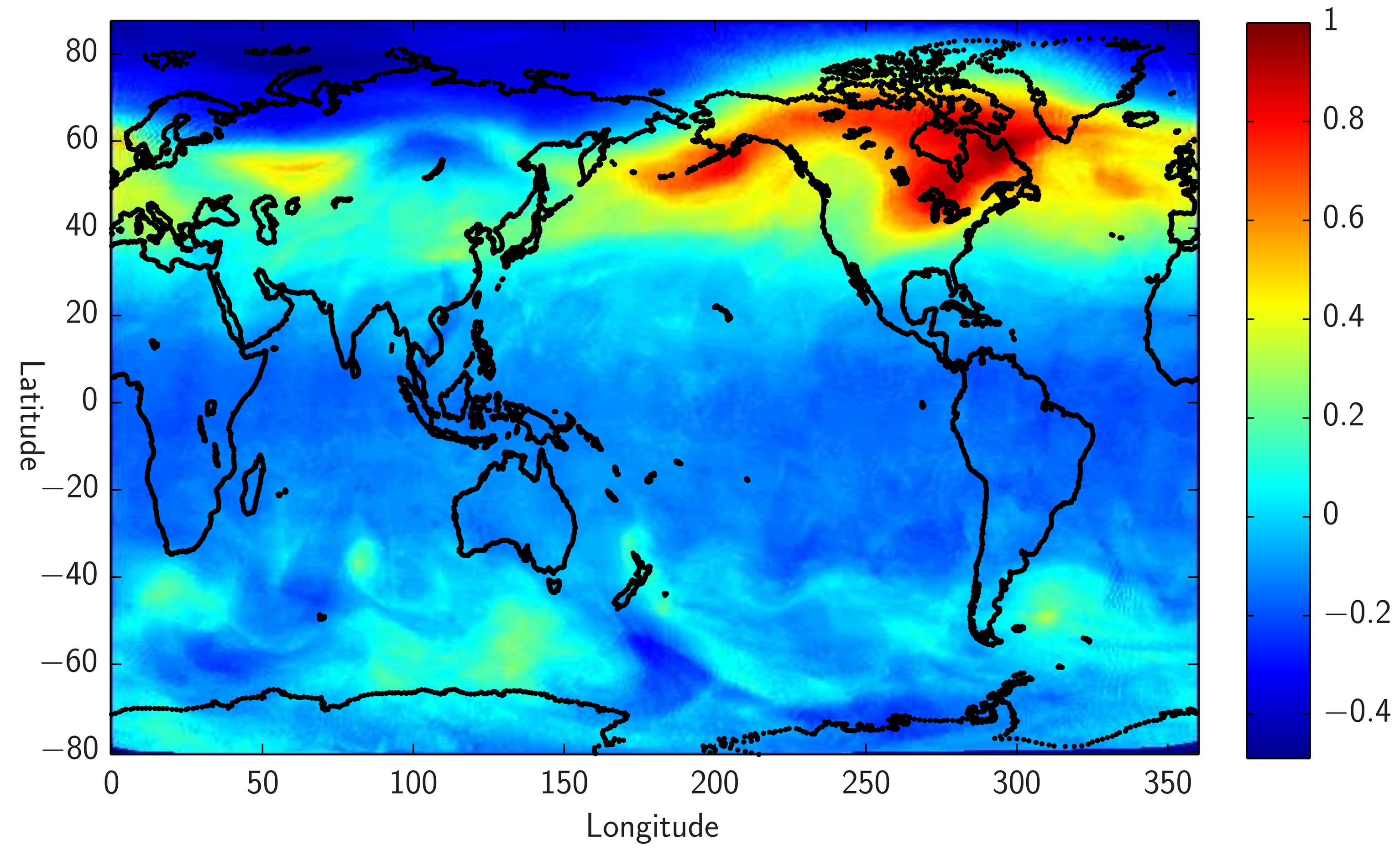
Playing Russian Roulette with Large-Scale GMRF

Heiko Strathmann[1,2], Daniel Simpson[3], and Mark Girolami[1]

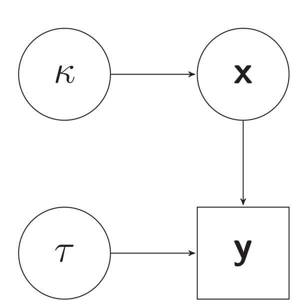
Department of Statistical Science[1] & Gatsby Computational Neuroscience Unit[2], University College London. Norwegian University of Science and Technology[3]

A Large-Scale Dataset: Total Column Ozone Data



- Popular dataset. We follow modelling approach in [2].
- Large-Scale: n=173,405 orbiting satellite measurements.
- Example model: stationary model using approximate Matérn SPDE on a fixed triangulation of the globe.
- Note: full analysis would require modelling observation process and uncertainty of field.

Our Example: The Common Latent Gaussian Model



$\tau \sim \log_2 \mathcal{N}(0, 100)$
$\kappa \sim \log_2 \mathcal{N}(0, 100)$
$\mathbf{x}|\kappa \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(\kappa)^{-1})$
$\mathbf{y}|\mathbf{x}, \tau \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \tau^{-1}\mathbf{I})$
$\mathbf{x} \in \mathbb{R}^m$ – latent field
$\mathbf{y} \in \mathbb{R}^n$ – observations

Challenging dimensions — though sparse:

- $n = 173,405$ $m = 196,002$
- $\mathbf{A} \in \mathbb{R}^{n \times m}$ – piecewise linear basis
- $\mathbf{Q}(\kappa)^{-1} \in \mathbb{R}^{m \times m}$ – precision matrix SPDE
- κ controls range of correlation
- τ observation noise

Goal: Exact-Approximate Bayesian Inference for Parameters

We are interested in the posterior over the parameters

$$\pi(\kappa, \tau|\mathbf{y}) \propto \pi(\mathbf{y}|\kappa, \tau)\pi(\kappa)\pi(\tau),$$

where the log-marginal-likelihood $\pi(\mathbf{y}|\kappa, \tau)$ can be shown to be (using $\boldsymbol{\theta} := \{\kappa, \tau\}$),

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) := \log(\pi(\mathbf{y}|\kappa, \tau)) = \log(\det(\mathbf{Q}(\kappa))) + n\log(\tau) - \log(\det(\mathbf{Q}(\kappa) + \tau\mathbf{A}^T\mathbf{A})) \\ - \tau\mathbf{y}^T\mathbf{y} + \tau^2\mathbf{y}^T\mathbf{A}(\mathbf{Q}(\kappa) + \tau\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} + C.\end{aligned}$$

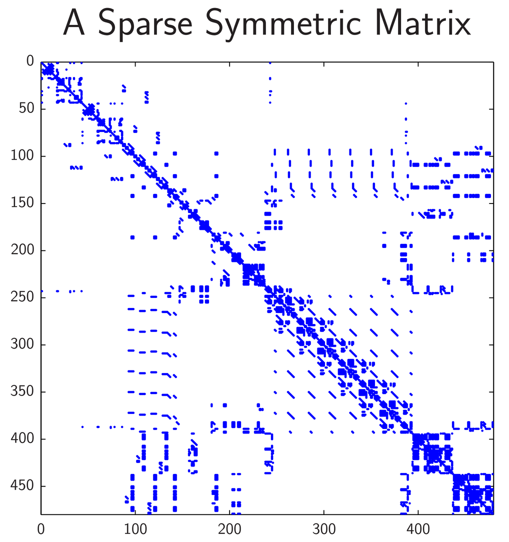
Infeasible Cholesky — Infeasible Log-Determinants – Infeasible Exact Inference?

Standard method for computing log-determinant of psd. matrix $\mathbf{Q}$:

1. Compute triangular Cholesky factor $\mathbf{L}$, such that $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$
2. Compute $\log(\det(\mathbf{Q})) = \log(\det(\mathbf{L}\mathbf{L}^T)) = \log((\prod_i L_{ii})(\prod_j L_{jj})) = 2\sum_i \log(L_{ii})$

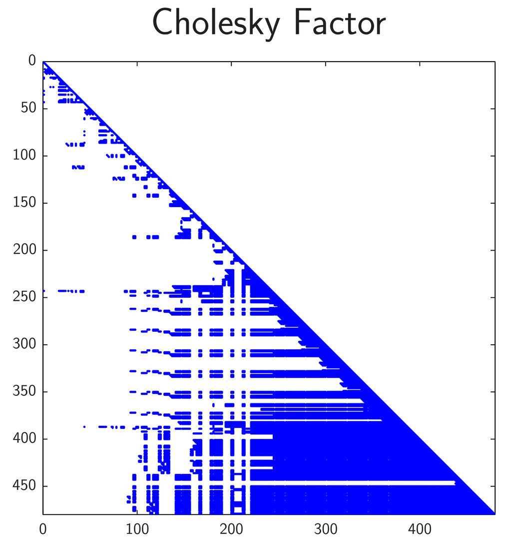
Does this work for Sparse matrices?

A Sparse Symmetric Matrix



Non-zeros: 3.29 percent

Cholesky Factor



Non-zeros: 13.23 percent

Problem: Fill-in effect limits dimensionality due to finite memory. Intractable Likelihood!

Suggested Approach: High-Level

1. Compute a Monte-Carlo estimate of $\log(\det(\mathbf{Q}))$ as in [2] and plug it into $\mathcal{L}(\boldsymbol{\theta}) = \log(\pi(\mathbf{y}|\kappa, \tau))$
2. Use Russian Roulette [3] to get an unbiased estimator of $\exp(\mathcal{L}(\boldsymbol{\theta}))$, i.e., $\hat{\pi}(\mathbf{y}|\boldsymbol{\theta})$
3. Use the Pseudo-Marginal-MCMC scheme [1] to sample from $\pi(\boldsymbol{\theta}|\mathbf{y})$, i.e., use the Metropolis Hastings acceptance probability

$$\min\left(1, \frac{\hat{\pi}(\mathbf{y}|\boldsymbol{\theta}^{\text{new}}) \times \pi(\boldsymbol{\theta}^{\text{new}}) \times q(\boldsymbol{\theta}^{\text{old}}|\boldsymbol{\theta}^{\text{new}})}{\hat{\pi}(\mathbf{y}|\boldsymbol{\theta}^{\text{old}}) \times \pi(\boldsymbol{\theta}^{\text{old}}) \times q(\boldsymbol{\theta}^{\text{new}}|\boldsymbol{\theta}^{\text{old}})}\right)$$

for some proposal q. $\Rightarrow$ Exact-Approximate Bayesian inference is possible!

Russian Roulette for the Exponential Function

Have: unbiased estimator $\widehat{\mathcal{L}(\theta)}$. Want: unbiased estimator for $\pi(\mathbf{y}|\boldsymbol{\theta}) = \exp(\mathcal{L}(\theta))$.

1. Infinite series representation of exponential

$$\exp(\mathcal{L}(\boldsymbol{\theta})) = 1 + \sum_{i=1}^{\infty}\prod_{j=1}^{i}\frac{\mathcal{L}(\boldsymbol{\theta})}{j} = 1 + \sum_{i=1}^{\infty}\frac{\mathcal{L}(\boldsymbol{\theta})^i}{i!} = 1 + \frac{\mathcal{L}(\boldsymbol{\theta})}{1} + \frac{\mathcal{L}(\boldsymbol{\theta})\mathcal{L}(\boldsymbol{\theta})}{2} + \frac{\mathcal{L}(\boldsymbol{\theta})\mathcal{L}(\boldsymbol{\theta})\mathcal{L}(\boldsymbol{\theta})}{6} + \dots$$

2. Unbiased estimate: Replace every $\mathcal{L}(\boldsymbol{\theta})$ by independent unbiased estimate $X^{(i)} \sim \widehat{\mathcal{L}(\theta)}$ for $i = 1, 2, \dots$

$$\exp(\mathcal{L}(\boldsymbol{\theta})) = 1 + \sum_{i=1}^{\infty}\prod_{j=1}^{i}\frac{X^{(j,i)}}{j} =: 1 + \sum_{i=1}^{\infty}\alpha_i.$$

3. Truncate the infinite sum $\sum_{i=1}^{\infty}\alpha_i$ unbiasedly via a set of decreasing stopping probabilities
   - Define threshold r and evaluate α_i until finding j such that $|\alpha_j| < r$
   - With some diminishing probability, for example $q_j := \frac{|\alpha_j|}{r} < 1$, continue evaluating with weight $\frac{1}{q_j}$
   - Truncated terms, on decision not to continue, is the desired unbiased estimator

Estimating Log-Determinants – Rational Approximations and Krylov-Methods

Approach suggested in [2], based on sparse matrix-vector products.

1. Compute Monte Carlo estimates of log-determinants as

$$\log(\det(\mathbf{Q})) = \text{tr}(\log(\mathbf{Q})) = \mathbb{E}_{\mathbf{s}}(\mathbf{s}^T\log(\mathbf{Q})\mathbf{s}) \approx \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i^T\log(\mathbf{Q})\mathbf{s}_i,$$

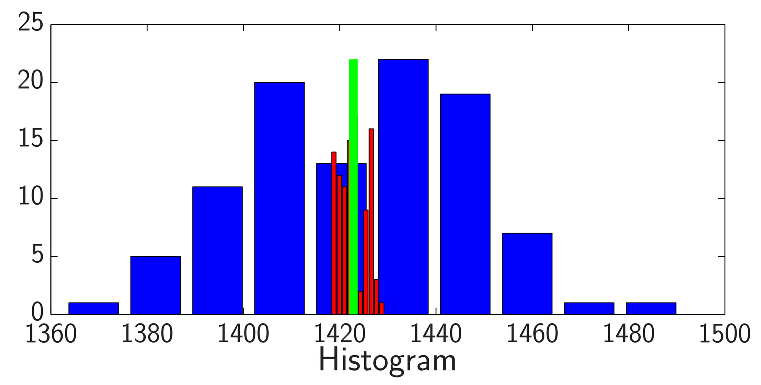
where the source-vectors $\mathbf{s}_i$ are random realisations with zero mean and unit variance, e.g., $\mathbf{s}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

2. Need $\log(\mathbf{Q})\mathbf{s}$. Use rational approximation of complex Cauchy contour integral (up to machine precision!).

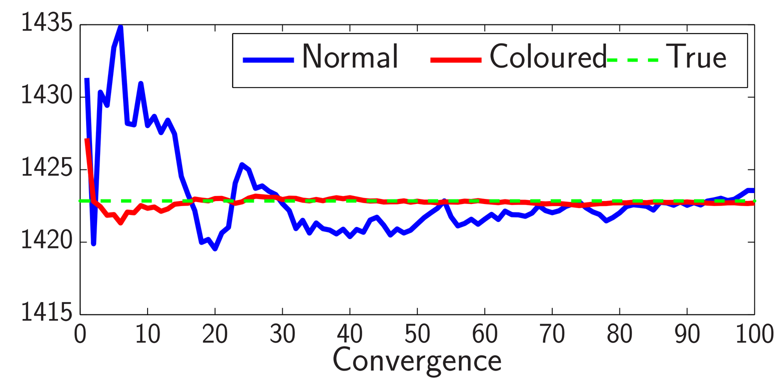
$$\log(\mathbf{Q})\mathbf{s} \approx \frac{1}{2\pi i}\sum_{i=1}^{M}\alpha_i(\mathbf{Q} - \sigma_i\mathbf{I})^{-1}\mathbf{s},$$

where $\boldsymbol{\alpha}, \boldsymbol{\sigma} \in \mathbb{C}^M$ are complex integration weights and shifts. Error bound for resolution M and accuracy ϵ: $M \propto \log\left(\frac{\lambda_{\max}(\mathbf{Q})}{\lambda_{\min}(\mathbf{Q})}\right)\log(\epsilon)$. Use Krylov-subspace methods to solve for $(\mathbf{Q} - \sigma_i I)^{-1}\mathbf{s}$.

3. Use graph-colourings to exploit sparsity structure of $\mathbf{Q}$ when generating source vectors $\mathbf{s}_i$.



Histogram

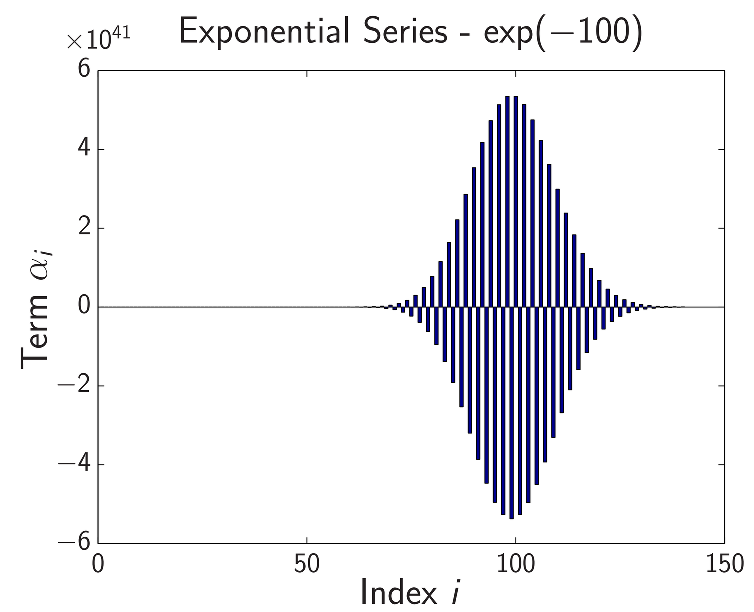


Convergence

Challenges in Russian Roulette – Stickiness and Computation Time

- Pseudo-Marginal MCMC is very sensible to variance in $\hat{\pi}(\mathbf{y}|\boldsymbol{\theta})$, which affects mixing.
- Exponential function's interesting places are at index of same order as argument (see plot)
- Need many Russian Roulette iterations, otherwise variance is catastrophic.

Extremely challenging to obtain good mixing!

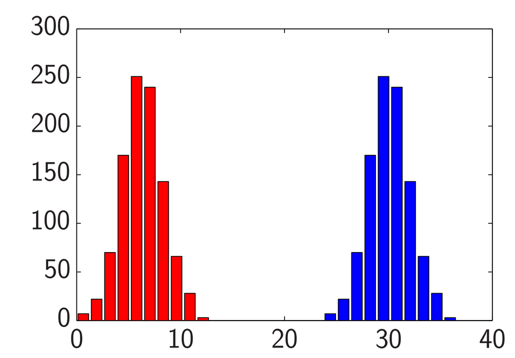


Approaching Feasibility: Reducing the Number of Estimates for Russian Roulette

1. Reduce absolute value. Find bound $\mathcal{U}$ such that $\mathcal{U} < X^{(i)} < 0$ and perform RR on

$$\exp(\mathcal{L}(\boldsymbol{\theta})) = \exp(\mathcal{U})\exp(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{U}),$$

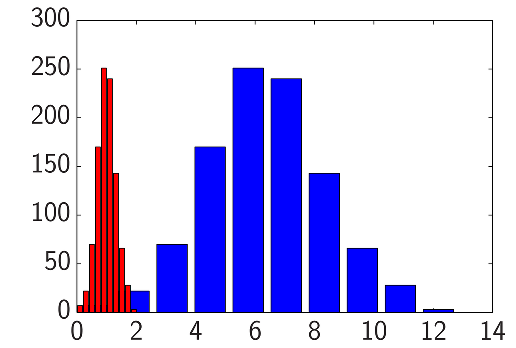
whose interesting parts are now closer to 0, i.e., need less RR iterations to reach.



2. Scale estimates. Find a positive integer $E \approx |\mathcal{L}(\boldsymbol{\theta}) - \mathcal{U}|$ and rescale

$$\exp(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{U}) = \left(\exp\left(\frac{\mathcal{L}(\boldsymbol{\theta}) - \mathcal{U}}{E}\right)\right)^E$$

Now need E RR estimates of $\exp\left(\frac{\mathcal{L}(\boldsymbol{\theta}) - \mathcal{U}}{E}\right) \approx \exp(-1)$, which is much better behaved.

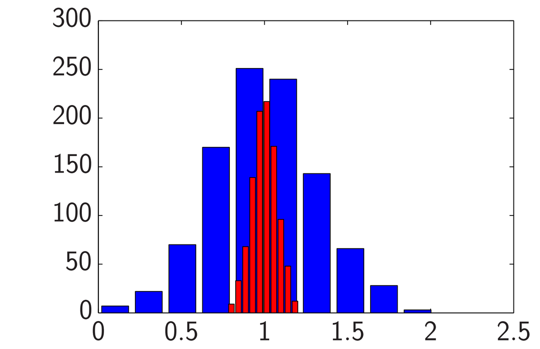


3. Average independent samples $X^{(i)}$ to reduce variance. For the practitioner, there is a faster alternative with controllable bias.

4. Given estimates $\{X^{(i)}\}_{j=1}^N$, select group size $d < N$ and create $\tilde{N}$ index sets $\mathcal{I}_i$ that contain d unique indices $j \in \{1, \dots, N\}$. Generate $\tilde{N}$ pseudo-independent estimates

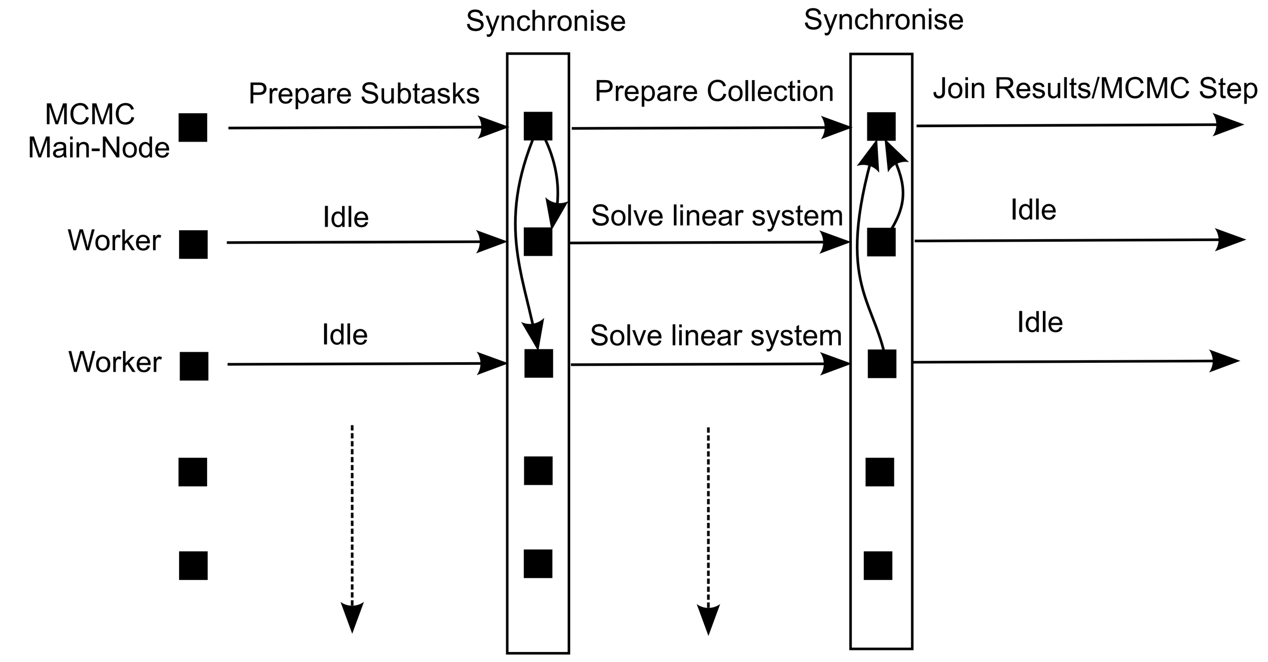
$$\tilde{X}^{(i)} = \frac{1}{d}\sum_{j \in \mathcal{I}_i} X^{(j)} \qquad (1 \le i \le \tilde{N}),$$

which have lower variance. Introduced dependence is effectively broken by permuting over different RR denominators $\prod_{n=1}^{\infty} n = 1 \cdot 2 \cdot 6 \cdot 24 \cdot \dots$ and *bias can be controlled!*

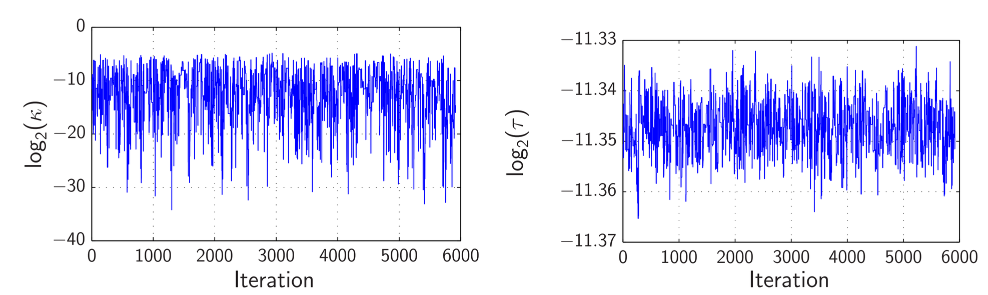


Approaching Feasibility: Distributed Russian Roulette

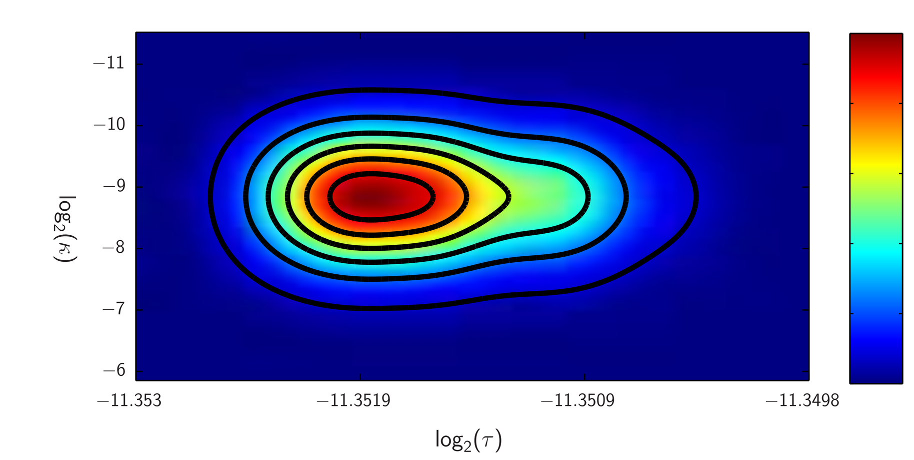
- Approximating $\mathbf{s}_i^T\log(\mathbf{Q})\mathbf{s}_i$ corresponds to solving $N \approx 20$ to 30 linear systems.
- Each estimate in $\sum_{i=1}^N \mathbf{s}_i^T\log(\mathbf{Q})\mathbf{s}_i$ is independent. $M \approx 20$ to 50.
- Averaging multiple estimates of $\mathcal{L}(\boldsymbol{\theta})$, which are independent. Needed ~ 50
- Given appropriate hardware, potential speed-up of factor ~ 75000. Exploit cluster computers.



Results – Current State – Work in Progress



Traces and kernel density estimate for $\boldsymbol{\theta} = \{\kappa, \tau\}$



Literature

[1] Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2):697–725, April 2009.

[2] Erlend Aune, DanielP. Simpson, and Jo Eidsvik. Parameter estimation in high dimensional gaussian distributions. *Statistics and Computing*, December 2012.

[3] Mark Girolami, Anne-Marie Lyne, Heiko Strathmann, Daniel Simpson, and Yves Atchade. Playing russian roulette with intractable likelihoods. Technical Report arXiv:1306.4032, June 2013.

M. Girolami is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) and a Royal Society Wolfson Research Merit Award. H. Strathmann is supported by the Gatsby Charitable Foundation.

Contact: heiko.strathmann@gmail.com