adaptive mcmc: informal overview

brooks

outline

1. what is mcmc?
2. what is adaptive mcmc?
3. is that even legal?
4. math is hard, let’s go sampling

this is primarily based on

- Atchadé, Fort, Moulines and Priouret, *Adaptive Markov chain Monte Carlo: Theory and Methods,* 2011

- Roberts and Rosenthal, *Examples of Adaptive MCMC*, 2009

- Andrieu and Thoms, *A Tutorial on Adaptive MCMC*, 2008

- Roberts and Rosenthal, *Coupling and Ergodicity of Adaptive MCMC*, 2005

mcmc in brief

- we want to sample from $\pi(x)$, but $\pi(x)$ is intractable
- mcmc: construct a markov chain transition operator $T(x \to x')$ which has $\pi(x)$ as its stationary (limiting, invariant, …) distribution
- invariant distribution means $\pi T = \pi$
- draw samples from T long enough and they will approximate π
- “ergodic theorem”: time average == space average

what do we need?

a reminder: necessary conditions for the sampler to correctly draw from the target distribution π are

- *invariance*: $\pi(x') = \int \pi(x) T(x \to x')\, dx$
- *ergodicity*: $p^{(n)}(x) \to \pi(x)$ as $n \to \infty$
  - an ergodic chain is *irreducible* and *aperiodic*
  - we'll come back to this later

metropolis–hastings

where does the transition operator T come from?
easiest approach: metropolis-hastings algorithm

- define a *proposal* distribution $q(x \to x')$
- define an *acceptance* distribution

$$A(x \to x') = \min[1, \pi(x')q(x \to x') / \pi(x)q(x' \to x)]$$

- *transition* distribution: “propose, maybe accept”

$$T(x \to x') = A(x \to x')\, q(x \to x')$$

metropolis–hastings

- easy to show $T(x \to x')$ generates a reversible markov chain with invariant distribution π: it satisfies detailed balance,

$$\pi(x)\, A(x \to x')\, q(x \to x') = \pi(x')\, A(x' \to x)\, q(x' \to x)$$

- typical choices for $q(x \to x')$, at least in continuous spaces: gaussian random walk proposals

metropolis–hastings

here is your proof:

$$\pi(x)\, T(x \to x')$$
$$= \pi(x)\, A(x \to x')\, q(x \to x')$$
$$= \min\{\, \pi(x) q(x' \to x),\ \pi(x') q(x \to x') \,\}$$
$$= \min\{\, \pi(x') q(x \to x'),\ \pi(x) q(x' \to x) \,\}$$
$$= \pi(x')\, A(x' \to x)\, q(x' \to x)$$
$$= \pi(x')\, T(x' \to x)$$

metropolis–hastings

typical choices for $q(x \to x')$, at least in continuous spaces: gaussian random walk proposals

QUESTIONS:

- gaussian random walk — what are the parameters?
- what other "valid" choices of $q(x \to x')$ are there?
- what happens in high dimensions?

the world is a sad place

“real world algorithm” used by practitioner to choose variance for a gaussian random walk:

1. guess an arbitrary step size / variance
2. see whether it looks like it is converging
3. if it isn’t converging, try another step size
4. after a while, just pick one and let it run for a while

adaptive mcmc

motivation

let's automate the parameter search on $q_\theta(x \to x')$:

1. guess an **initial** step size / variance θ
2. run one mcmc step
3. update step size parameter θ
4. repeat for all time

adaptive random walk mh

- the “adaptive metropolis” algorithm (Haario et al, 2001) follows just this scheme

- automatically tune parameters of MH step size, using the current empirical estimate of the covariance

- first true adaptive mcmc algorithm that i am aware of

optimization criterion

but, what are we trying to optimize? how do we measure proposal quality?

rely on theory / magic: Roberts, Gelman, Gilks 1997; others

- optimal acceptance rate, one dimension: **0.44**
- optimal scaling rate, **d** dimensions, **d** large: **(2.38)^2 Σ/d**, where **Σ** is the “true” covariance
- “rule of thumb” acceptance rate, for (say) when empirical covariance estimate is poor: **0.234**

let's panic

- warning: this is no longer "markov" chain monte carlo
- the proposal distributions $q(x \to x')$ now may depend on the full history, not just on the current state x
- ergodicity of these adaptive algorithms must be proved explicitly!
- adaptive metropolis (Haario et al, 2001): complicated proof for adaptive random walks
- active research: finding "simple" conditions for validity

what kind of adaptation is valid?

- many ad-hoc schemes to update the metropolis-hastings proposal may change the target density
- i will draw a toy counterexample on the whiteboard

explicit counterexample

- discrete state space: $X = \{1, 2\}$

$$\theta \in \Theta := (0, 1)$$

$$P_\theta = \begin{bmatrix} P_\theta(X_i = 1, X_{i+1} = 1) & P_\theta(X_i = 1, X_{i+1} = 2) \\ P_\theta(X_i = 2, X_{i+1} = 1) & P_\theta(X_i = 2, X_{i+1} = 2) \end{bmatrix}$$

$$= \begin{bmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{bmatrix}.$$

- we have $\pi = [0.5, 0.5]$ for any parameter choice.
  but, suppose that θ is a function of x_i:

$$\check{P} := \begin{bmatrix} \theta(1) & 1-\theta(1) \\ 1-\theta(2) & \theta(2) \end{bmatrix} \qquad \check{\pi} = \left(\frac{1-\theta(2)}{2-\theta(1)-\theta(2)}, \quad \frac{1-\theta(1)}{2-\theta(1)-\theta(2)} \right) \neq \pi$$

[example from andrieu and thoms, 2008]

diminishing adaptation

- turns out *diminishing adaptation* is the key general requirement for adaptive mcmc algorithms to still sample from $\pi(x)$

- if the parameters converge to be near some good value eventually, then things are probably okay

- but, be careful not to “overfit” to a crappy value!

- *caveat: diminishing adaptation is neither a sufficient nor necessary condition for ergodicity!*
  … see Bai et al 2008 for examples

diminishing adaptation

how do we design algorithms which exhibit diminishing adaptation? [see Andrieu and Thoms 2008]

- if there is a parameter, ensure that the parameter converges in the limit of infinite samples
- this can be done using robbins-munro updates
- … OR, by dependence on the full history, such that each new sample has o(1/n) impact on the parameter value
- … OR, by only adapting at intervals which become less and less frequent

can we prove our samplers work?

yes, but it's difficult.

theorems / proofs are in Roberts and Rosenthal 2005.

- **Theorem**: diminishing adaptation, along with "simultaneous uniform ergodicity", implies ergodicity

- **Theorem**: diminishing adaptation, along with "containment", implies ergodicity

these are both fairly technical conditions…

ergodicity, containment

taken from K. Latuszynskis slides, at http://www.stats.ox.ac.uk/~evans/CDT/Adaptive.pdf

- **(Diminishing Adaptation)** Let $D_n = \sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\|$ and assume $\lim_{n \to \infty} D_n = 0$ in probability
- **(Simultaneous uniform ergodicity)** For all $\varepsilon > 0$, there exists $N = N(\varepsilon)$ s.t. $\|P^N_\gamma(x, \cdot) - \pi(\cdot)\| \leq \varepsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$
- **(Containment condition)** Let $M_\varepsilon(x, \gamma) = \inf\{n \geq 1 : \|P^n_\gamma(x, \cdot) - \pi(\cdot)\| \leq \varepsilon\}$ and assume $\{M_\varepsilon(X_n, \gamma_n)\}_{n=0}^\infty$ is bounded in probability, i.e. given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$, for all $\delta > 0$, there exists N s.t. $\mathbb{P}[M_\varepsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \geq 1 - \delta$ for all $n \in \mathbb{N}$.

(simultaneous uniform ergodicity is stronger than containment)

help, help

okay: in (Bai et al 2008) we get some other conditions under which containment holds, including:

- finite state space and parameter space
- “compact in some topology in which the transition kernels or proposal kernels have jointly continuous densities”
- …

fairy godmother

… you can always get away with **finite adaptation**!

1. adapt for a while (“learning phase”)
2. stop adapting; exploit

as long as any parameter setting yields a valid sampler, things will be okay

- hopefully the parameter is “not bad”
- … but, how do you decide stopping times?

examples!

adaptive metropolis

propose using the empirical covariance matrix, plus a little bit of regularization:

$$Q_n(x, \cdot) \;=\; (1-\beta)\, N(x,\, (2.38)^2\, \Sigma_n \,/\, d) \;+\; \beta\, N(x,\, (0.1)^2\, I_d \,/\, d)$$

common trick: run it for a while with some initial Σ_0 before beginning adaptation

set ß small, like 0.05-ish

adaptive metropolis

how do we prove it?

- Haario et al, 2001: direct argument based on mixingales
- Bai et al, 2008 (theorem 6): diminishing adaptation + containment.
    - the containment argument hinges on assumptions about target distribution: i.e. lighter than exponential tails, regularity of target, continuity of first derivatives, …

adaptive metropolis within gibbs

what if you are running a metropolis-within-gibbs setup?

i.e. suppose it is a non-conjugate model, and we sample single dimensions one at a time.

- idea: for each single dimension, learn a separate one-dimensional random walk scaling parameter

- optimal acceptance rate, per dimension: **0.44**

adaptive metropolis within gibbs

how do we prove it?

- diminishing adaptation: for each dimension, update the parameter σ by $\delta(n)$
  - ensure $\delta(n) \to 0$ as $n \to \infty$, e.g. $\delta(n) = \min\{ 0.01, n^{-1/2} \}$
- containment: choose a global maximum parameter M, and enforce σ is in the interval $[-M, M]$
  - ensures containment for “large class of target densities”, including any which are “log-concave outside an arbitrary bounded region”
  - in practice the step sizes stabilize nicely; M is just for the proof

regional adaptive metropolis

maybe you want to have different proposals in different "regions" of the space?

- idea: partition the state space into a finite number of disjoint regions
- proposal: $q(x) = \text{Normal}(x, \exp(2a_i))$ for each region i
- target acceptance rate per region: 0.234
- **proof**: ensure diminishing adaptation by decreasing magnitude of updates to a_i, ensure containment looking at drift conditions, and assuming log-concave target

adaptive random scan gibbs

here's something a little different: suppose we're running a gibbs sampler — for D dimensions, we sample in turn from each $p(x_d \mid x_{\setminus d})$

which dimension should we sample from next?

- standard: deterministic ordering
- slightly less standard: uniform distribution over D
- bold: non-uniform distribution over D
- crazy: dynamically updating distribution over D

adaptive random scan gibbs

how do we prove it?

- Levine and Casella, *Optimizing random scan Gibbs samplers*, 2006

- conditions are just (1) a.e. convergence in the weights across dimensions, and (2) ergodic sampler for any fixed set of weights

- also suggests a minimax-optimal criterion for updating the scan weights after each update!

adaptive random scan gibbs

how do we *really* prove it?

- Latuszynski: “the above theorem is simple, neat and wrong”

- Latuszynski, Roberts, Rosenthal, 2013: provides a counterexample, a case where the proposal distribution can converge “too slowly”

- alternate proof: diminishing adaptation + simultaneous uniform ergodicity

  - requires assumption that some “very good” parameter exists, somewhere, we just have to find it

“adaptive random sampling adaptive metropolis within gibbs”

take this setup to the extreme:

- metropolis-within-gibbs setup, where we adapt each individual metropolis proposal $q(x'_d \mid x)$
- … AND we adapt the random sampling distribution over the D dimensions
- proof in Latuszynski et al requires strong uniform ergodicity for each of the individual kernels

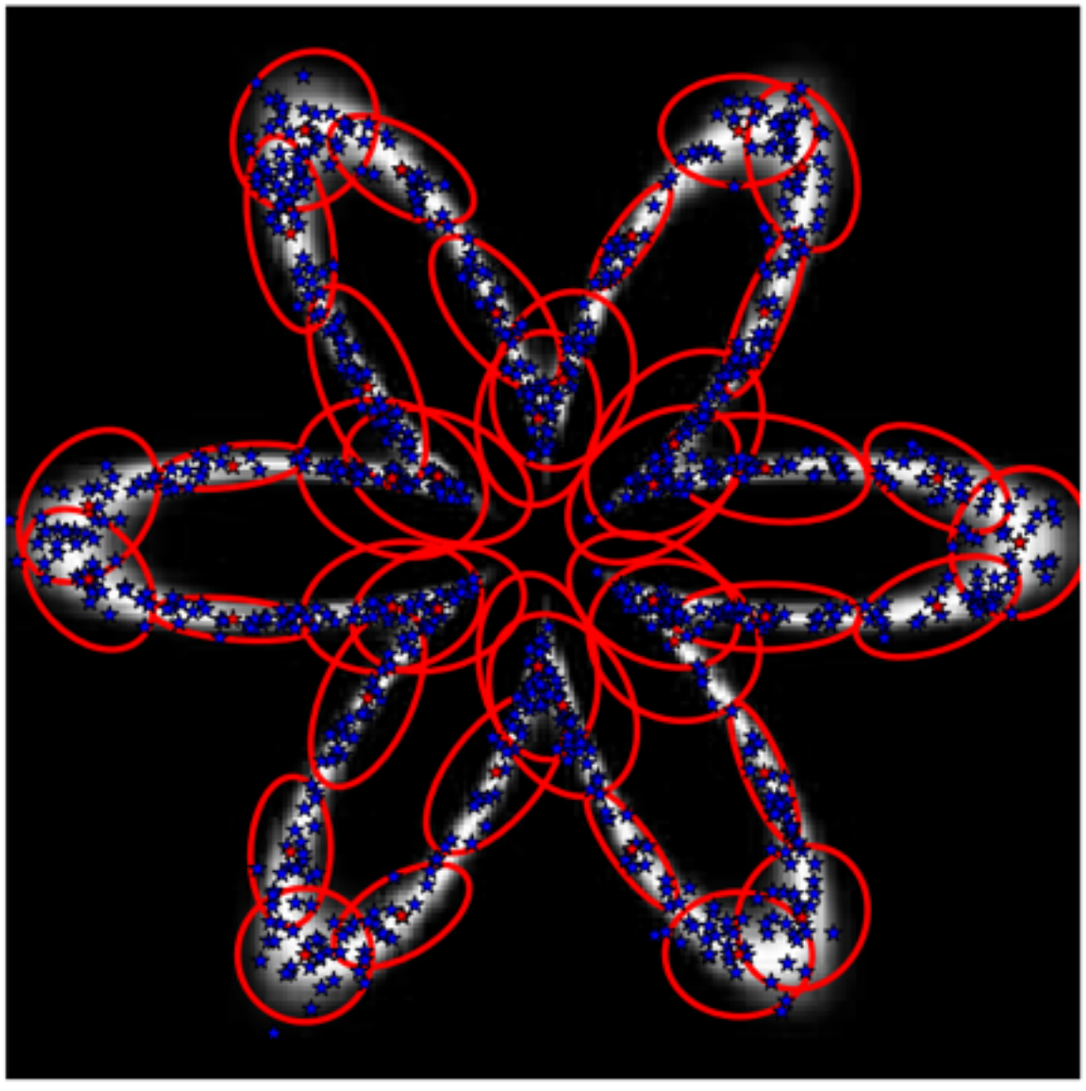
kernel adaptive metropolis hastings

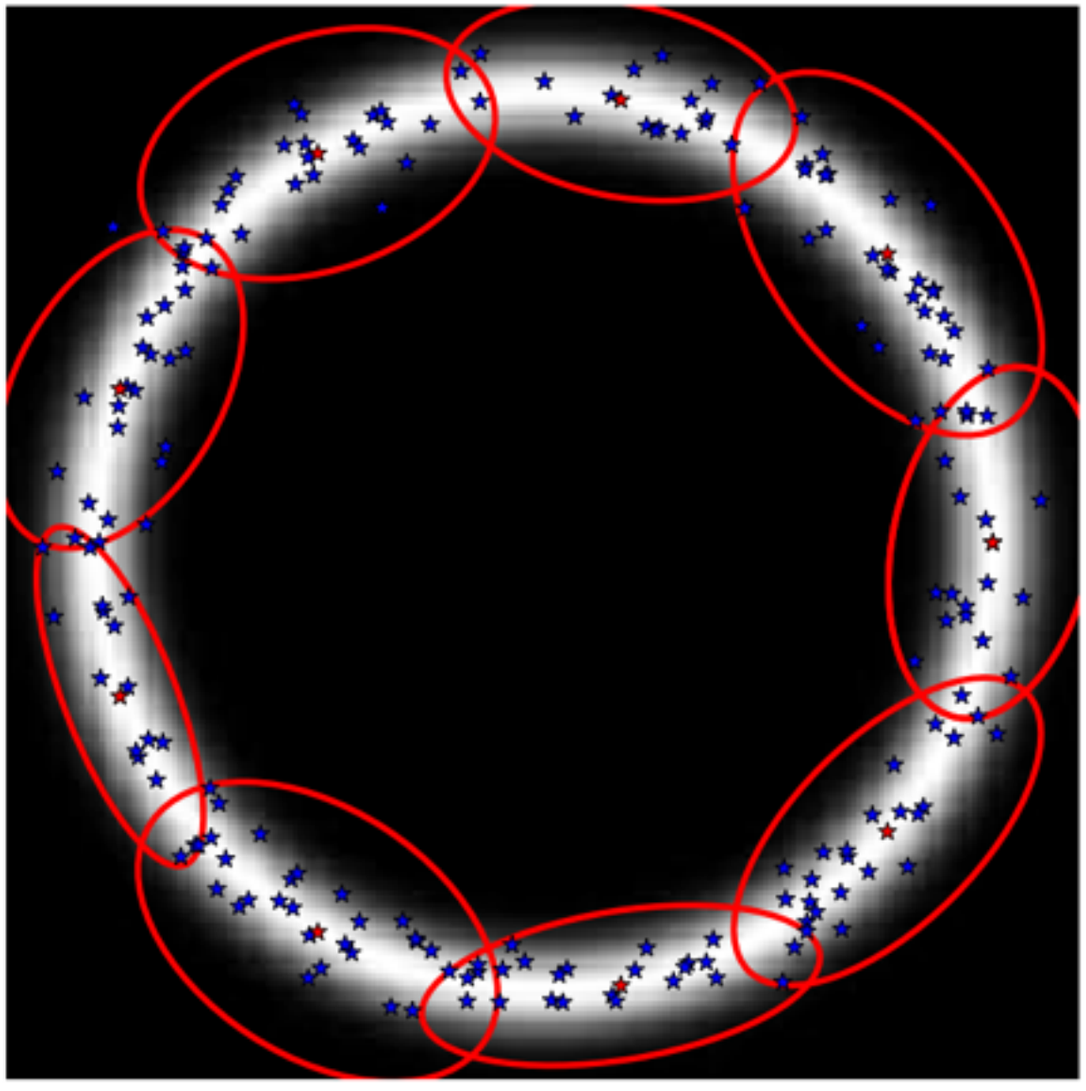
new paper (Sejdinovic et al, icml 2014).

- idea: kernel over the state space
- data points local to proposal region define a covariance operator
- use that covariance as the empirical covariance in the adaptive m–h scheme
- **proof**: ensure diminishing adaptation by adapting with lower and lower probability as time goes on — note that “adapting” here just means learning the general shape of the distribution in different areas

kernel adaptive metropolis hastings

pictures are pretty:





adaptive independent metropolis hastings

Holden et al, 2009.

- idea: what if, instead of a random walk, we just make *independent* proposals and tune the parameter based on our history?

- if we have a decent estimate of the full posterior, independent proposals allow for potentially longer jumps than random walks

- goal: minimize **q(x|history) / π(x)**

- **proof one**: ensure diminishing adaptation, and make sure each proposal is uniformly ergodic

- **proof two**: assume all proposals have uniformly heavier tails than the target — then, we have geometric convergence

adaptive mcmc w/ bayesian optimization

Mahendran et al, 2010.

- idea: instead of optimizing for acceptance rate, what if (instead) we try to minimize the area under the autocorrelation function?

- this is pretty intractable, but bayesian (black-box) optimization can tune parameters

- **proof**: finite adaptation, if only for computational reasons (GP scales poorly with number of samples).

- **proof two**: discrete spaces (Boltzmann machines, …)

thanks

questions?

(relevance to pp?)