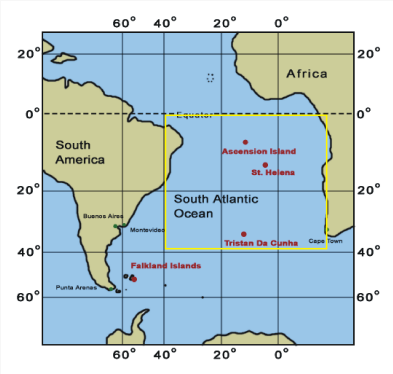
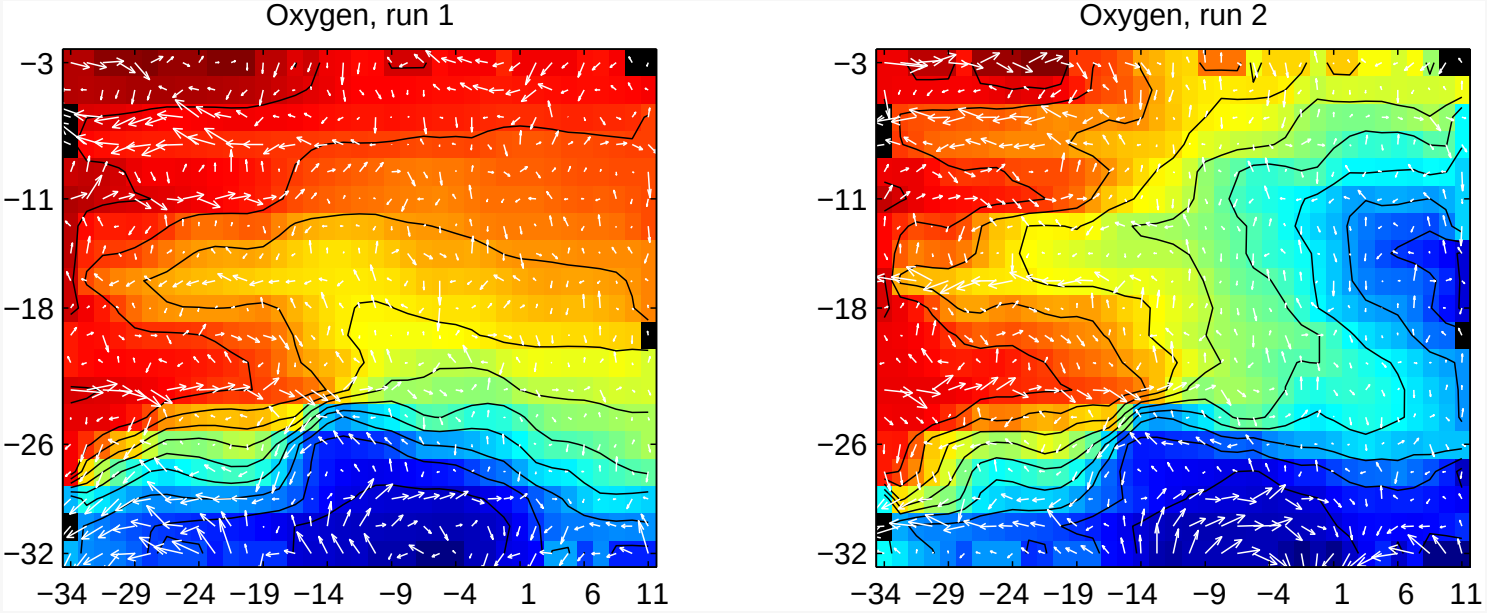
Polynomial acceleration of Gibbs sampling (sampling using lessons from CSE)

Colin Fox, Al Parker
fox@physics.otago.ac.nz

Ocean circulation :: 2 samples from the posterior



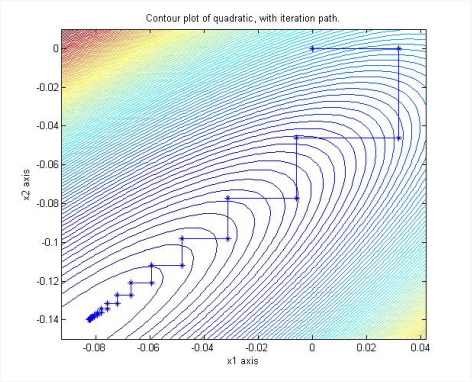
Data on traces, assert physics and observational models, infer abyssal advection



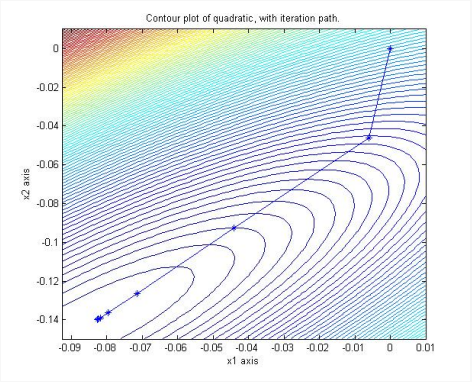
McKeague Nicholls Speer Herbei 2005 *Statistical Inversion of South Atlantic Circulation in an Abyssal Neutral Density Layer*

Adapting computational linear algebra to sampling

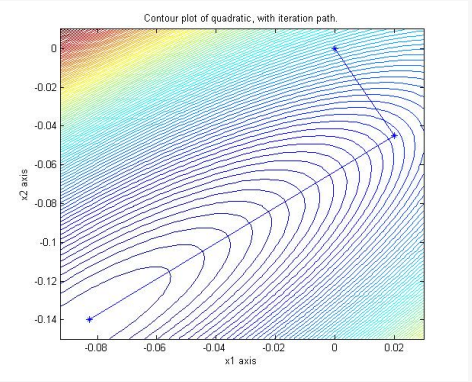
Optimization ...



Gauss-Seidel

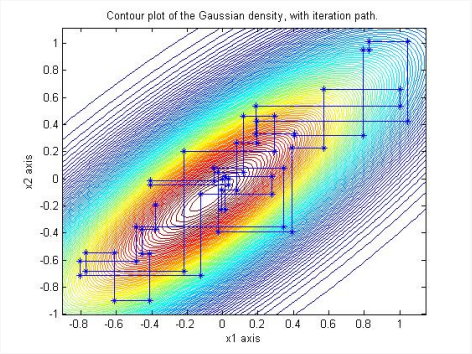


Cheby-GS

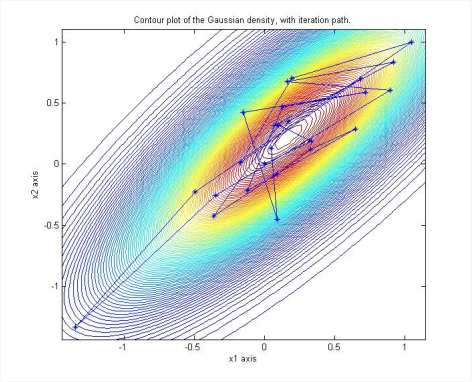


CG/Lanczos

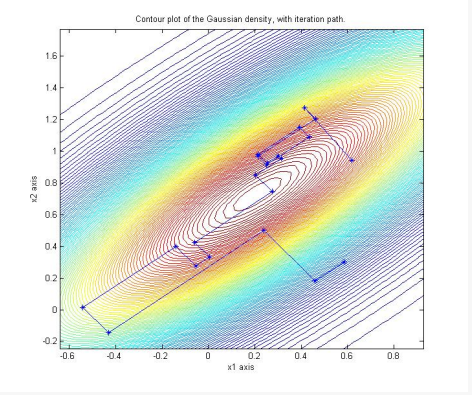
Sampling ...



Gibbs



Cheby-Gibbs



Lanczos

Normal distributions, quadratic forms, linear systems

We want to sample from Gaussian density with *precision matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$, SPD, $\dim(\mathbf{x}) = n$

$$\pi(\mathbf{x}) = \sqrt{\frac{\det(\mathbf{A})}{2\pi^n}} \exp\left\{-\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} + \mathbf{b}^\mathsf{T}\mathbf{x}\right\}$$

Covariance matrix is $\Sigma = \mathbf{A}^{-1}$ is also SPD.

Write $x \sim \mathrm{N}(\mu, \mathbf{A}^{-1})$ where mean is

$$\begin{aligned}\mu &= \arg\min_{\mathbf{x}} \left\{\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{b}^\mathsf{T}\mathbf{x}\right\} \\ &= \mathbf{x}^* : \mathbf{A}\mathbf{x}^* = \mathbf{b}\end{aligned}$$

Particularly interested in case where $\mathbf{A}$ is sparse (GMRF) and n large

Matrix formulation of Gibbs sampling from $\mathrm{N}(0, \mathbf{A}^{-1})$

Let $\mathbf{y} = (y_1, y_2, ..., y_n)^T$

Component-wise Gibbs updates each component in sequence from the (normal) conditional distributions.

One 'sweep' over all n components can be written

$$\mathbf{y}^{(k+1)} = -\mathbf{D}^{-1}\mathbf{L}\mathbf{y}^{(k+1)} - \mathbf{D}^{-1}\mathbf{L}^T\mathbf{y}^{(k)} + \mathbf{D}^{-1/2}\mathbf{z}^{(k)}$$

where: $\mathbf{D} = \mathrm{diag}(\mathbf{A})$, $\mathbf{L}$ is the strictly lower triangular part of $\mathbf{A}$, $\mathbf{z}^{(k-1)} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$

$$\mathbf{y}^{(k+1)} = \mathbf{G}\mathbf{y}^{(k)} + \mathbf{c}^{(k)}$$

$\mathbf{c}^{(k)}$ is iid 'noise' with zero mean, finite covariance

(stochastic AR(1) process = first order stationary iteration plus noise)

Goodman & Sokal, 1989

Matrix splitting form of stationary iterative methods

The *splitting* $\mathbf{A} = \mathbf{M} - \mathbf{N}$ converts linear system $\mathbf{Ax} = \mathbf{b}$ to $\mathbf{Mx} = \mathbf{Nx} + \mathbf{b}$.

If $\mathbf{M}$ is nonsingular

$$\mathbf{x} = \mathbf{M}^{-1}\mathbf{Nx} + \mathbf{M}^{-1}\mathbf{b}.$$

Iterative methods compute successively better approximations by

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{M}^{-1}\mathbf{N}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b} \\ &= \mathbf{G}\mathbf{x}^{(k)} + \mathbf{g}\end{aligned}$$

Many splittings use terms in $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$. Gauss-Seidel sets $\mathbf{M} = \mathbf{L} + \mathbf{D}$

$$\mathbf{x}^{(k+1)} = -\mathbf{D}^{-1}\mathbf{L}\mathbf{x}^{(k+1)} - \mathbf{D}^{-1}\mathbf{L}^{\mathsf{T}}\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}$$

Matrix splitting form of stationary iterative methods

The *splitting* $\mathbf{A} = \mathbf{M} - \mathbf{N}$ converts linear system $\mathbf{Ax} = \mathbf{b}$ to $\mathbf{Mx} = \mathbf{Nx} + \mathbf{b}$.

If $\mathbf{M}$ is nonsingular

$$\mathbf{x} = \mathbf{M}^{-1}\mathbf{N}\mathbf{x} + \mathbf{M}^{-1}\mathbf{b}.$$

Iterative methods compute successively better approximations by

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{M}^{-1}\mathbf{N}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b} \\ &= \mathbf{G}\mathbf{x}^{(k)} + \mathbf{g}\end{aligned}$$

Many splittings use terms in $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$. Gauss-Seidel sets $\mathbf{M} = \mathbf{L} + \mathbf{D}$

$$\mathbf{x}^{(k+1)} = -\mathbf{D}^{-1}\mathbf{L}\mathbf{x}^{(k+1)} - \mathbf{D}^{-1}\mathbf{L}^{\mathsf{T}}\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}$$

spot the similarity to Gibbs

$$\mathbf{y}^{(k+1)} = -\mathbf{D}^{-1}\mathbf{L}\mathbf{y}^{(k+1)} - \mathbf{D}^{-1}\mathbf{L}^{T}\mathbf{y}^{(k)} + \mathbf{D}^{-1/2}\mathbf{z}^{(k)}$$

Goodman & Sokal 1989; Amit & Grenander 1991

Gibbs converges $\iff$ solver converges

Theorem 1 *Let* $\mathbf{A} = \mathbf{M} - \mathbf{N}$, $\mathbf{M}$ *invertible. The stationary linear solver*

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{M}^{-1}\mathbf{N}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b} \\ &= \mathbf{G}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b}\end{aligned}$$

converges, if and only if the random iteration

$$\begin{aligned}\mathbf{y}^{(k+1)} &= \mathbf{M}^{-1}\mathbf{N}\mathbf{y}^{(k)} + \mathbf{M}^{-1}\mathbf{c}^{(k)} \\ &= \mathbf{G}\mathbf{y}^{(k)} + \mathbf{M}^{-1}\mathbf{c}^{(k)}\end{aligned}$$

converges in distribution. Here $\mathbf{c}^{(k)} \overset{iid}{\sim} \pi$ *has zero mean and finite variance.*

Proof. Both converge iff $\varrho(\mathbf{G}) < 1$. $\square$

Convergent splittings generate convergent (generalized) Gibbs samplers

Mean converges with asymptotic convergence factor $\varrho(\mathbf{G})$, covariance with $\varrho(\mathbf{G})^2$

Young 1971 Thm 3-5.1, Duflo 1997 Thm 2.3.18-4, Goodman & Sokal, 1989, Galli & Gao 2001 Parker F 2011

Some not so common Gibbs samplers for $\mathrm{N}(0, \mathbf{A}^{-1})$

splitting/sampler	$\mathbf{M}$	$\mathbf{Var}\left(\mathbf{c}^{(k)}\right) = \mathbf{M}^T + \mathbf{N}$	converge if
Richardson	$\frac{1}{\omega}\mathbf{I}$	$\frac{2}{\omega}\mathbf{I} - \mathbf{A}$	$0 < \omega < \frac{2}{\varrho(\mathbf{A})}$
Jacobi	$\mathbf{D}$	$2\mathbf{D} - \mathbf{A}$	$\mathbf{A}$ SDD
GS/Gibbs	$\mathbf{D} + \mathbf{L}$	$\mathbf{D}$	always
SOR/B&F	$\frac{1}{\omega}\mathbf{D} + \mathbf{L}$	$\frac{2-\omega}{\omega}\mathbf{D}$	$0 < \omega < 2$
SSOR/REGS	$\frac{\omega}{2-\omega}\mathbf{M}_{\mathsf{SOR}}\mathbf{D}^{-1}\mathbf{M}_{\mathsf{SOR}}^{\mathsf{T}}$	$\frac{\omega}{2-\omega}\left(\mathbf{M}_{\mathsf{SOR}}\mathbf{D}^{-1}\mathbf{M}_{\mathsf{SOR}}^{T} + \mathbf{N}_{\mathsf{SOR}}^{T}\mathbf{D}^{-1}\mathbf{N}_{\mathsf{SOR}}\right)$	$0 < \omega < 2$

Want: convenient to solve $\mathbf{Mu} = \mathbf{r}$ and sample from $\mathrm{N}(0, \mathbf{M}^T + \mathbf{N})$

Relaxation parameter ω can accelerate Gibbs.

SSOR is a forwards and backwards sweep of SOR to give a *symmetric* splitting

SOR: Adler 1981; Barone & Frigessi 1990, Amit & Grenander 1991, SSOR: Roberts & Sahu 1997

A closer look at convergence

To sample from $\mathrm{N}(\mu, \mathbf{A}^{-1})$ where $\mathbf{A}\mu = \mathbf{b}$

Split $\mathbf{A} = \mathbf{M} - \mathbf{N}$, $\mathbf{M}$ invertible. $\mathbf{G} = \mathbf{M}^{-1}\mathbf{N}$, and $\mathbf{c}^{(k)} \overset{\text{iid}}{\sim} \mathrm{N}(0, \mathbf{M}^T + \mathbf{N})$

The iteration

$$\mathbf{y}^{(k+1)} = \mathbf{G}\mathbf{y}^{(k)} + \mathbf{M}^{-1}\left((\mathbf{c}^{(k)} + \mathbf{b}\right)$$

implies

$$\mathrm{E}\left(\mathbf{y}^{(m)}\right) - \mu = \mathbf{G}^m\left[\mathrm{E}\left(\mathbf{y}^{(0)}\right) - \mu\right]$$

and

$$\mathrm{Var}\left(\mathbf{y}^{(m)}\right) - \mathbf{A}^{-1} = \mathbf{G}^m\left[\mathrm{Var}\left(\mathbf{y}^{(0)}\right) - \mathbf{A}^{-1}\right]\mathbf{G}^m$$

(Hence asymptotic average convergence factors $\varrho(\mathbf{G})$ and $\varrho(\mathbf{G})^2$)

Errors go down as the polynomial

$$P_m\left(\mathbf{I} - \mathbf{G}\right) = \left(\mathbf{I} - \left(\mathbf{I} - \mathbf{G}\right)\right)^m = \left(\mathbf{I} - \mathbf{M}^{-1}\mathbf{A}\right)^m$$

$$P_m(\lambda) = (1 - \lambda)^m$$

note $P_m(0) = 1$

A closer look at convergence

To sample from $\mathrm{N}(\mu, \mathbf{A}^{-1})$ where $\mathbf{A}\mu = \mathbf{b}$

Split $\mathbf{A} = \mathbf{M} - \mathbf{N}$, $\mathbf{M}$ invertible. $\mathbf{G} = \mathbf{M}^{-1}\mathbf{N}$, and $\mathbf{c}^{(k)} \overset{\text{iid}}{\sim} \mathrm{N}(0, \mathbf{M}^T + \mathbf{N})$

The iteration

$$\mathbf{y}^{(k+1)} = \mathbf{G}\mathbf{y}^{(k)} + \mathbf{M}^{-1}\left((\mathbf{c}^{(k)} + \mathbf{b}\right)$$

implies

$$\mathrm{E}\left(\mathbf{y}^{(m)}\right) - \mu = \mathbf{G}^m\left[\mathrm{E}\left(\mathbf{y}^{(0)}\right) - \mu\right]$$

and

$$\mathrm{Var}\left(\mathbf{y}^{(m)}\right) - \mathbf{A}^{-1} = \mathbf{G}^m\left[\mathrm{Var}\left(\mathbf{y}^{(0)}\right) - \mathbf{A}^{-1}\right]\mathbf{G}^m$$

(Hence asymptotic average convergence factors $\varrho(\mathbf{G})$ and $\varrho(\mathbf{G})^2$)

Errors go down as the polynomial

$$P_m\left(\mathbf{I} - \mathbf{G}\right) = \left(\mathbf{I} - \left(\mathbf{I} - \mathbf{G}\right)\right)^m = \left(\mathbf{I} - \mathbf{M}^{-1}\mathbf{A}\right)^m$$

$$P_m(\lambda) = (1 - \lambda)^m$$

note $P_m(0) = 1$

can we do better?

Controlling the error polynomial

Consider the splitting

$$\mathbf{A} = \frac{1}{\tau}\mathbf{M} + \left(1 - \frac{1}{\tau}\right)\mathbf{M} - \mathbf{N}$$

giving the iteration operator

$$\mathbf{G}_\tau = \left(\mathbf{I} - \tau\mathbf{M}^{-1}\mathbf{A}\right)$$

and error polynomial $P_m(\lambda) = (1 - \tau\lambda)^m$.

Taking the *sequence* of parameters $\tau_1, \tau_2, \ldots, \tau_m$ gives the error polynomial

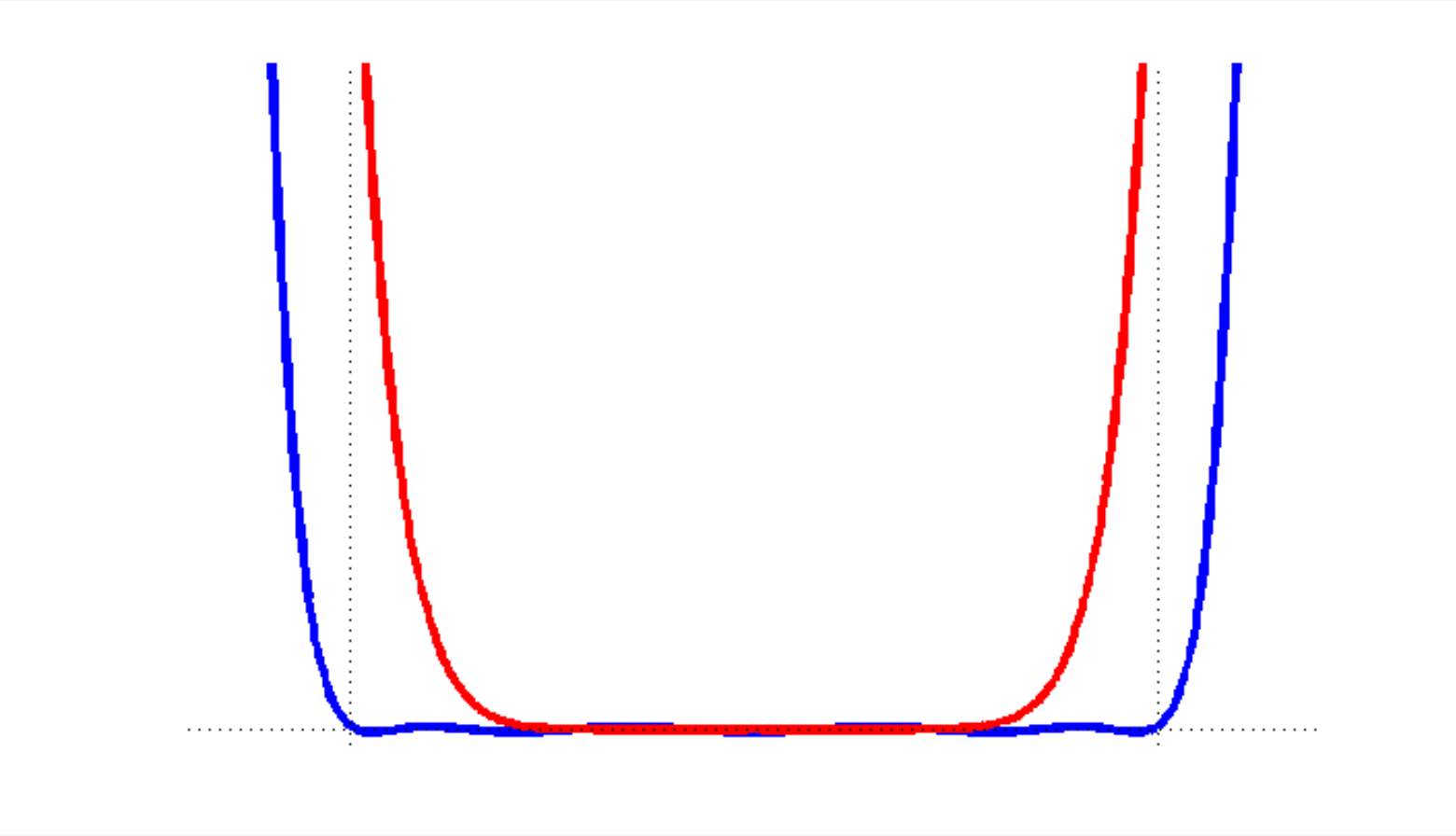
$$P_m(\lambda) = \prod_{l=1}^{m} (1 - \tau_l \lambda)$$

... we can choose the zeros of P_m !

Equivalently, can post-process chain by taking linear combination of states.

Golub & Varga 1961, Golub & van Loan 1989, Axelsson 1996, Saad 2003, Parker & F 2011

The best (Chebyshev) polynomial



Choose

$$\frac{1}{\tau_l} = \frac{\lambda_n + \lambda_1}{2} + \frac{\lambda_n - \lambda_1}{2} \cos\left(\pi \frac{2l+1}{2p}\right) \qquad l = 0, 1, 2, \ldots, p-1$$

where λ_1 λ_n are extreme eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$.

Second-order accelerated sampler

First-order accelerated iteration turns out to be unstable (iteration operators can have spectral radius $\gg 1$)

Numerical stability, and optimality at each step, is given by the second-order iteration

$$\mathbf{y}^{(k+1)} = (1-\alpha_k)\mathbf{y}^{(k-1)} + \alpha_k \mathbf{y}^{(k)} + \alpha_k \tau_k \mathbf{M}^{-1}(\mathbf{c}^{(k)} - \mathbf{A}\mathbf{y}^{(k)})$$

with α_k and τ_k chosen so error polynomial satisfies Chebyshev recursion.

Theorem 2 *Solver converges $\Rightarrow$ sampler converges (given correct noise distribution) Error polynomial is optimal for both mean and covariance.*

Asymptotic average reduction factor (Axelsson 1996) is

$$\sigma = \frac{1-\sqrt{\lambda_1/\lambda_n}}{1+\sqrt{\lambda_1/\lambda_n}}$$

F & Parker 2011

Algorithm 1: Chebyshev accelerated SSOR sampling from $\mathrm{N}(\mathbf{0}, \mathbf{A}^{-1})$

input : The SSOR splitting $\mathbf{M}$, $\mathbf{N}$ of $\mathbf{A}$; smallest eigenvalue $\lambda_{\min}$ of $\mathbf{M}^{-1}\mathbf{A}$; largest eigenvalue $\lambda_{\max}$ of $\mathbf{M}^{-1}\mathbf{A}$; relaxation parameter ω; initial state $\mathbf{y}^{(0)}$; $k_{\max}$

output: $\mathbf{y}^{(k)}$ approximately distributed as $N(\mathbf{0}, \mathbf{A}^{-1})$

set $\gamma = \left(\frac{2}{\omega} - 1\right)^{1/2}$, $\delta = \left(\frac{\lambda_{\max} - \lambda_{\min}}{4}\right)^2$, $\theta = \frac{\lambda_{\max} + \lambda_{\min}}{2}$;

set $\alpha = 1$, $\beta = 2/\theta$, $\tau = 1/\theta$, $\tau_c = \frac{2}{\tau} - 1$;

for $k = 1, \ldots, k_{\max}$ **do**

  if $k = 0$ **then**

    $b = \frac{2}{\alpha} - 1, \quad a = \tau_c b, \quad \kappa = \tau;$

  else

    $b = 2(1-\alpha)/\beta + 1, \quad a = \tau_c + (b-1)\left(1/\tau + 1/\kappa - 1\right), \quad \kappa = \beta + (1-\alpha)\kappa;$

  end

  sample $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$;

  $\mathbf{c} = \gamma b^{1/2} \mathbf{D}^{1/2} \mathbf{z}$;

  $\mathbf{x} = \mathbf{y}^{(k)} + \mathbf{M}^{-1}(\mathbf{c} - \mathbf{A}\mathbf{y}^{(k)})$;

  sample $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$;

  $\mathbf{c} = \gamma a^{1/2} \mathbf{D}^{1/2} \mathbf{z}$;

  $\mathbf{w} = \mathbf{x} - \mathbf{y}^{(k)} + \mathbf{M}^{-T}(\mathbf{c} - \mathbf{A}\mathbf{x})$;

  $\mathbf{y}^{(k+1)} = \alpha(\mathbf{y}^{(k)} - \mathbf{y}^{(k-1)} + \tau\mathbf{w}) + \mathbf{y}^{(k-1)}$;

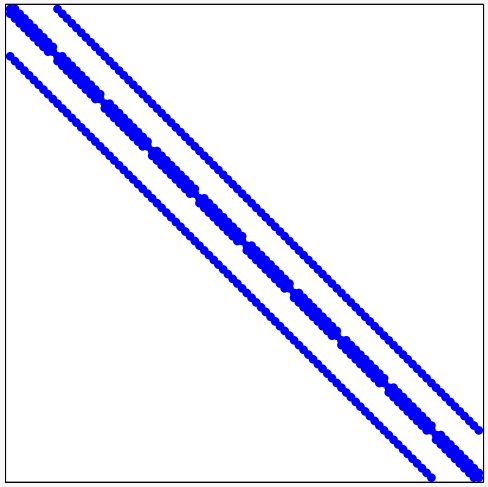
  $\beta = (\theta - \beta\delta)^{-1}$;

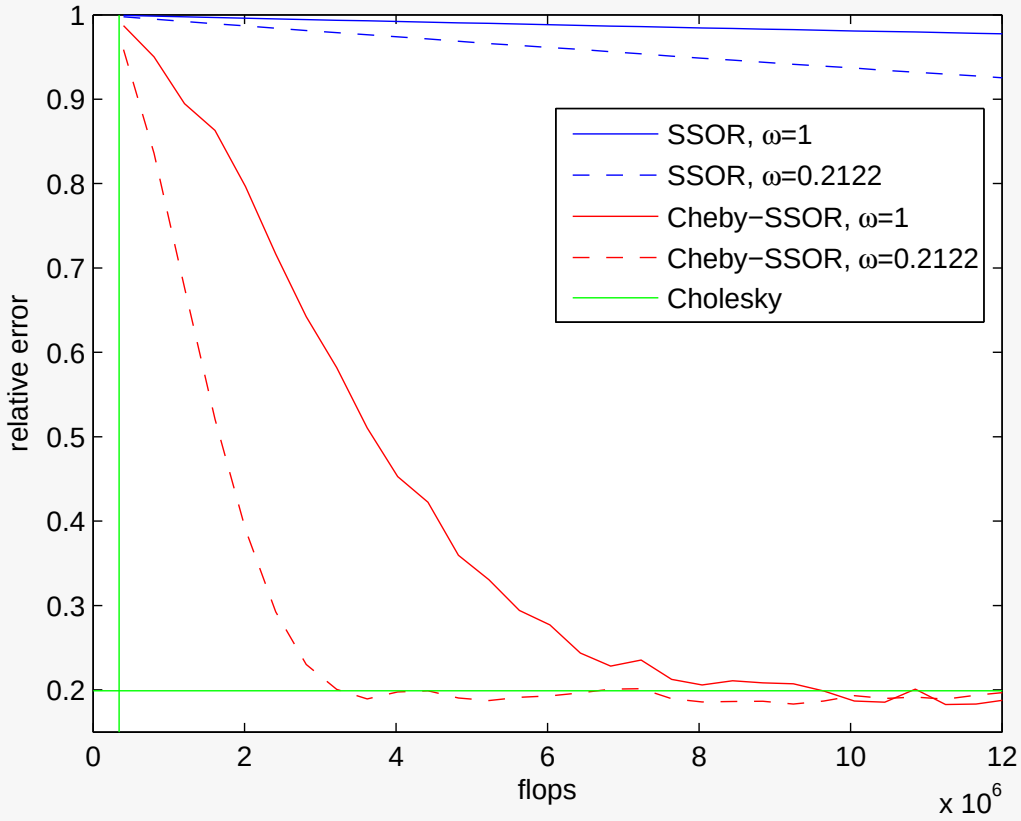
  $\alpha = \theta\beta$;

end

10×10 lattice ($n = 100$) sparse precision matrix

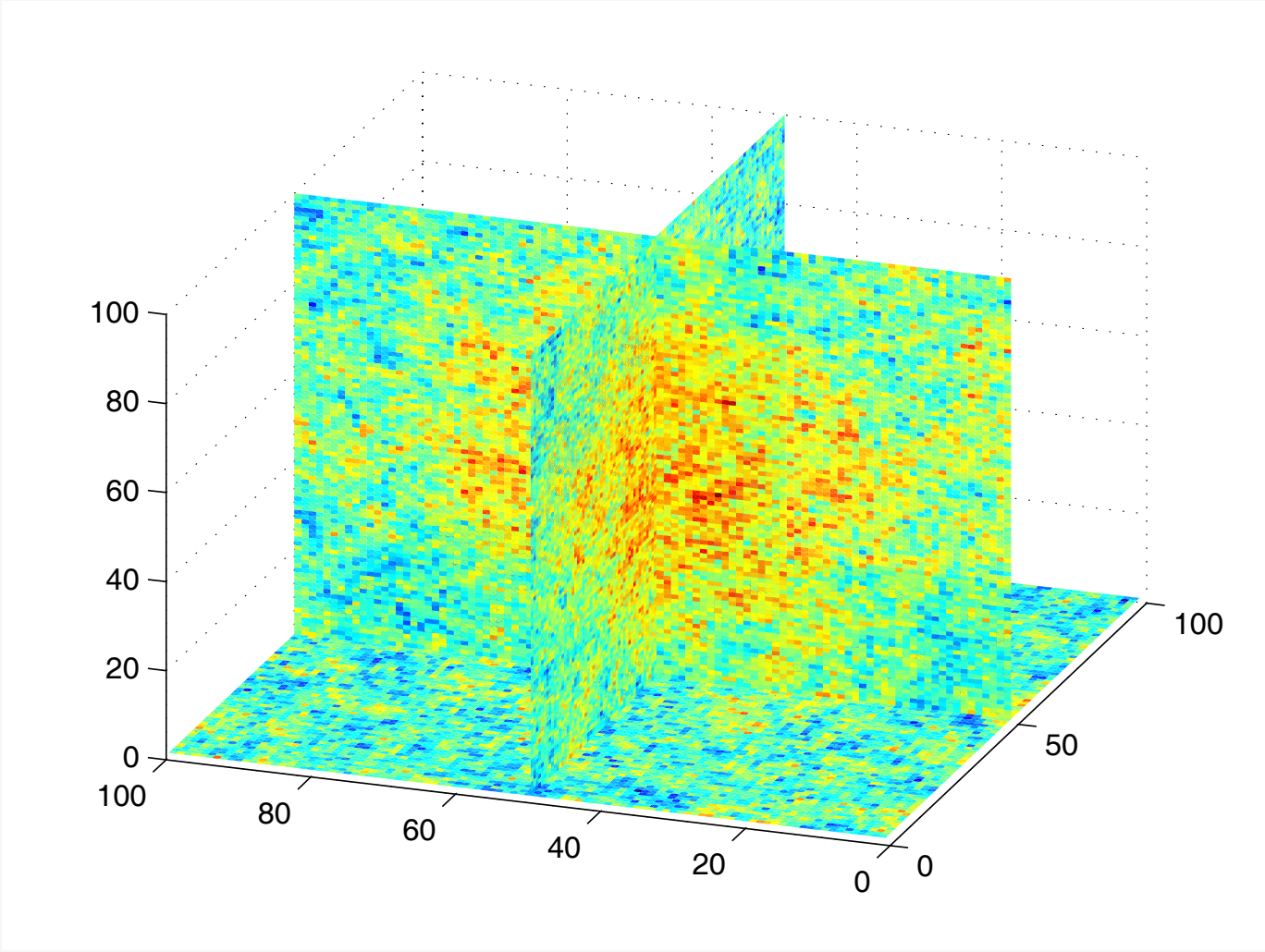
$$[\mathbf{A}]_{ij} = 10^{-4}\delta_{ij} + \begin{cases} n_i & \text{if} \quad i = j \\ -1 & \text{if} \quad i \neq j \quad \text{and} \quad ||s_i - s_j||_2 \leq 1 \\ 0 & \text{otherwise} \end{cases} .$$





$\approx 10^4$ times faster

$100 \times 100 \times 100$ lattice ($n = 10^6$) sparse precision matrix



only used sparsity, no other special structure

Some observations

In the Gaussian setting

- *stochastic relaxation* is fundamentally equivalent to classical *relaxation*
- If you can solve it then you can sample it
- ... with the *same computational cost*
- A sequence of (sub-optimal) kernels can outperform repeated application of the optimal kernel

more generally

- acceleration of convergence in mean and covariance is not limited to Gaussian targets
- ... but is unlikely to hold for densities without special structure
- Convergence also follows for bounded perturbation of a Gaussian (Amit 1991 1996)
- ... but no results for convergence rate

MCM'beach-BBQ-surf