Integration

Usually $\pi(\theta \mid Y)$ cannot be obtained analytically.

We can summarize a general distribution $\pi(X)$ by drawing samples $X^{(1)}, \ldots X^{(n)} \sim \pi(X)$, make histograms.

In addition, sampling enables *Monte Carlo integration:*

$$\hat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f(X^{(i)}) \xrightarrow{\text{a.s.}} E_\pi f(X)$$

under mild conditions (SLLN).

Enables approximation of most quantities of interest by choosing appropriate f: e.g. x, x^2, $1_A(x)$, etc.

Markov Chain Monte Carlo

Often sampling from $\pi(X)$ is itself non-trivial.

MCMC: Construct a Markov chain with limiting distribution $\pi(X)$, and run it "long enough" that samples are (approximately) from π.

If MC is *ergodic* (aperiodic, irreducible, π-invariant) we get a similar SLLN (ergodic theorem).

How to construct?

Metropolis Algorithm

General case: $\pi(dx) = \pi(x)\mu(dx)$ for some σ-finite μ on $\mathcal{X}$.

To draw samples from $\pi(x)$:

Choose *proposal* kernel $q(x, x')$.

Metropolis-Hastings

- Draw $x^* \sim q(x^{(t)}, \cdot)$
- Set $x^{(t+1)} = \begin{cases} x^* & \text{w/ prob } \alpha = \min\left(1, \frac{\pi(x^*)q(x^*,x)}{\pi(x)q(x,x^*)}\right) \\ x^{(t)} & \text{otherwise} \end{cases}$

Result: reversible MC with stationary distribution $\pi(x)$.

Random-walk Metropolis

How to choose proposal q? Common choices:

- Random walk: $x^* = x + \epsilon$
  - e.g. if $\mathcal{X} = \mathbb{R}^d$, take $\epsilon \sim N(0, \sigma^2 I_d)$.

Random-walk Metropolis

How to choose proposal q? Common choices:

- Random walk: $x^* = x + \epsilon$
  - e.g. if $\mathcal{X} = \mathbb{R}^d$, take $\epsilon \sim N(0, \sigma^2 I_d)$.
- Independent: $q(x, x') \equiv q(x')$ (MIS)

Random-walk Metropolis

How to choose proposal q? Common choices:

- Random walk: $x^* = x + \epsilon$
  - e.g. if $\mathcal{X} = \mathbb{R}^d$, take $\epsilon \sim N(0, \sigma^2 I_d)$.
- Independent: $q(x, x') \equiv q(x')$ (MIS)

Works under simple conditions (support).

Random-walk Metropolis

How to choose proposal q? Common choices:

- Random walk: $x^* = x + \epsilon$
  - e.g. if $\mathcal{X} = \mathbb{R}^d$, take $\epsilon \sim N(0, \sigma^2 I_d)$.
- Independent: $q(x, x') \equiv q(x')$ (MIS)

Works under simple conditions (support). May not be efficient.

Example: Suppose $\pi(x) = N_2(0, \Sigma)$

Consider $\Sigma = \begin{bmatrix} \sigma_1 & \rho \\ \rho & \sigma_2 \end{bmatrix}$, with $\sigma_1 = 2$, $\sigma_2 = 1$ and $\rho = .95$

Efficiency

Statistical Efficiency: $var(\hat{f})$

Under reasonably weak conditions*, for any function f with $var_\pi(f) \leq \infty$, we obtain a CLT:

$$\sqrt{n}(\bar{f}_n - \mu_f) \to N(0, \sigma^2_{\bar{f}_n})$$

where

$$\sigma^2_{\bar{f}_n} = \sigma^2_f(1 + 2\sum_{j=1}^{n}(1 - \frac{j}{n})\rho_j)$$

and ρ_j is lag-j autocorrelation:

$$\rho_j = \frac{1}{\sigma^2_f} E\left((f(X^{(n)}) - \mu_f)(f(X^{(n+j)}) - \mu_f)\right)$$

A first look: Adaptive Metropolis and AMIS