Advances and Applications in Perfect Sampling

Ph.D. Dissertation Defense

Ulrike Schneider

advisor: Jem Corcoran

May 8, 2003

Department of Applied Mathematics

University of Colorado

Outline

Introduction

(1) MCMC Methods

(2) Perfect Sampling

Advances

(3) Slice Coupling

(4) Variants on the IMH Algorithm

Applications

(5) Bayesian Variable Selection

(6) Computing Self Energy Using Feynman Diagrams

1. MCMC methods

- MCMC = Markov Chain Monte Carlo
- Originally developed for the use in Monte Carlo techniques (approximating deterministic quantities using random processes).
- Main method is the Metropolis-Hastings algorithm (METROPOLIS, 1953, HASTINGS 1970).
- MCMC has been used extensively outside Monte Carlo methods.

1. MCMC methods

- Simulate from non-standard distributions with the help of Markov chains.
- Artificially create a Markov chain that has the desired distribution as equilibrium distribution.
- Start the chain in some state at time $t = 0$ and run "long enough" (yields approximate samples).

HOW LONG IS LONG ENOUGH??

The problem of assessing convergence is a major drawback in the use of MCMC methods.

2. Perfect Sampling

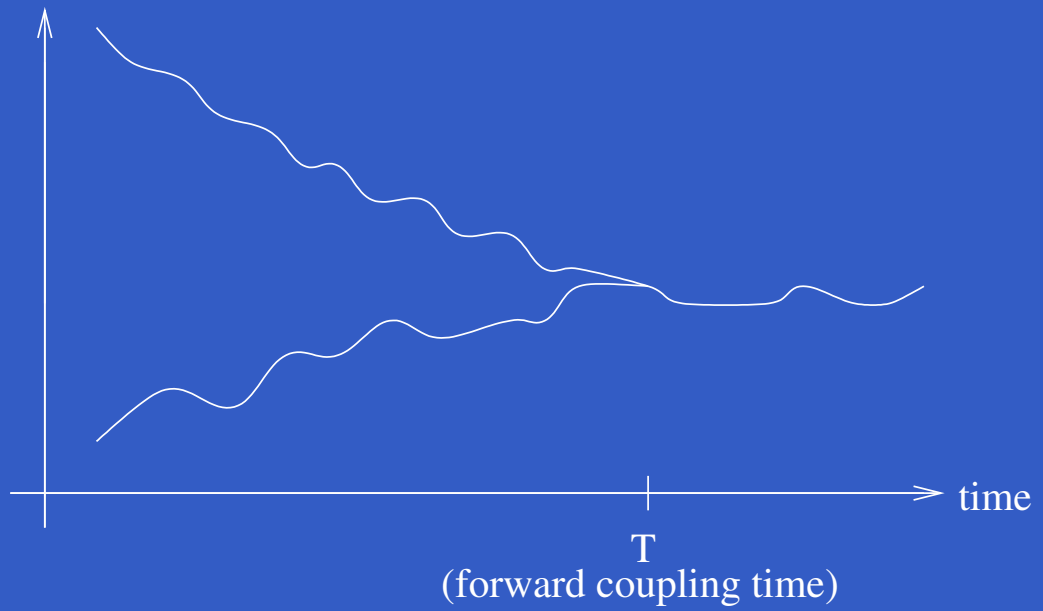
!! MCMC WITHOUT STATISTICAL ERROR !!

Enables exact simulation from the stationary distribution of certain Markov chains. First paper by PROPP and WILSON in 1996.

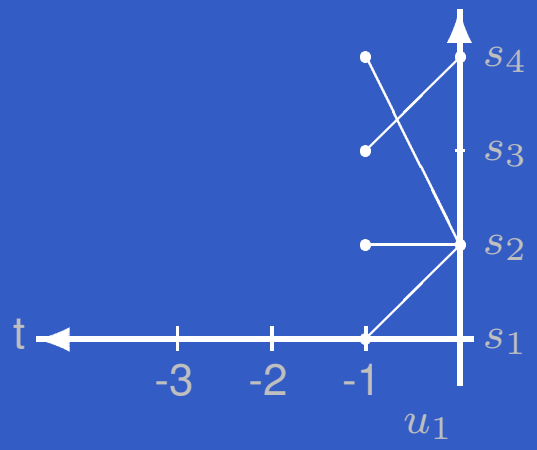
- Parallel chains are started in each state.
- Chains are run as if started at time $t = -\infty$ and stopped at time $t = 0$.
- Can be done in finite time for uniformly ergodic Markov chains.

Essential Idea – Coupling Sample Paths

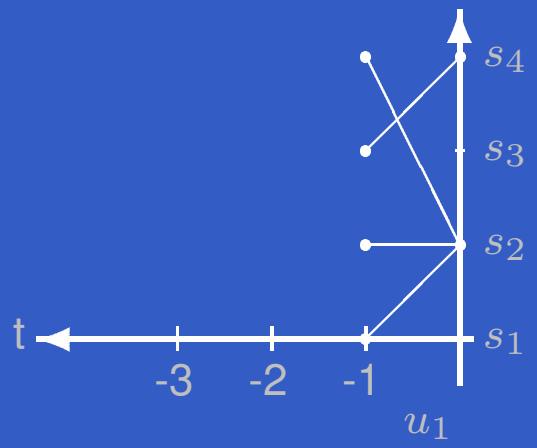
Once two sample paths of a Markov chain couple or coalesce, they stay together.



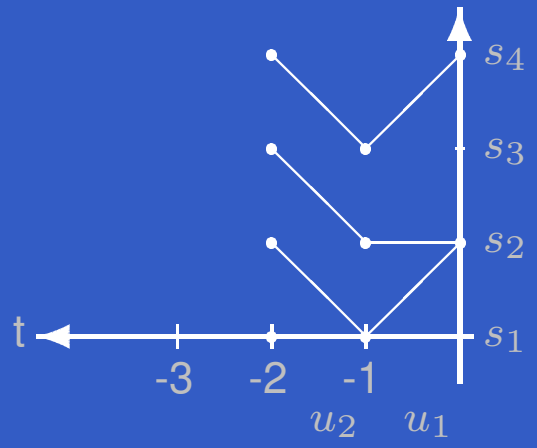
Perfect Sampling: Go back far enough in time, so that by time $t = 0$, all chains have coalesced (backward coupling time).



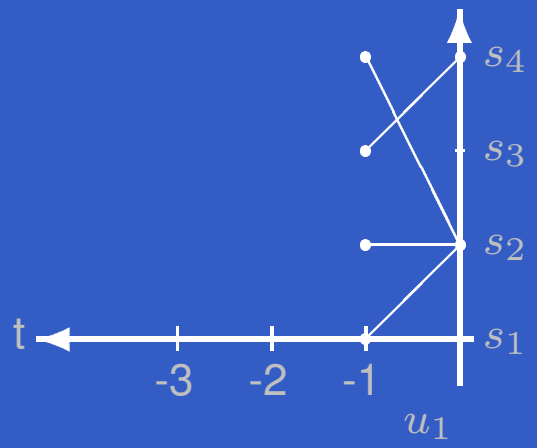
Starting at -1 : no coalescence by $t = 0$.



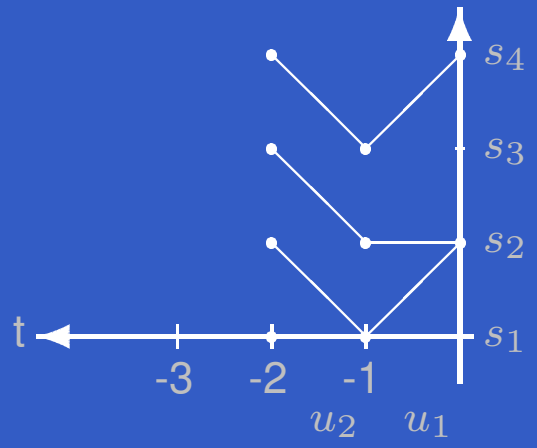
Starting at -1 : no coalescence by $t = 0$.



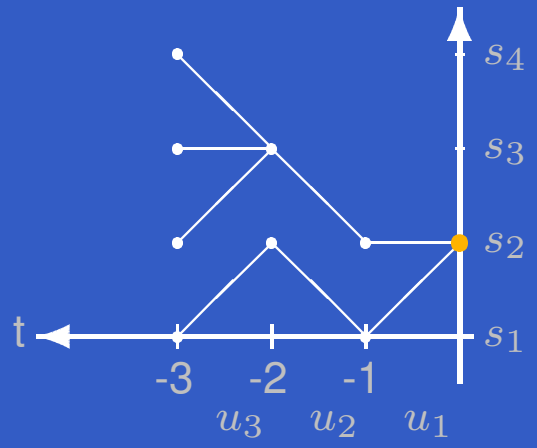
Starting at -2 : no coalescence by $t = 0$.



Starting at -1 : no coalescence by $t = 0$.



Starting at -2 : no coalescence by $t = 0$.

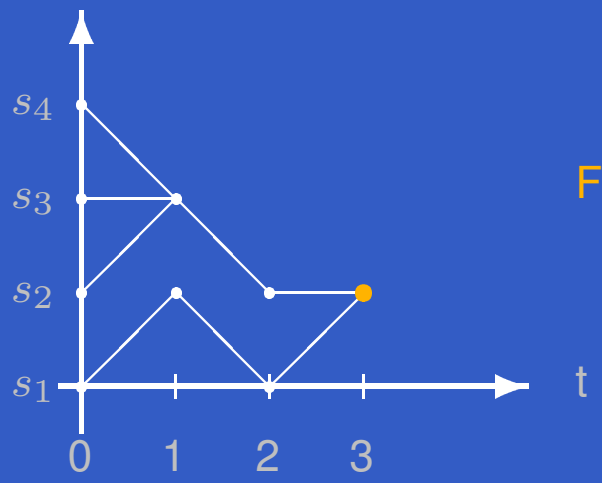


Starting at -3 : coalescence by $t = 0$.

s_2 is a draw from the stationary dist.!

Why backwards?

- If started stationary, the Markov chain is stationary at all fixed time steps.
- Time of coalescence is random.
- Reporting states at the random forward coupling time no longer necessarily gives draws from the stationary distribution.



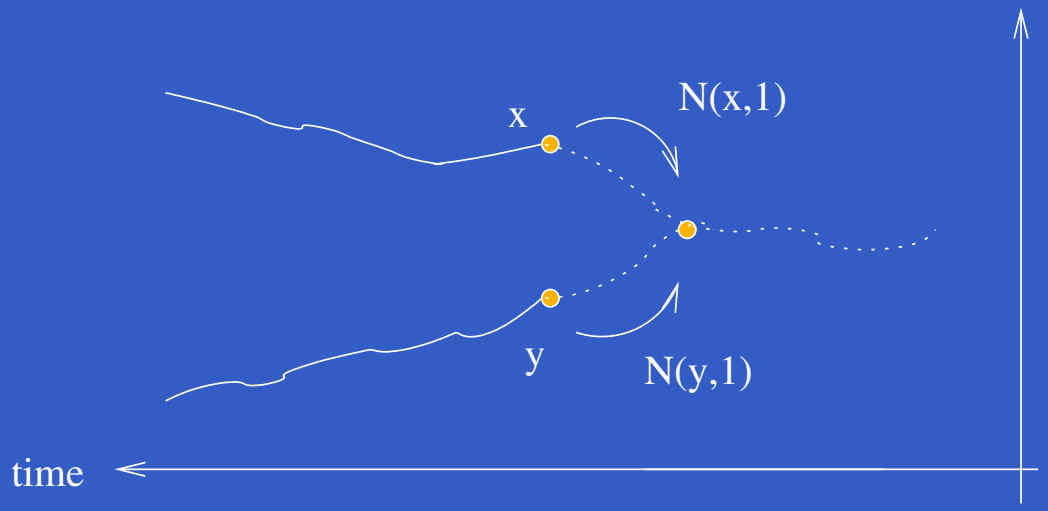
Forward coupling time $T = 3$.

The Challenge

- What about infinite and continuous state spaces?
- Theoretically works for all uniformly ergodic chains, BUT we need a way to detect a backward coupling time!
- Ideas include minorization criteria, bounding processes, perfect slice sampling, Harris coupler, IMH algorithm, slice coupling, ...
- Theoretical development has slowed down.
- Focus has shifted towards applying perfect sampling algorithms to relevant problems – applications are non-trivial ...

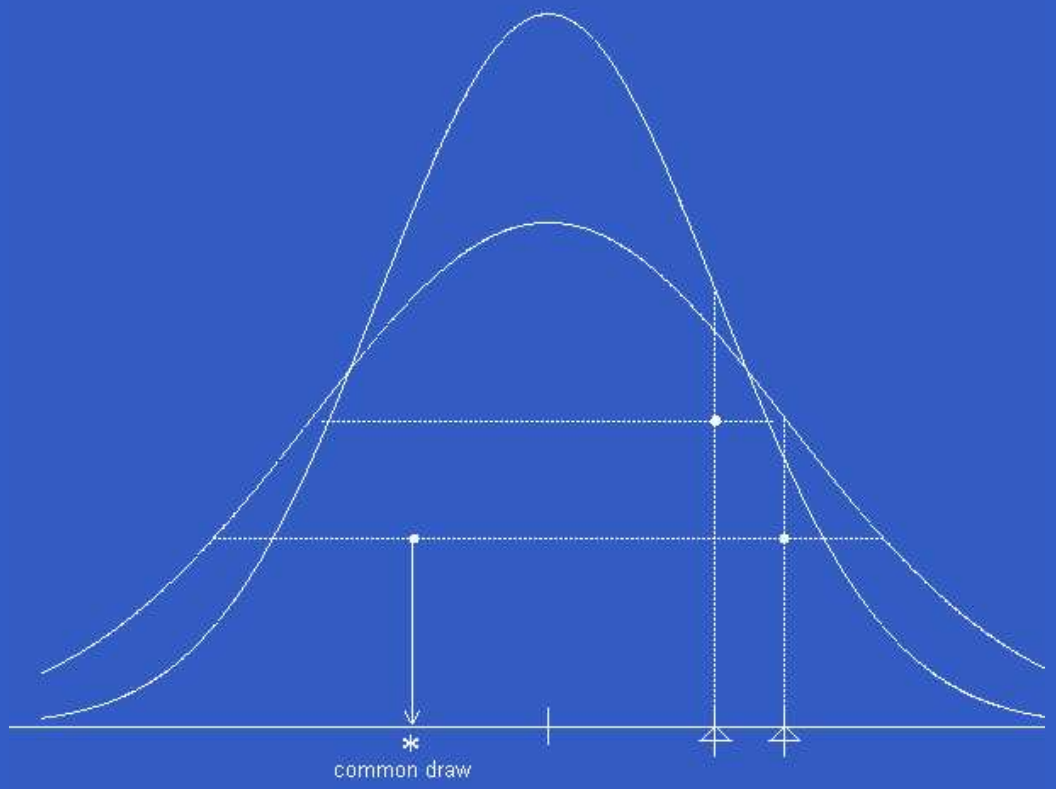
3. Slice Coupling

- Get a potentially common draw from two different continuous distributions!
- Will enable us to couple continuous sample paths in perfect sampling algorithms.



3. Slice Coupling

"“Slicing” uniformly under the curve of a density function (area represents probability!)



3. Slice Coupling

"Slicing" uniformly under the curve of a density function (area represents probability!)

Different techniques

- Layered multishift coupler (Wilson, 2000).
- Folding coupler, shift-and-fold step (Corcoran and Schneider, 2003).
- Shift-and-patch algorithm for non-invertible densities (Corcoran and Schneider, 2003).

4. The IMH-algorithm

- IMH = independent Metropolis-Hastings
- IMH is the perfect counterpart (CORCORAN AND TWEEDIE, 2000) to the Metropolis-Hastings algorithm using an independent candidate density.

Regular Metropolis-Hastings

- Want sample from $\pi(x)$ (possibly unnormalized).
- Use instead a candidate density $q(x)$.
- Create the Metropolis-Hastings chain X_t according to the transition law:
  - Assume $X_t = x$ and draw a candidate $y \sim q(.)$
  - Accept ($X_{t+1} = y$) the candidate with probability

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

  - Otherwise, reject ($X_{t+1} = x$).

This Markov chain has stationary distribution $\pi(x)$.

Perfect IMH

- Reorder the state space S, assume that there exists $l \in S$, such that

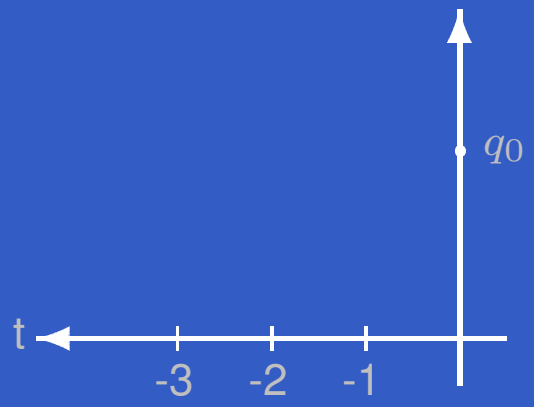
$$\frac{\pi(l)}{q(l)} = \max_{x \in S} \frac{\pi(x)}{q(x)}.$$

- This l (lowest element of S) satisfies

$$\alpha(l, y) \leq \alpha(x, y) \quad \forall x \in S,$$

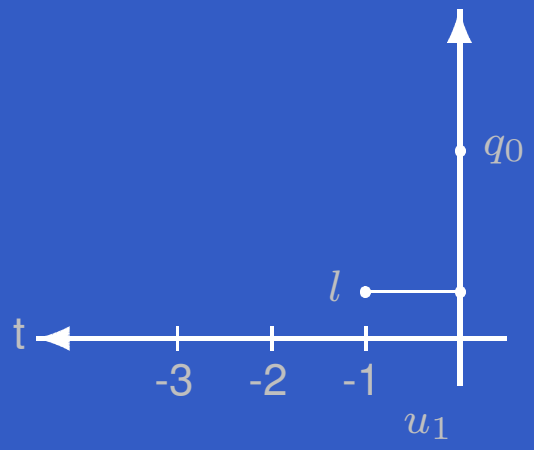
i.e. if we accept to move from state l to state y, any other $x \in S$ will also move to y. This allows us to detect a backward coupling time.

Illustration of IMH



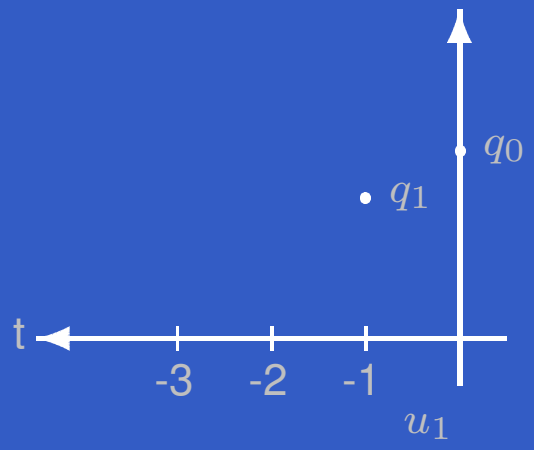
Draw candidate q_0 at time $t = 0$.

Illustration of IMH



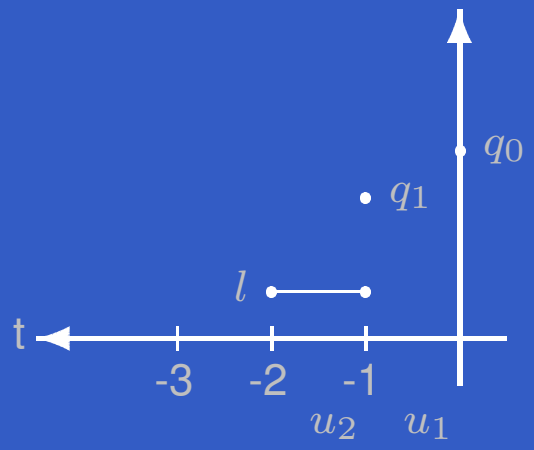
l rejects candidate q_0 at time $t = 0$.

Illustration of IMH



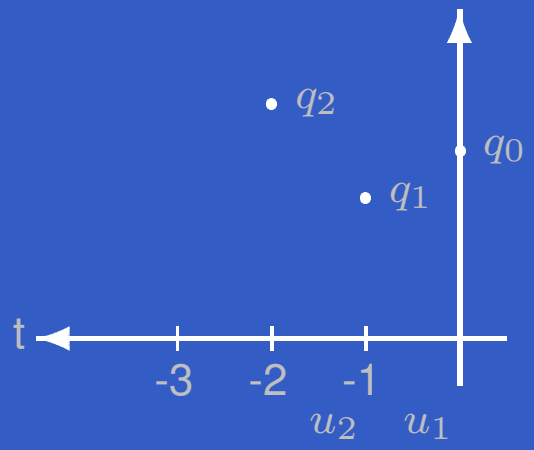
Draw candidate q_1 at time $t = -1$.

Illustration of IMH



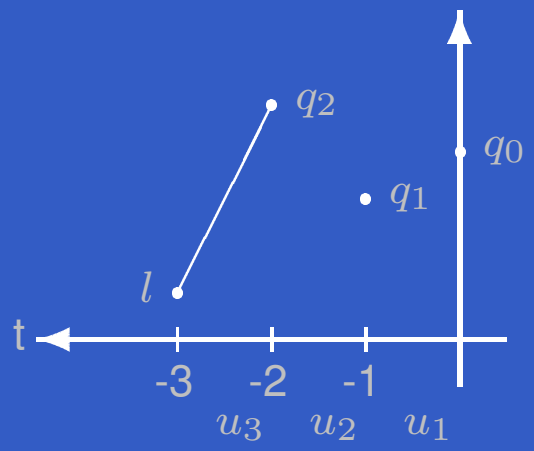
l rejects candidate q_1 at time $t = -1$.

Illustration of IMH



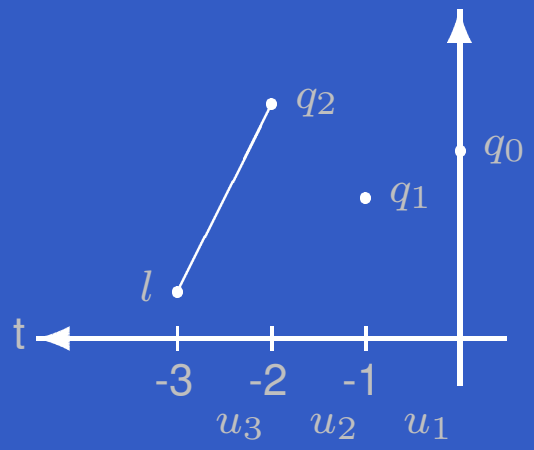
Draw candidate q_2 at time $t = -2$.

Illustration of IMH

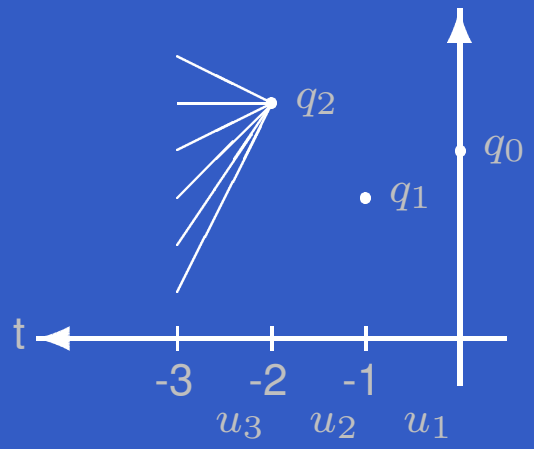


l accepts candidate q_2 at time $t = -2$!

Illustration of IMH

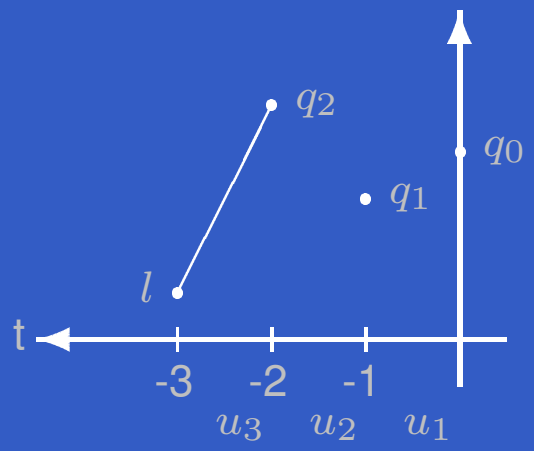


l accepts candidate q_2 at time $t = -2$!

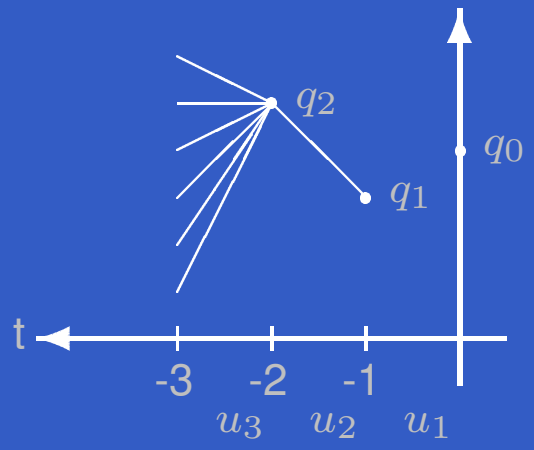


Any state accepts q_2 at time $t = -2$!

Illustration of IMH

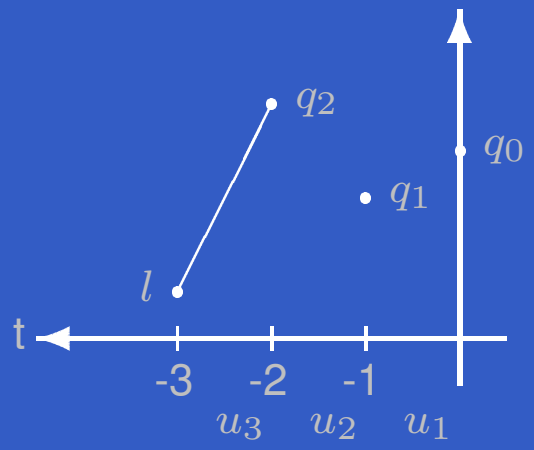


l accepts candidate q_2 at time $t = -2$!

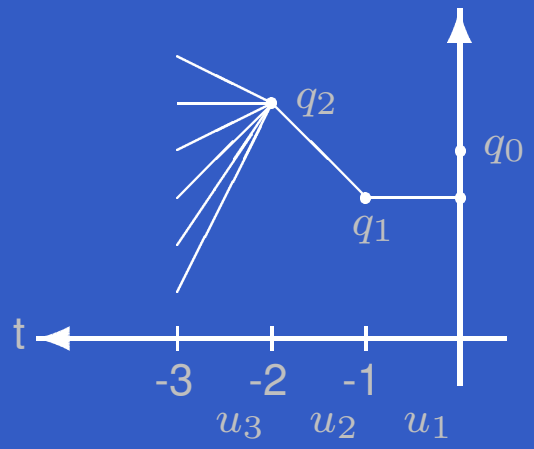


Transition forward – accept candidate q_1.

Illustration of IMH

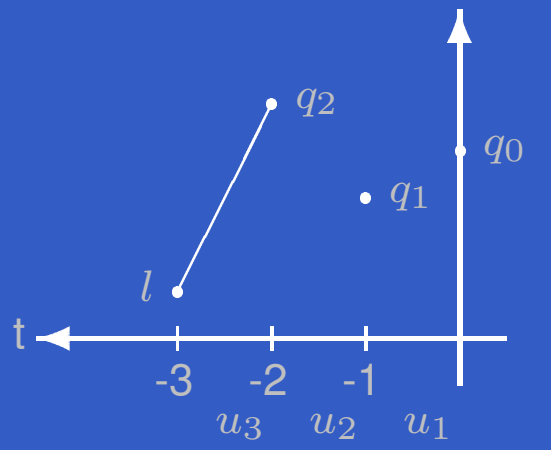


l accepts candidate q_2 at time $t = -2$!

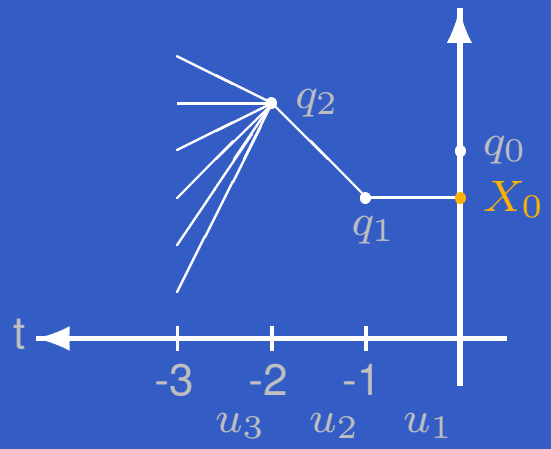


Transition forward – reject candidate q_0.

Illustration of IMH



l accepts candidate q_2 at time $t = -2$!



Transition forward – reject candidate q_0.

X_0 is a draw from π!

Variants on IMH

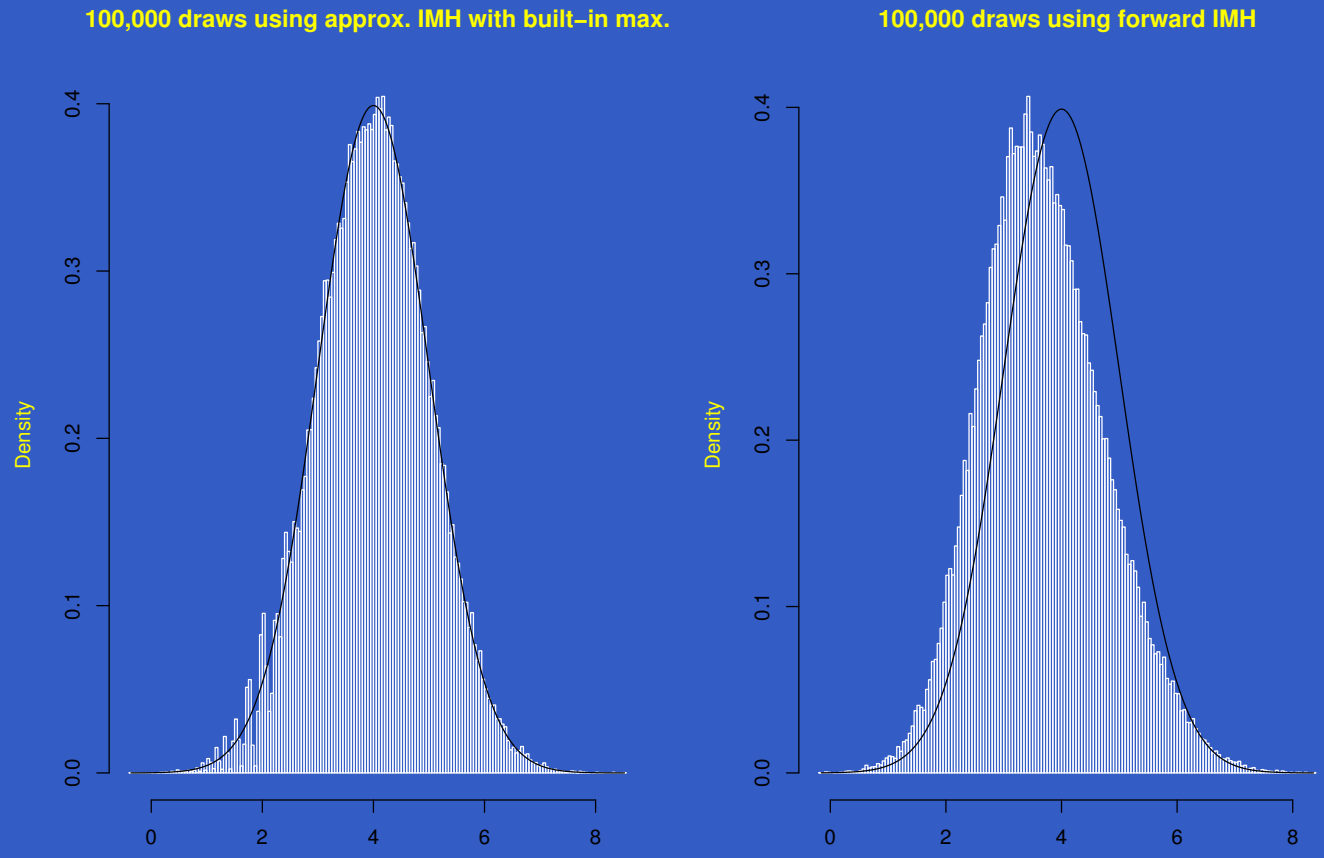
We only need to know the value of the maximum $\frac{\pi}{q}$-ratio, but not where it occurs.

(1) Bounded IMH (SCHNEIDER and CORCORAN, 2002)
We do not even need to know the maximum exactly!
An upper bound for $\frac{\pi}{q}$ is a lower bound for $\alpha(l, y)$ –
still obtain can therefore still obtain exact draws.
Used in

- Variable selection problem
- Computing self-energy for the interacting fermion problem

Variants on IMH

(2) Approximate IMH (CORCORAN and SCHNEIDER, 2003)
If no upper bound is available, a built-in random search appears to outperform "regular forward" IMH at the same computational cost.



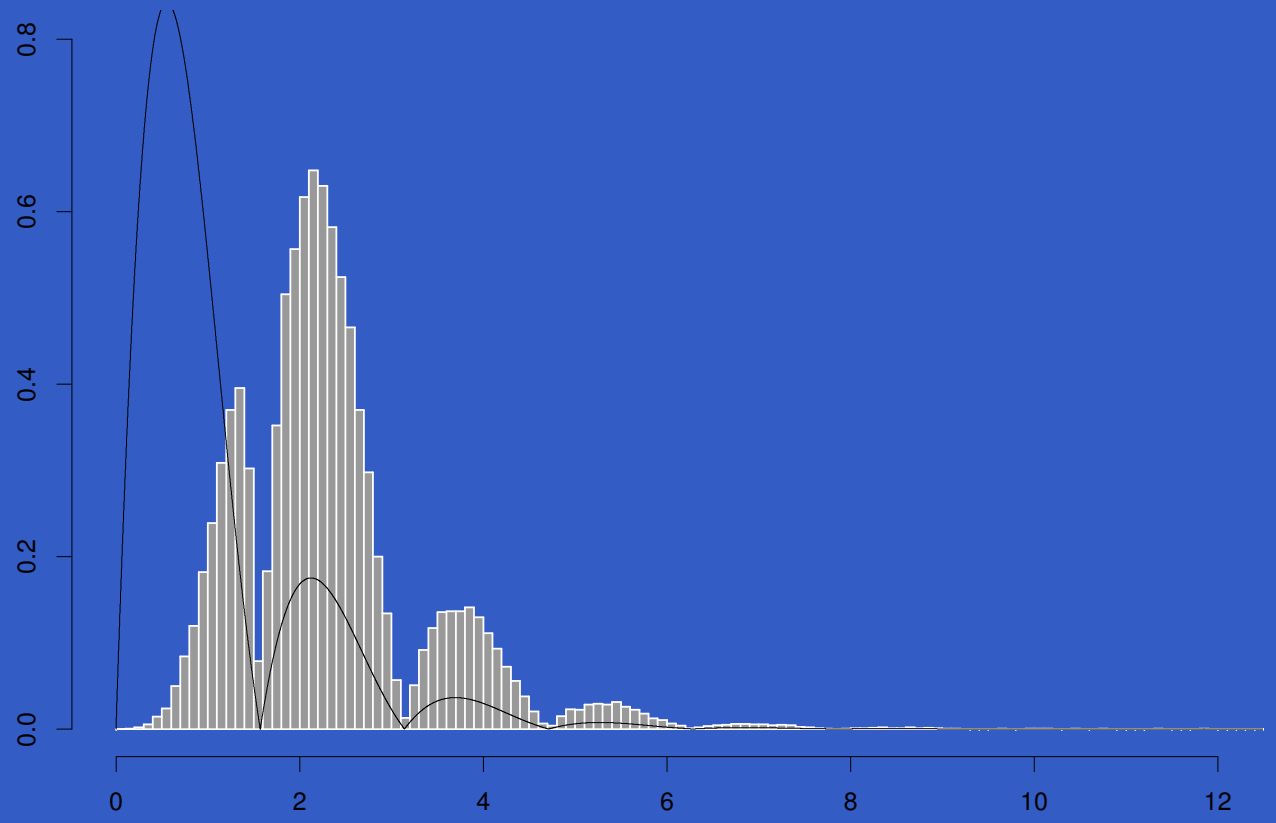
Variants on IMH

(3) Adaptive IMH (CORCORAN and SCHNEIDER, 2003)
Forward IMH with a self-targeting candidate.
Appears to converge very rapidly.

Variants on IMH

(3) Adaptive IMH (CORCORAN and SCHNEIDER, 2003)
Forward IMH with a self-targeting candidate.
Appears to converge very rapidly.

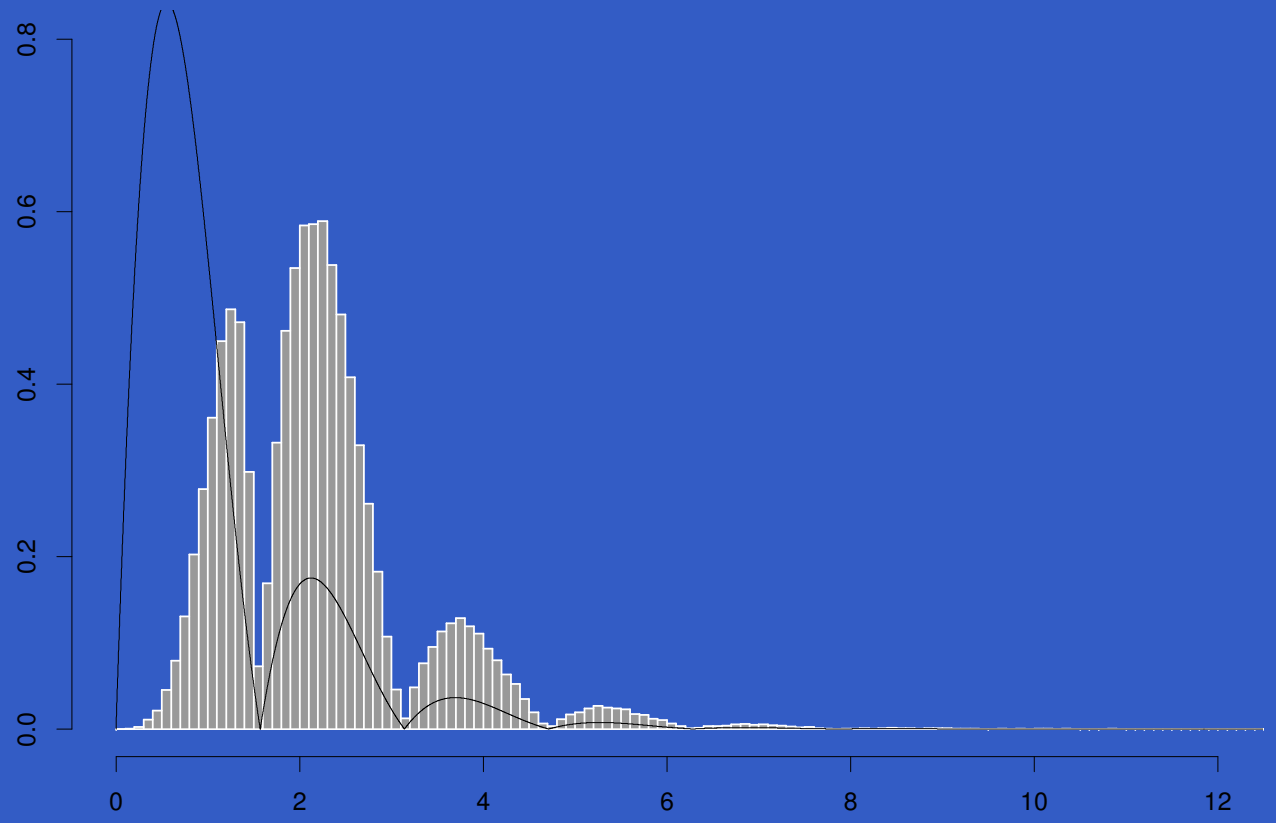
forward MH: 100,000 draws, 300 time steps



Variants on IMH

(3) Adaptive IMH (CORCORAN and SCHNEIDER, 2003)
Forward IMH with a self-targeting candidate.
Appears to converge very rapidly.

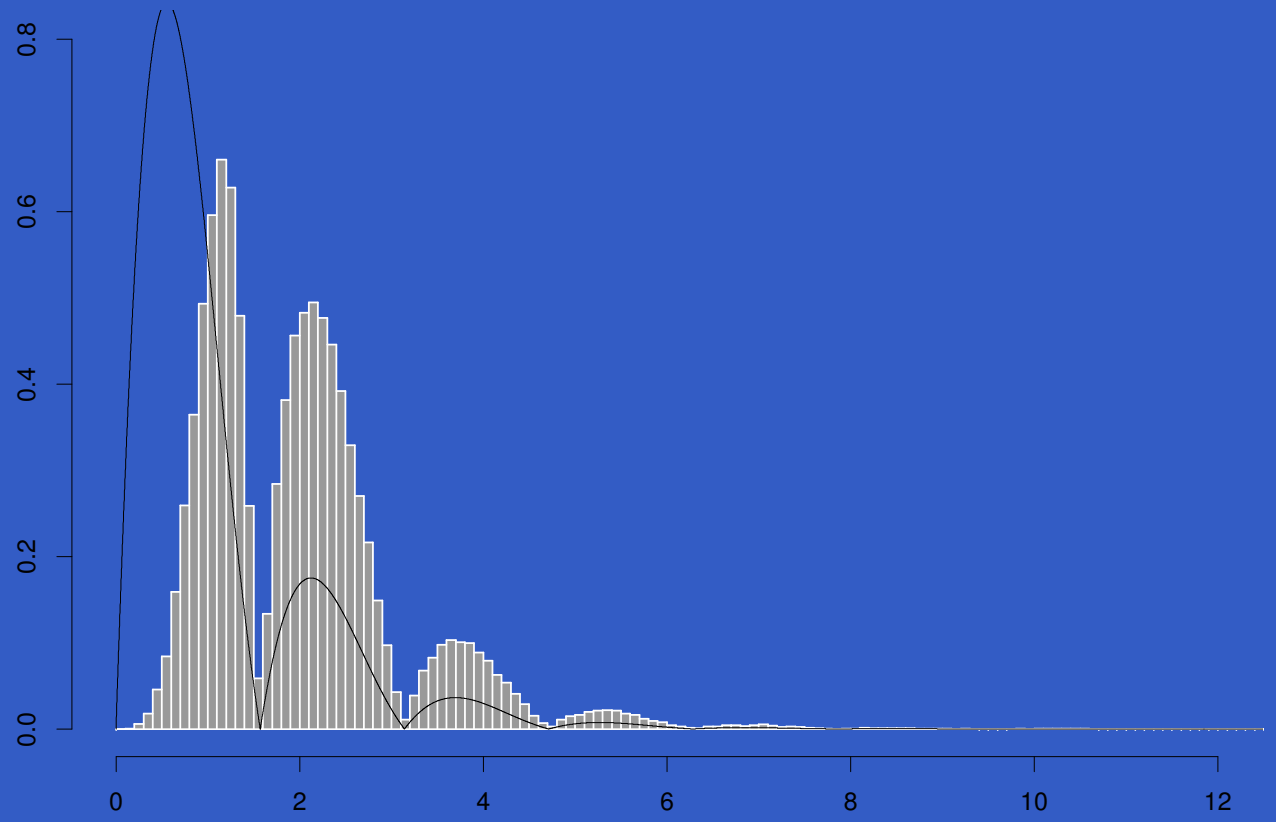
forward MH: 100,000 draws, 500 time steps



Variants on IMH

(3) Adaptive IMH (Corcoran and Schneider, 2003)
Forward IMH with a self-targeting candidate.
Appears to converge very rapidly.

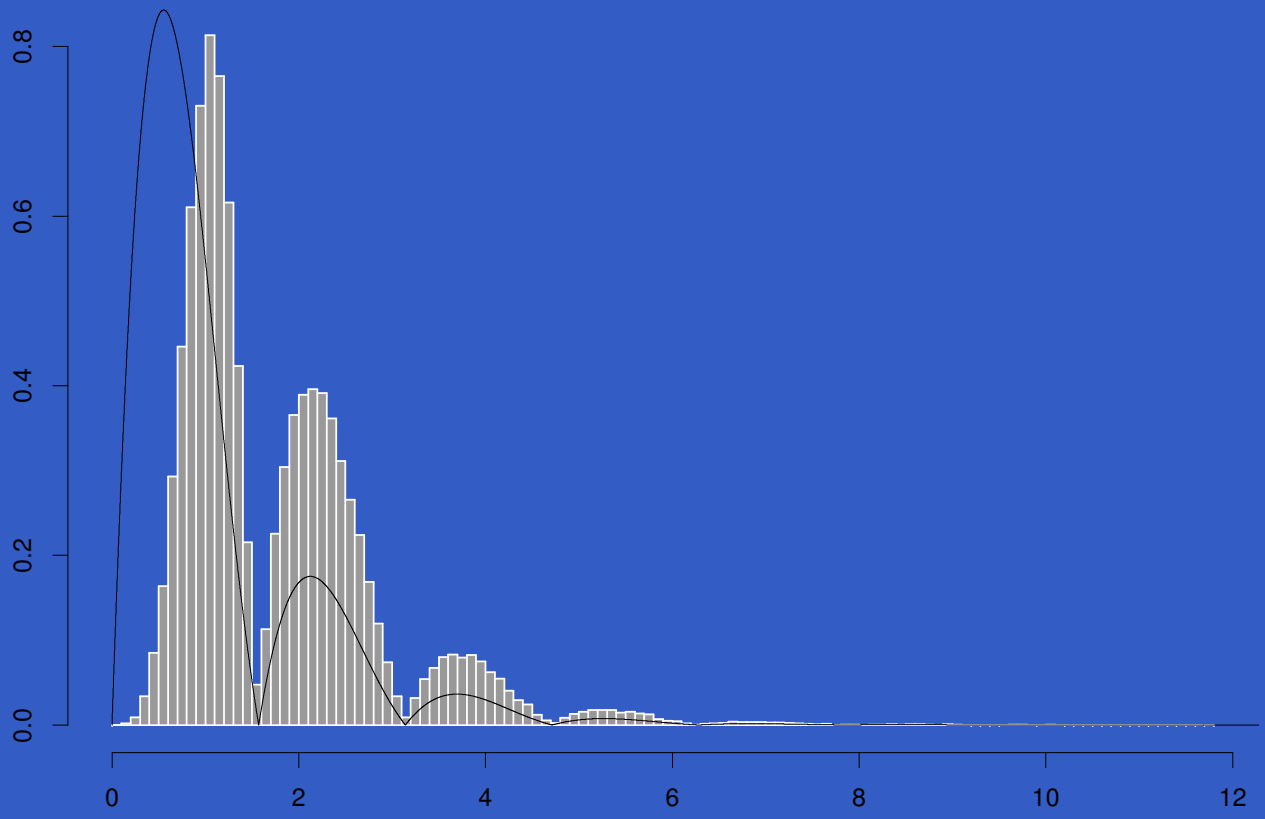
forward MH: 100,000 draws, 1000 time steps



Variants on IMH

(3) Adaptive IMH (CORCORAN and SCHNEIDER, 2003)
Forward IMH with a self-targeting candidate.
Appears to converge very rapidly.

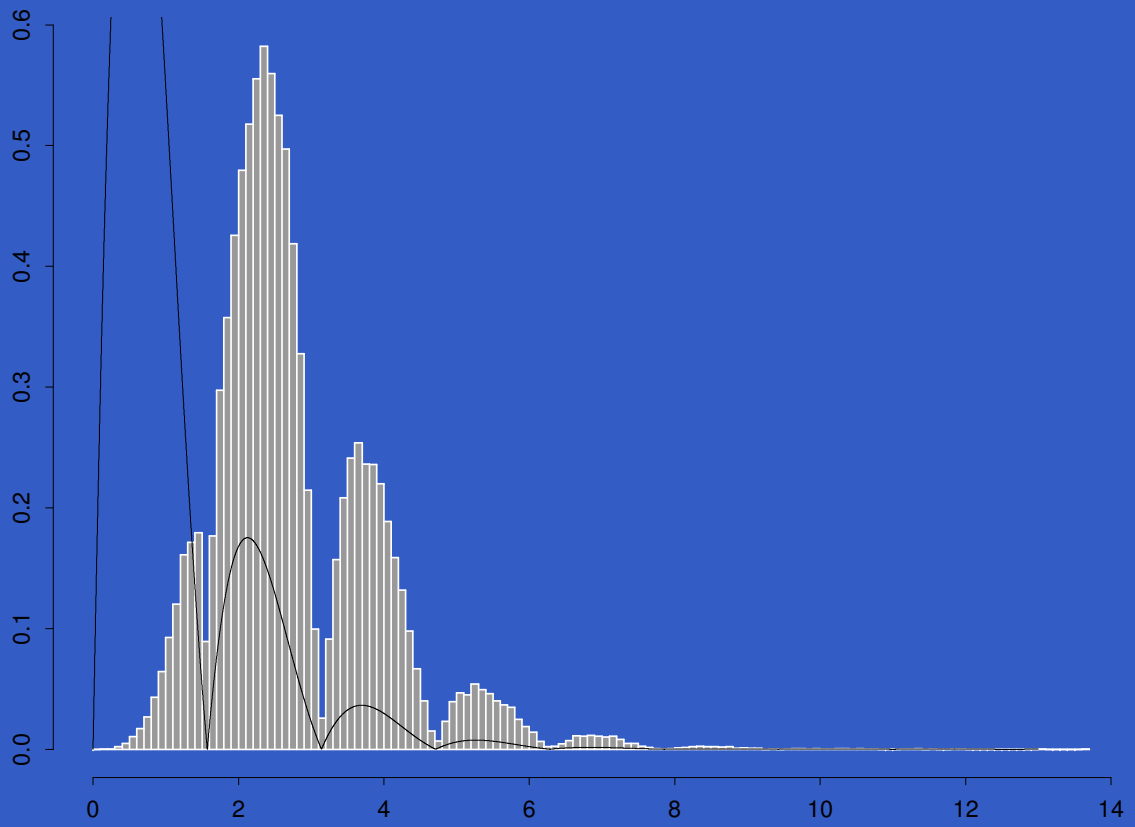
forward MH: 100,000 draws, 2000 time steps



Variants on IMH

(3) Adaptive IMH (CORCORAN and SCHNEIDER, 2003)
Forward IMH with a self-targeting candidate.
Appears to converge very rapidly.

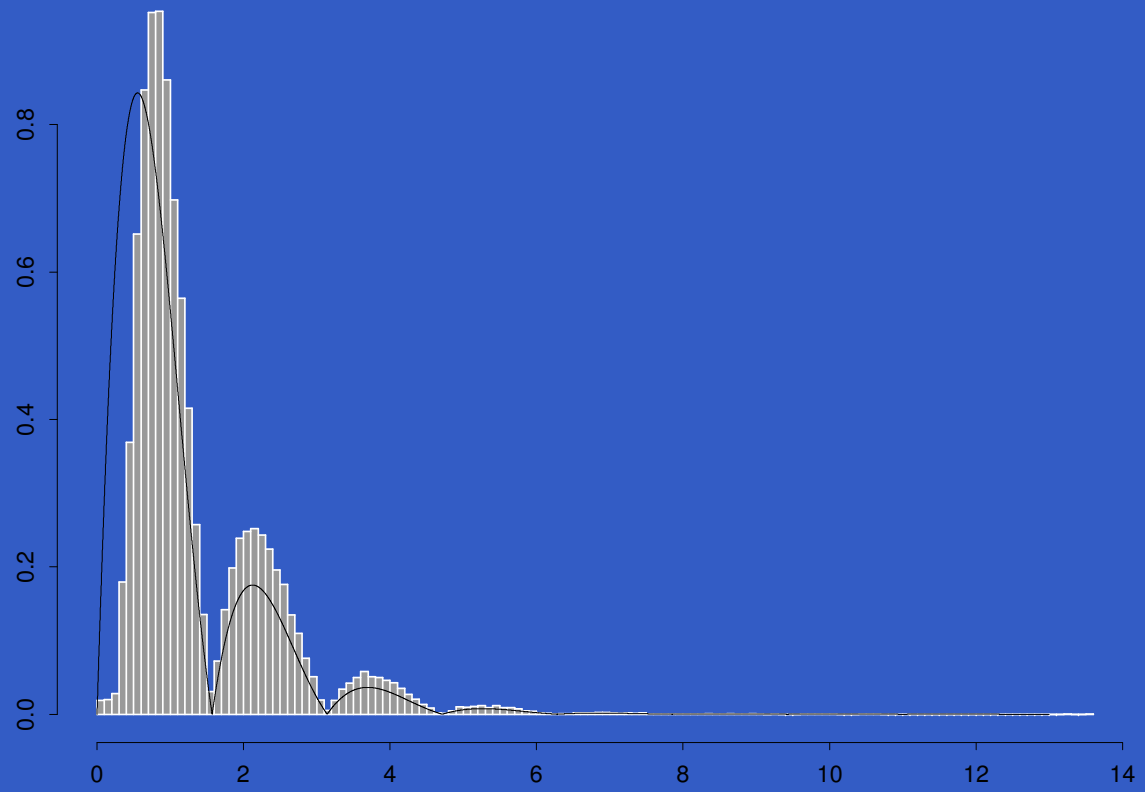
adaptive IMH: 100,000 draws, 100 time steps



Variants on IMH

(3) Adaptive IMH (CORCORAN and SCHNEIDER, 2003)
Forward IMH with a self-targeting candidate.
Appears to converge very rapidly.

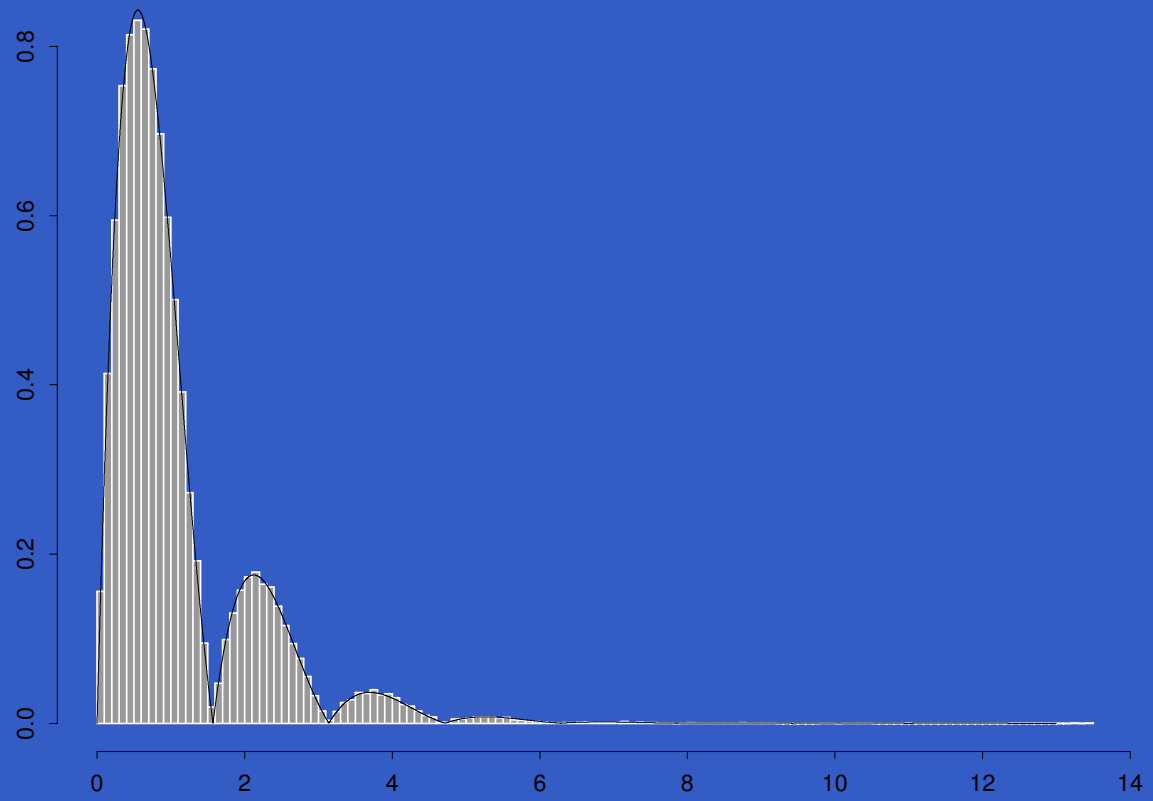
adaptive IMH: 100,000 draws, refinement 1



Variants on IMH

(3) Adaptive IMH (CORCORAN and SCHNEIDER, 2003)
Forward IMH with a self-targeting candidate.
Appears to converge very rapidly.

adaptive IMH: 100,000 draws, refinement 2



5. Bayesian Variable Selection

Linear regression model with Gaussian noise

$$\boldsymbol{y} = \gamma_1\theta_1\boldsymbol{x}_1 + \cdots + \gamma_q\theta_q\boldsymbol{x}_q + \boldsymbol{\varepsilon}$$

where:

- $\boldsymbol{x}_i \in \mathbb{R}^n, i = 1, \ldots, q$ predictors (fixed, known)
- $\theta_i \in \mathbb{R}, i = 1, \ldots, q$ coefficients (random)
- $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I})$ noise vector (random)
- $\gamma_i \in \{0, 1\}, i = 1, \ldots, q$ indicators (random)

The Goal

Given an observation y, choose the "best" subset of the predictors – which predictors were part of the model? Amounts to finding values of $\gamma = (\gamma_1, \ldots, \gamma_q)$!

The Goal

Given an observation y, choose the "best" subset of the predictors – which predictors were part of the model? Amounts to finding values of $\gamma = (\gamma_1, \ldots, \gamma_q)$!

Best? Bayesian perspective:

The Goal

Given an observation y, choose the "best" subset of the predictors – which predictors were part of the model? Amounts to finding values of $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)$!

Best? Bayesian perspective:

Select $\boldsymbol{\gamma}$ that appears most frequently when sampling from the posterior of the model.

The Goal

Given an observation y, choose the "best" subset of the predictors – which predictors were part of the model? Amounts to finding values of $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)$!

Best? Bayesian perspective:

Select $\boldsymbol{\gamma}$ that appears most frequently when sampling from the posterior of the model.

We need to be able to simulate from the posterior $\pi_{\boldsymbol{\Gamma},\sigma^2,\boldsymbol{\Theta}|\boldsymbol{Y}}$!

Usually, Bayesian approaches use regular MCMC methods – question of convergence!

The Model: Priors

Want to incorporate the following "standard normal gamma conjugate" priors:

$$\begin{aligned} \frac{\lambda\nu}{\sigma^2} &\sim \chi^2(\nu) \longrightarrow Z := \frac{1}{\sigma^2} \sim \Gamma(\frac{\nu}{2}, \frac{\lambda\nu}{2}) \\ \boldsymbol{\theta}|Z &\sim N(\boldsymbol{\xi}, \sigma^2 \boldsymbol{V}) \\ \gamma_i &\overset{i.i.d.}{\sim} \text{Bernoulli}(\frac{1}{2}),\ i = 1 \ldots, q \end{aligned}$$

The variance σ^2 for the $\boldsymbol{\varepsilon}$ and $\boldsymbol{\theta}$ is random.
$\boldsymbol{V}, \boldsymbol{\xi}, \lambda$, and ν are hyperparameters (fixed and known).

The Model: Posterior

Linear regression with Gaussian noise yields likelihood

$$L(\boldsymbol{\gamma}, \boldsymbol{\theta}, z) = z^{\frac{n}{2}} \exp\{-\frac{1}{2} z (\boldsymbol{y} - \sum_{i=1}^{q} \gamma_i \theta_i \boldsymbol{x}_i)^T (\boldsymbol{y} - \sum_{i=1}^{q} \gamma_i \theta_i \boldsymbol{x}_i)\}$$

Posterior – proportional to likelihood $\times$ priors:

$$\pi_{\boldsymbol{\Gamma}, Z, \boldsymbol{\Theta}|\boldsymbol{Y}}(\boldsymbol{\gamma}, z, \boldsymbol{\theta}|\boldsymbol{y}) \propto \; z^{\frac{n+q+\nu}{2}-1} \times$$

$$\exp\{-\frac{1}{2} z [(\boldsymbol{y} - \sum_{i=1}^{q} \gamma_i \theta_i \boldsymbol{x}_i)^T (\boldsymbol{y} - \sum_{i=1}^{q} \gamma_i \theta_i \boldsymbol{x}_i) + (\boldsymbol{\theta} - \boldsymbol{\xi})^T \boldsymbol{V}^{-1} (\boldsymbol{\theta} - \boldsymbol{\xi}) + \lambda\nu]\}$$

Increasing Layers of Complexity

- Fixed variance, fixed coefficients – sampling $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)$ using support set coupling within a Gibbs sampler (HUANG and DJURIC, 2002)
- Random variance, fixed coefficients – sampling $(\boldsymbol{\gamma}, z) = (\gamma_1, \ldots, \gamma_q, z)$ using slice coupling within a Gibbs sampler (SCHNEIDER and CORCORAN, 2002)
- Random variance, random coefficients – sampling $(\boldsymbol{\gamma}, z, \boldsymbol{\theta}) = (\gamma_1, \ldots, \gamma_q, z, \theta_1, \ldots, \theta_q)$ using bounded IMH (SCHNEIDER and CORCORAN, 2002)

The General Case

- Incorporate random variance and random coefficents.
- Reduce the size of the state space – define $\beta_i := \gamma_i \theta_i$ to have a mixture prior distribution.
- The values of γ can be recovered from $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ by setting:

$$\gamma_i = \begin{cases} 0 & \text{if } \beta_i = 0 \\ 1 & \text{if } \beta_i \neq 0 \end{cases} \qquad i = 1, \ldots, q.$$

The General Case – Using bounded IMH

Posterior

$$\pi_{\boldsymbol{B},Z|\boldsymbol{Y}}(\boldsymbol{\beta}, z|\boldsymbol{y}) \propto L(\boldsymbol{\beta}, z) g_{\boldsymbol{B}|Z}(\boldsymbol{\beta}|z) g_Z(z)$$

where $L(\boldsymbol{\beta}, z) = z^{\frac{n}{2}} \exp\{-\frac{1}{2} z (\boldsymbol{y} - \sum_{i=1}^{q} \beta_i \boldsymbol{x}_i)^T (\boldsymbol{y} - \sum_{i=1}^{q} \beta_i \boldsymbol{x}_i)\}$

- Choose candidate density

$$q(\boldsymbol{\beta}, z) \propto z^{\frac{n}{2}} g_{\boldsymbol{B}|Z}(\boldsymbol{\beta}|z) g_Z(z).$$

- Then

$$\frac{\pi}{q} = \frac{L(\boldsymbol{\beta}, z)}{z^{\frac{n}{2}}} = \exp\{-\frac{1}{2} z (\boldsymbol{y} - \sum_{i=1}^{q} \beta_i \boldsymbol{x}_i)^T (\boldsymbol{y} - \sum_{i=1}^{q} \beta_i \boldsymbol{x}_i)\} \leq 1$$

The General Case – Using bounded IMH

To get the candidate values $(z, \boldsymbol{\beta})$ according to

$$q(z, \boldsymbol{\beta}) \propto z^{\frac{n}{2}} g_{\boldsymbol{B}|Z}(\boldsymbol{\beta}|z) g_Z(z),$$

sample "hierarchically".

- Draw $Z \sim \Gamma(\frac{n+\nu}{2}, \frac{\lambda\nu}{2})$
- Draw $B|Z \sim N(\xi, \frac{1}{z}V)$
- Set $\beta_i = 0$ with probability $\frac{1}{2}$ $(i = 1 \ldots, q)$

Can now use bounded IMH to get exact draws from the posterior.

Some results

Hald data set $q = 4$, $n = 13$, y contains heat measurements of cement, predictor variables describe composition of the cement (aluminate, silicate, ferrite, dicalcium)

γ	percentage
(0,1,0,0)	69 %
(1,1,0,0)	14 %
(1,0,1,0)	13 %
(0,1,1,0)	3 %
(0,1,0,1)	1 %

component	percentage
$P(\gamma_1 = 1)$	27 %
$P(\gamma_2 = 1)$	87 %
$P(\gamma_3 = 1)$	16 %
$P(\gamma_4 = 1)$	1 %

6. Computing Self Energy

Compute self energy for the interacting fermion problem.

- Create and destroy a particle on a lattice of atoms (such as a crystal).
- Particle interacts with other electrons, “wake” of energy created around the movement of the particle.
- Quantify this self energy with the help of Feynman diagrams.

Approximate the sum using Monte Carlo methods and perfect sampling (Corcoran, Schneider and Schüttler, 2003)

$$\sigma(k) = \sum_{n=1}^{n_{max}} \sum_{g \in \mathcal{G}_n} \left(\frac{-T}{N} \right)^n \sum_{k_1, \ldots, k_n \in \mathcal{K}} F_g^{(n)}(k, k_1, \ldots, k_n).$$

7. Future Research

- Address large backward coupling times of the IMH algorithm (Bayesian variable selection) – multistage coupling?
- Find analytical error bounds and employ convergence diagnostics for the approximate and adaptive IMH algorithm.

More to come ...

References

[1] J. Corcoran and U. Schneider. Variants on the independent Metropolis-Hastings algorithm: approximate and adaptive IMH. *In preparation*, 2003.

[2] J.N. Corcoran and U. Schneider. Shift and scale coupling methods for perfect simulation. *Prob. Eng. Inf. Sci.*, 17:277–303, 2003.

[3] J.N. Corcoran, U. Schneider, and H.-B. Schüttler. Perfect stochastic summation for high order Feynman diagrams. *Submitted for publication*, 2003. Preprint at `http://amath.colorado.edu/student/schneidu/feynman.html`.

[4] Y. Huang and P. Djurić. Variable selection by perfect sampling. *EURASIP J. Appl. Si. Pr.*, 1:38–45, 2002.

[5] A.E. Raferty, D. Madigan, and J.A. Hoeting. Bayesian model averaging for linear regression models. *J. Am. Stat. Ass.*, 92:179–191, 1997.

[6] U. Schneider and J.N. Corcoran. Perfect sampling for Bayesian variable selection in a linear regression model. *J. Stat. Plann. Inference*, 2003. To appear. Preprint at `http://www.cgd.ucar.edu/stats/staff/uli/modelselect.html`.