# Transport map-accelerated adaptive importance sampling

Simon Cotter

University of Manchester

13th November 2018

EPSRC

Engineering and Physical Sciences
Research Council

# Collaborators

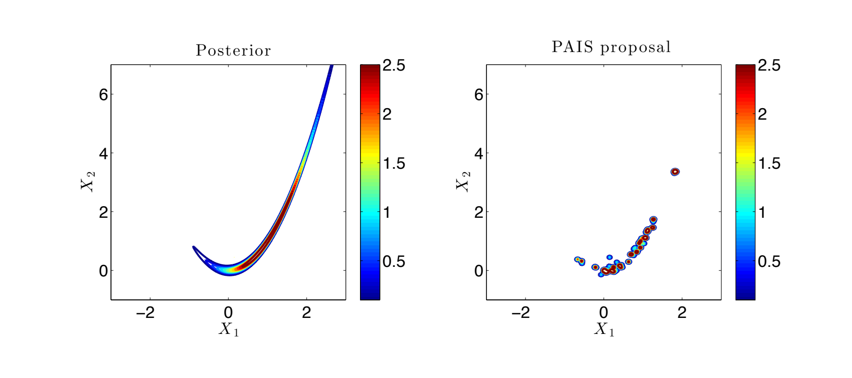
Left: Colin Cotter (Imperial College, UK), Centre: Yannis Kevrekidis (John Hopkins, US) Right: Paul Russell (University of Manchester, UK)

SLC is grateful to EPSRC for First Grant award EP/L023393/1

EPSRC
Engineering and Physical Sciences
Research Council

# Motivation

# Motivation

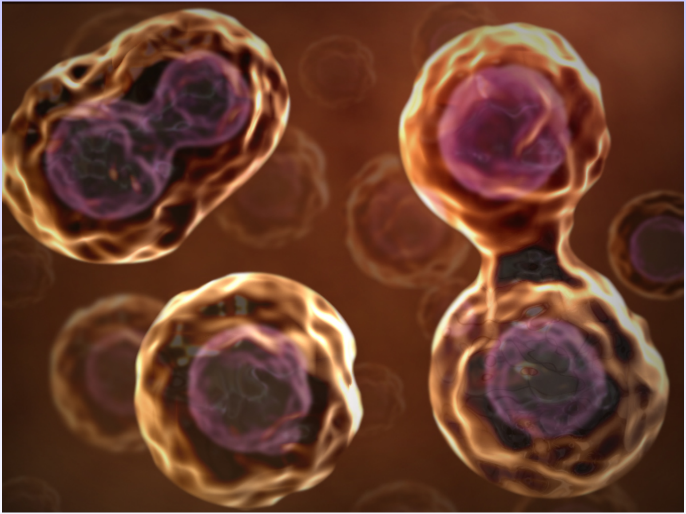


# Motivation

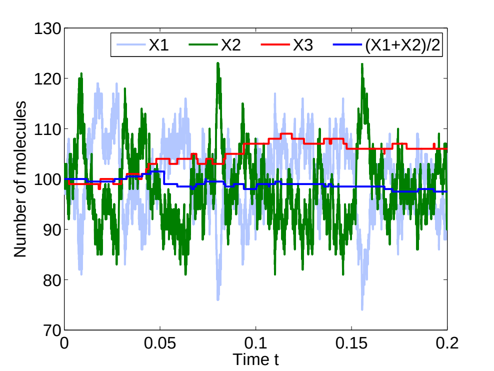
# Outline

1. Motivating example: multiscale stochastic reaction networks
2. Parallel Adaptive Importance Sampling
3. Transport maps
4. Numerical Results

# Motivation



# Multiscale Systems




$$\emptyset \xrightarrow{k_1} X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftarrows}} X_2 \xrightarrow{k_4 x_2} X_3 \xrightarrow{k_5 x_3} \emptyset$$

# Inverse problems for multiscale chemical reaction networks

- Not able to accurately observe the fast variables (POMP model)
- Subset of the reaction parameters will be unobservable
- Likelihood is invariant to moves along manifolds in parameter space
- Posterior distribution concentrated close to such a manifold
- Without knowledge of the manifold:
  - Metropolis-Hastings and other single-state algorithms perform poorly, proposing off the manifold frequently, slow mixing along manifold
  - Importance sampling schemes have poor proposal distributions
  - Slow convergence, or even instability (importance weight collapse)

# Inverse problems for multiscale chemical reaction networks

- Not able to accurately observe the fast variables (POMP model)
- Subset of the reaction parameters will be unobservable
- Likelihood is invariant to moves along manifolds in parameter space
- Posterior distribution concentrated close to such a manifold
- Without knowledge of the manifold:
  - Metropolis-Hastings and other single-state algorithms perform poorly, proposing off the manifold frequently, slow mixing along manifold
  - Importance sampling schemes have poor proposal distributions
  - Slow convergence, or even instability (importance weight collapse)

# Inverse problems for multiscale chemical reaction networks

- Not able to accurately observe the fast variables (POMP model)
- Subset of the reaction parameters will be unobservable
- Likelihood is invariant to moves along manifolds in parameter space
- Posterior distribution concentrated close to such a manifold
- Without knowledge of the manifold:
  - Metropolis-Hastings and other single-state algorithms perform poorly, proposing off the manifold frequently, slow mixing along manifold
  - Importance sampling schemes have poor proposal distributions
  - Slow convergence, or even instability (importance weight collapse)

# Inverse problems for multiscale chemical reaction networks

- Not able to accurately observe the fast variables (POMP model)
- Subset of the reaction parameters will be unobservable
- Likelihood is invariant to moves along manifolds in parameter space
- Posterior distribution concentrated close to such a manifold
- Without knowledge of the manifold:
  - Metropolis-Hastings and other single-state algorithms perform poorly, proposing off the manifold frequently, slow mixing along manifold
  - Importance sampling schemes have poor proposal distributions
  - Slow convergence, or even instability (importance weight collapse)

# Inverse problems for multiscale chemical reaction networks

- Not able to accurately observe the fast variables (POMP model)
- Subset of the reaction parameters will be unobservable
- Likelihood is invariant to moves along manifolds in parameter space
- Posterior distribution concentrated close to such a manifold
- Without knowledge of the manifold:
  - Metropolis-Hastings and other single-state algorithms perform poorly, proposing off the manifold frequently, slow mixing along manifold
  - Importance sampling schemes have poor proposal distributions
  - Slow convergence, or even instability (importance weight collapse)

# Inverse problems for multiscale chemical reaction networks

- Not able to accurately observe the fast variables (POMP model)
- Subset of the reaction parameters will be unobservable
- Likelihood is invariant to moves along manifolds in parameter space
- Posterior distribution concentrated close to such a manifold
- Without knowledge of the manifold:
  - Metropolis-Hastings and other single-state algorithms perform poorly, proposing off the manifold frequently, slow mixing along manifold
  - Importance sampling schemes have poor proposal distributions
  - Slow convergence, or even instability (importance weight collapse)

# Inverse problems for multiscale chemical reaction networks

- Not able to accurately observe the fast variables (POMP model)
- Subset of the reaction parameters will be unobservable
- Likelihood is invariant to moves along manifolds in parameter space
- Posterior distribution concentrated close to such a manifold
- Without knowledge of the manifold:
  - Metropolis-Hastings and other single-state algorithms perform poorly, proposing off the manifold frequently, slow mixing along manifold
  - Importance sampling schemes have poor proposal distributions
  - Slow convergence, or even instability (importance weight collapse)

# Inverse problems for multiscale chemical reaction networks

- Not able to accurately observe the fast variables (POMP model)
- Subset of the reaction parameters will be unobservable
- Likelihood is invariant to moves along manifolds in parameter space
- Posterior distribution concentrated close to such a manifold
- Without knowledge of the manifold:
  - Metropolis-Hastings and other single-state algorithms perform poorly, proposing off the manifold frequently, slow mixing along manifold
  - Importance sampling schemes have poor proposal distributions
  - Slow convergence, or even instability (importance weight collapse)

# QSSA: Simple Example

- Consider the system:

$$\emptyset \xrightarrow{k_1} X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftarrows}} X_2 \xrightarrow{k_4 x_2} \emptyset$$

- Effective system:

$$\emptyset \xrightarrow{k_1} S \xrightarrow{\hat{k}_4 s} \emptyset$$

- Fast subsystem: $k_1, k_4 \to 0$

$$X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftarrows}} X_2, \qquad S = X_1 + X_2$$

# QSSA: Simple Example

- Consider the system:

$$\emptyset \xrightarrow{k_1} X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftarrows}} X_2 \xrightarrow{k_4 x_2} \emptyset$$

- Effective system:

$$\emptyset \xrightarrow{k_1} S \xrightarrow{\hat{k}_4 s} \emptyset$$

- Fast subsystem: $k_1, k_4 \to 0$

$$X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftarrows}} X_2, \qquad S = X_1 + X_2$$

# QSSA: Simple Example

- Consider the system:

$$\emptyset \xrightarrow{k_1} X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftarrows}} X_2 \xrightarrow{k_4 x_2} \emptyset$$

- Effective system:

$$\emptyset \xrightarrow{k_1} S \xrightarrow{\hat{k}_4 s} \emptyset$$

- Fast subsystem: $k_1, k_4 \to 0$

$$X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftarrows}} X_2, \qquad S = X_1 + X_2$$

# QSSA: Simple Example

$$X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftharpoons}} X_2, \qquad S = X_1 + X_2$$

- Invariant distribution

$$X_2 \sim \mathcal{B}(S, \lambda_2) = \pi(X_2)$$

$[\lambda_1, \lambda_2] = \left[\frac{k_3}{k_2+k_3}, \frac{k_2}{k_2+k_3}\right]$ steady state solution of mean field ODE:

$$k_2\lambda_1 = k_3\lambda_2, \qquad \lambda_1 + \lambda_2 = 1$$

- Compute expectation of the rate of reaction $R_4$

$$\hat{\alpha}_4 = \mathbb{E}(\alpha_4 | S) = k_4 \mathbb{E}(X_2 | S) = \frac{k_2 k_4 S}{k_2 + k_3}$$

# QSSA: Simple Example

$$X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftarrows}} X_2, \qquad S = X_1 + X_2$$

- Invariant distribution

$$X_2 \sim \mathcal{B}(S, \lambda_2) = \pi(X_2)$$

$[\lambda_1, \lambda_2] = \left[\frac{k_3}{k_2+k_3}, \frac{k_2}{k_2+k_3}\right]$ steady state solution of mean field ODE:

$$k_2\lambda_1 = k_3\lambda_2, \qquad \lambda_1 + \lambda_2 = 1$$

- Compute expectation of the rate of reaction $R_4$

$$\hat{\alpha}_4 = \mathbb{E}(\alpha_4|S) = k_4\mathbb{E}(X_2|S) = \frac{k_2 k_4 S}{k_2 + k_3}$$

# QSSA: Simple Example

$$X_1 \underset{k_3 x_2}{\overset{k_2 x_1}{\rightleftarrows}} X_2, \qquad S = X_1 + X_2$$

- Invariant distribution

$$X_2 \sim \mathcal{B}(S, \lambda_2) = \pi(X_2)$$

$[\lambda_1, \lambda_2] = \left[\frac{k_3}{k_2+k_3}, \frac{k_2}{k_2+k_3}\right]$ steady state solution of mean field ODE:

$$k_2\lambda_1 = k_3\lambda_2, \qquad \lambda_1 + \lambda_2 = 1$$

- Compute expectation of the rate of reaction $R_4$

$$\hat{\alpha_4} = \mathbb{E}(\alpha_4|S) = k_4\mathbb{E}(X_2|S) = \frac{k_2 k_4 S}{k_2 + k_3}$$

# Multiscale approximations: Simple Example

- Therefore if we only observe the slow variable $S = X_1 + X_2$
  - $k_1$ observable
  - $k_2, k_3, k_4$ unobservable
  - QSSA: $\frac{k_2 k_4}{k_2 + k_3}$ observable, effective degradation rate of $S$
- Constrained method (details omitted)
  - Effective rate (and observable): $\frac{k_2 k_4}{k_2 + k_3 + k_4}$
- Multiscale approximations required in order to approximate intractable likelihood
- Likelihood is invariant to moves along the manifolds defined by effective rates

---

SLC, "Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling", Journal of Computational Physics, 2016

# Multiscale approximations: Simple Example

- Therefore if we only observe the slow variable $S = X_1 + X_2$
  - $k_1$ observable
  - $k_2, k_3, k_4$ unobservable
  - QSSA: $\frac{k_2 k_4}{k_2 + k_3}$ observable, effective degradation rate of $S$
- Constrained method (details omitted)
  - Effective rate (and observable): $\frac{k_2 k_4}{k_2 + k_3 + k_4}$
- Multiscale approximations required in order to approximate intractable likelihood
- Likelihood is invariant to moves along the manifolds defined by effective rates

---

SLC, "Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling", Journal of Computational Physics, 2016

# Multiscale approximations: Simple Example

- Therefore if we only observe the slow variable $S = X_1 + X_2$
  - $k_1$ observable
  - $k_2, k_3, k_4$ unobservable
  - QSSA: $\frac{k_2 k_4}{k_2+k_3}$ observable, effective degradation rate of $S$
- Constrained method (details omitted)
  - Effective rate (and observable): $\frac{k_2 k_4}{k_2+k_3+k_4}$
- Multiscale approximations required in order to approximate intractable likelihood
- Likelihood is invariant to moves along the manifolds defined by effective rates

SLC, "Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling", Journal of Computational Physics, 2016

- Therefore if we only observe the slow variable $S = X_1 + X_2$
  - $k_1$ observable
  - $k_2, k_3, k_4$ unobservable
  - QSSA: $\frac{k_2 k_4}{k_2+k_3}$ observable, effective degradation rate of $S$
- Constrained method (details omitted)
  - Effective rate (and observable): $\frac{k_2 k_4}{k_2+k_3+k_4}$
- Multiscale approximations required in order to approximate intractable likelihood
- Likelihood is invariant to moves along the manifolds defined by effective rates

SLC, "Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling", Journal of Computational Physics, 2016

# Multiscale approximations: Simple Example

- Therefore if we only observe the slow variable $S = X_1 + X_2$
  - $k_1$ observable
  - $k_2, k_3, k_4$ unobservable
  - QSSA: $\frac{k_2 k_4}{k_2 + k_3}$ observable, effective degradation rate of $S$
- Constrained method (details omitted)
  - Effective rate (and observable): $\frac{k_2 k_4}{k_2 + k_3 + k_4}$
- Multiscale approximations required in order to approximate intractable likelihood
- Likelihood is invariant to moves along the manifolds defined by effective rates

SLC, "Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling", Journal of Computational Physics, 2016

# Multiscale approximations: Simple Example

- Therefore if we only observe the slow variable $S = X_1 + X_2$
  - $k_1$ observable
  - $k_2, k_3, k_4$ unobservable
  - QSSA: $\frac{k_2 k_4}{k_2 + k_3}$ observable, effective degradation rate of $S$
- Constrained method (details omitted)
  - Effective rate (and observable): $\frac{k_2 k_4}{k_2 + k_3 + k_4}$
- Multiscale approximations required in order to approximate intractable likelihood
- Likelihood is invariant to moves along the manifolds defined by effective rates

SLC, "Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling", Journal of Computational Physics, 2016

# Multiscale approximations: Simple Example

- Therefore if we only observe the slow variable $S = X_1 + X_2$
  - $k_1$ observable
  - $k_2, k_3, k_4$ unobservable
  - QSSA: $\frac{k_2 k_4}{k_2 + k_3}$ observable, effective degradation rate of $S$
- Constrained method (details omitted)
  - Effective rate (and observable): $\frac{k_2 k_4}{k_2 + k_3 + k_4}$
- Multiscale approximations required in order to approximate intractable likelihood
- Likelihood is invariant to moves along the manifolds defined by effective rates

SLC, "Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling", Journal of Computational Physics, 2016

# Multiscale approximations: Simple Example

- Therefore if we only observe the slow variable $S = X_1 + X_2$
  - $k_1$ observable
  - $k_2, k_3, k_4$ unobservable
  - QSSA: $\frac{k_2 k_4}{k_2+k_3}$ observable, effective degradation rate of $S$
- Constrained method (details omitted)
  - Effective rate (and observable): $\frac{k_2 k_4}{k_2+k_3+k_4}$
- Multiscale approximations required in order to approximate intractable likelihood
- Likelihood is invariant to moves along the manifolds defined by effective rates

---

SLC, "Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling", Journal of Computational Physics, 2016

# Constrained approximation: Simple Example

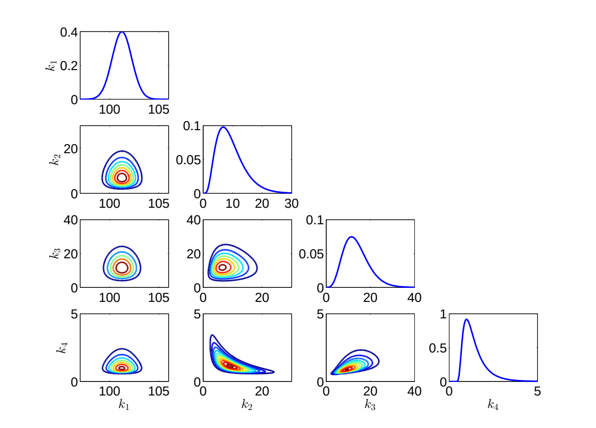



Figure: CMA approximation of the posterior arising from observations of the slow variable $S = X_1 + X_2$, concentrated around a manifold $\frac{k_1(k_2+k_3+k_4)}{k_2 k_4} = C$, i.e. more challenging than this plot suggests. (Any visualisation suggestions?)

# Outline


1 Motivating example: multiscale stochastic reaction networks

2 Parallel Adaptive Importance Sampling

3 Transport maps

4 Numerical Results

# Importance Sampling

- Posterior measure has density $\pi$
- Proposal density $\nu$
- Take $N$ samples from $\nu$, $\{x_i\}_{i=1}^N$
- Compute respective weights $w_i = \pi(x_i)/\nu(x_i)$

$$\mathbb{E}_\pi(f) \approx \frac{1}{\sum_j w_j} \sum_{i=1}^{N} f(x_i) w_i$$

- The $x_i$ are unequally weighted samples from $\pi$
- Very efficient when $\pi$ and $\nu$ are close

# Importance Sampling

- Posterior measure has density $\pi$
- Proposal density $\nu$
- Take $N$ samples from $\nu$, $\{x_i\}_{i=1}^N$
- Compute respective weights $w_i = \pi(x_i)/\nu(x_i)$

$$\mathbb{E}_\pi(f) \approx \frac{1}{\sum_j w_j} \sum_{i=1}^{N} f(x_i) w_i$$

- The $x_i$ are unequally weighted samples from $\pi$
- Very efficient when $\pi$ and $\nu$ are close

# Importance Sampling

- Posterior measure has density $\pi$
- Proposal density $\nu$
- Take $N$ samples from $\nu$, $\{x_i\}_{i=1}^N$
- Compute respective weights $w_i = \pi(x_i)/\nu(x_i)$

$$\mathbb{E}_\pi(f) \approx \frac{1}{\sum_j w_j} \sum_{i=1}^{N} f(x_i) w_i$$

- The $x_i$ are unequally weighted samples from $\pi$
- Very efficient when $\pi$ and $\nu$ are close

# Importance Sampling

- Posterior measure has density $\pi$
- Proposal density $\nu$
- Take $N$ samples from $\nu$, $\{x_i\}_{i=1}^N$
- Compute respective weights $w_i = \pi(x_i)/\nu(x_i)$

$$\mathbb{E}_\pi(f) \approx \frac{1}{\sum_j w_j} \sum_{i=1}^{N} f(x_i) w_i$$

- The $x_i$ are unequally weighted samples from $\pi$
- Very efficient when $\pi$ and $\nu$ are close

# Importance Sampling

- Posterior measure has density $\pi$
- Proposal density $\nu$
- Take $N$ samples from $\nu$, $\{x_i\}_{i=1}^N$
- Compute respective weights $w_i = \pi(x_i)/\nu(x_i)$

$$\mathbb{E}_\pi(f) \approx \frac{1}{\sum_j w_j} \sum_{i=1}^{N} f(x_i) w_i$$

- The $x_i$ are unequally weighted samples from $\pi$
- Very efficient when $\pi$ and $\nu$ are close

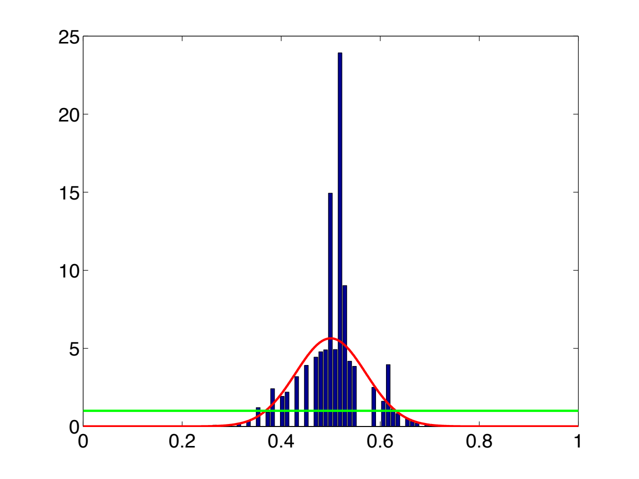
# Importance Sampling

- Posterior measure has density $\pi$
- Proposal density $\nu$
- Take $N$ samples from $\nu$, $\{x_i\}_{i=1}^N$
- Compute respective weights $w_i = \pi(x_i)/\nu(x_i)$

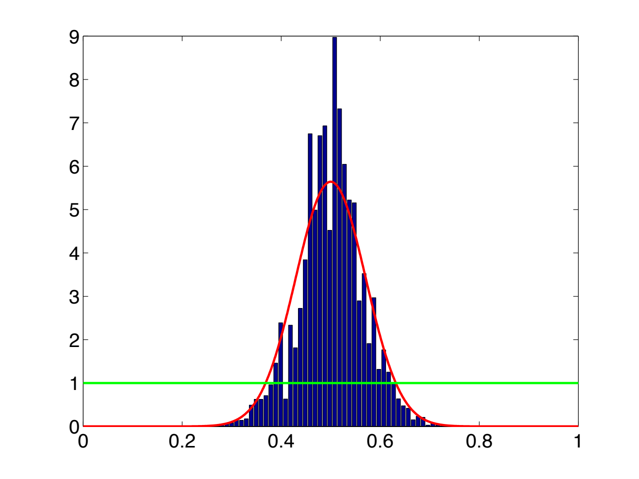
$$\mathbb{E}_\pi(f) \approx \frac{1}{\sum_j w_j} \sum_{i=1}^{N} f(x_i) w_i$$

- The $x_i$ are unequally weighted samples from $\pi$
- Very efficient when $\pi$ and $\nu$ are close

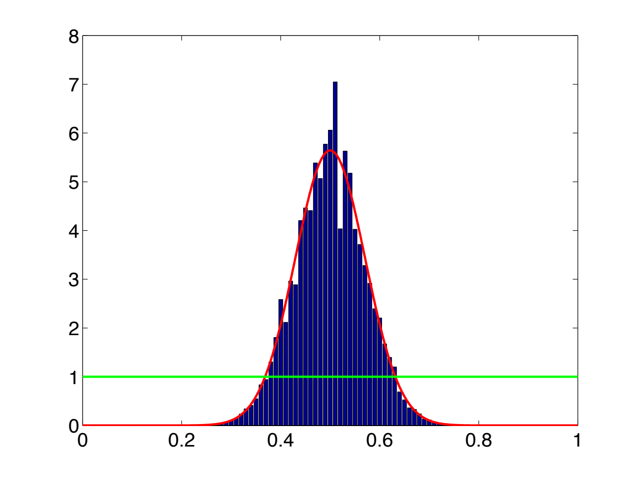
# Advantages of Importance Sampling: $10^2$ samples



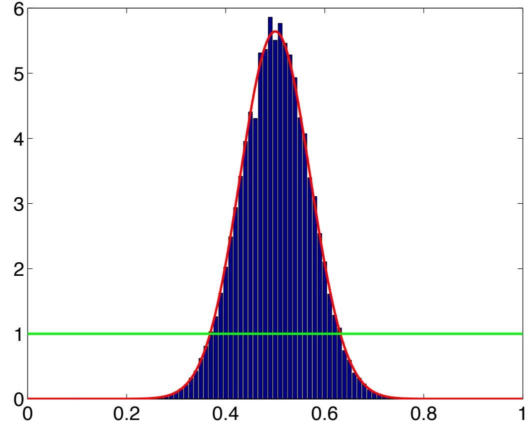
# Advantages of Importance Sampling: $10^3$ samples



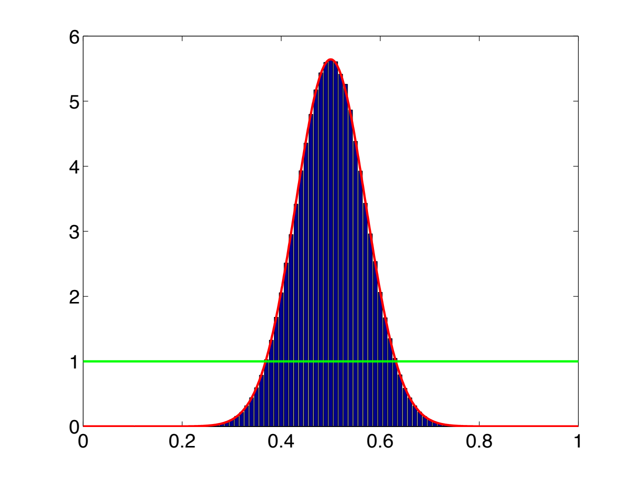
# Advantages of Importance Sampling: $10^4$ samples



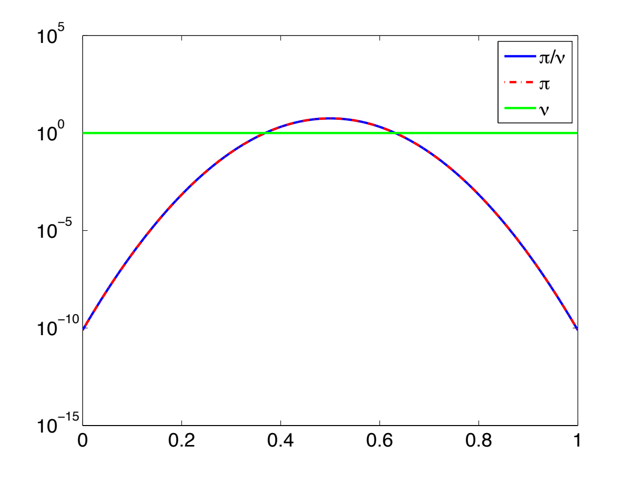
# Advantages of Importance Sampling: $10^5$ samples



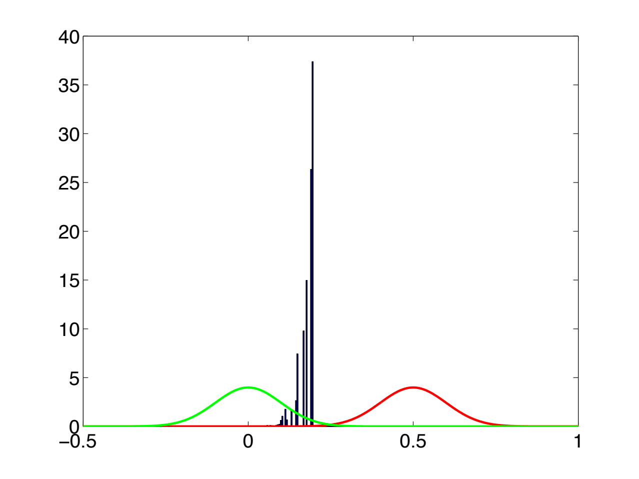
# Advantages of Importance Sampling: $10^6$ samples



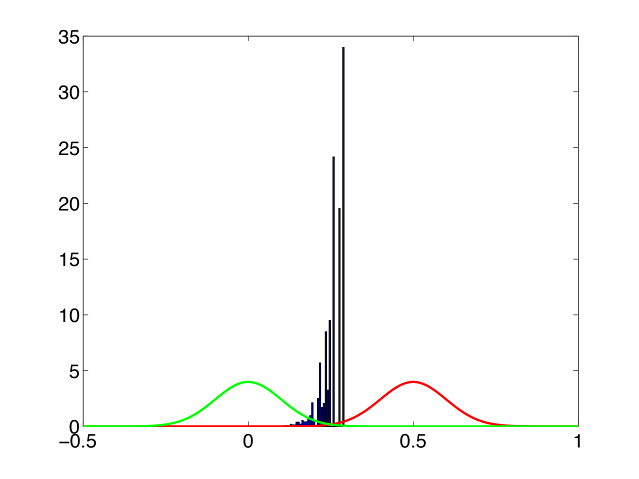
# Advantages of Importance Sampling: Weights



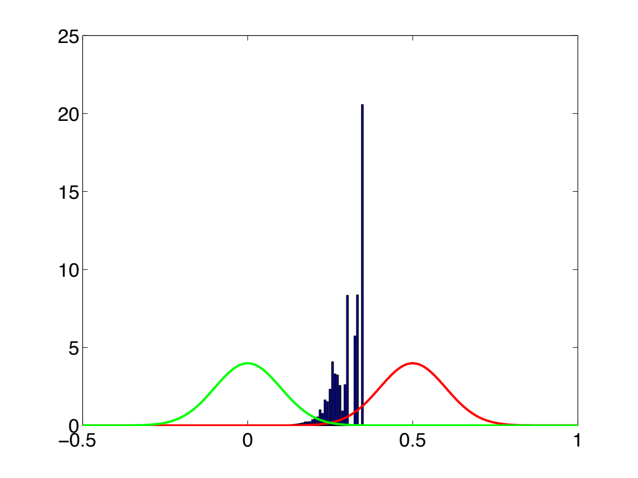
# Disadvantages of Importance Sampling: $10^2$ samples



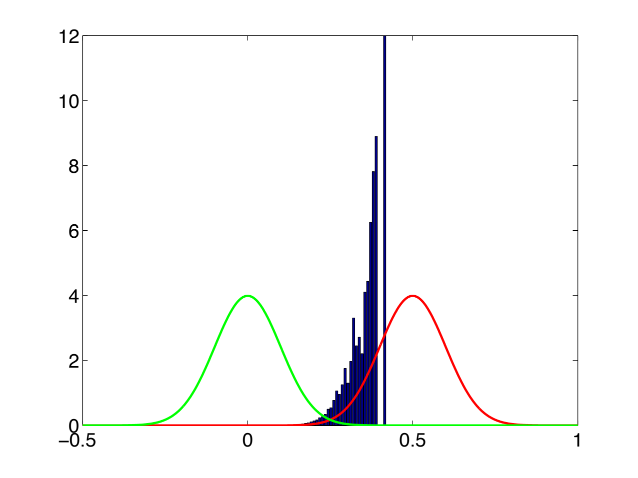
# Disadvantages of Importance Sampling: $10^3$ samples



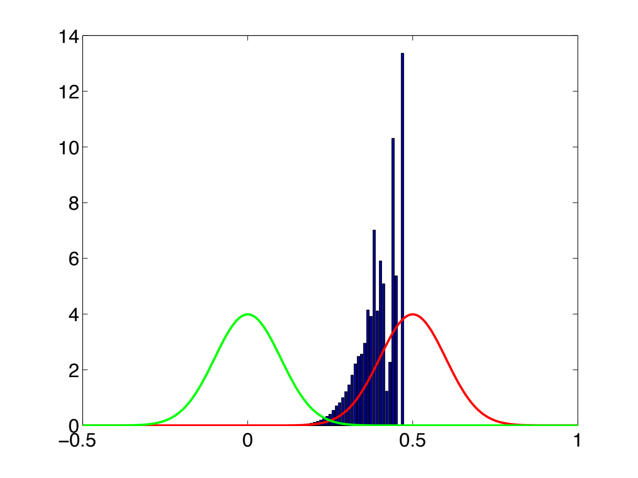
# Disadvantages of Importance Sampling: $10^4$ samples



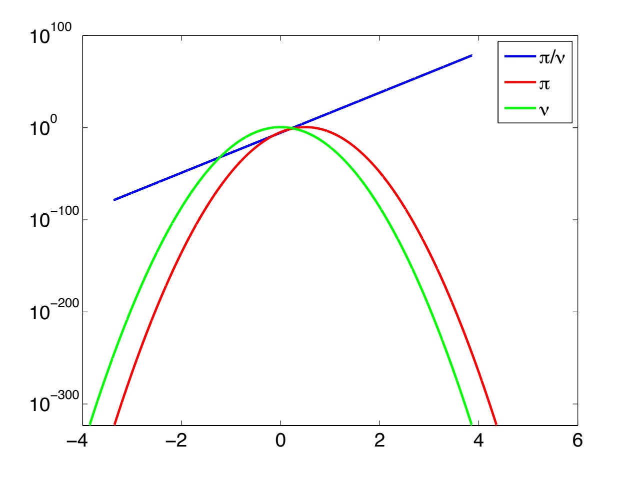
# Disadvantages of Importance Sampling: $10^5$ samples



# Disadvantages of Importance Sampling: $10^6$ samples



# Disadvantages of Importance Sampling: Weights



# Parallel Adaptive Importance Sampling

- An ensemble importance sampling method
- Proposal distribution in $k$th iteration informed by $M$ ensemble members

$$\chi^{(k)} = \frac{1}{M}\sum_{i=1}^{M} q(\cdot;\theta_i^{(k)},\beta)$$

- $q(\cdot;\cdot,\beta)$ a transition kernel, e.g. Gaussian, MALA proposal, etc
- Resampling step; ensemble transform method (or for large $M$, greedy approximation)
- If $C_{\text{overheads}} \ll C_{\text{likelihood}}$, big parallelisation payoff
- Error scales superlinearly with $M^{-1/2}$

---

C. Cotter, SLC, P. Russell, “Parallel adaptive importance sampling”, submitted to SIAM JUQ.

S. Reich, “A non-parametric ensemble transform method for Bayesian inference”, SISC 2013.

- An ensemble importance sampling method
- Proposal distribution in $k$th iteration informed by $M$ ensemble members

$$\chi^{(k)} = \frac{1}{M}\sum_{i=1}^{M} q(\cdot;\theta_i^{(k)},\beta)$$

- $q(\cdot;\cdot,\beta)$ a transition kernel, e.g. Gaussian, MALA proposal, etc
- Resampling step; ensemble transform method (or for large $M$, greedy approximation)
- If $C_{\text{overheads}} \ll C_{\text{likelihood}}$, big parallelisation payoff
- Error scales superlinearly with $M^{-1/2}$

C. Cotter, SLC, P. Russell, “Parallel adaptive importance sampling”, submitted to SIAM JUQ.

S. Reich, “A non-parametric ensemble transform method for Bayesian inference”, SISC 2013.

# Parallel Adaptive Importance Sampling

- An ensemble importance sampling method
- Proposal distribution in $k$th iteration informed by $M$ ensemble members

$$\chi^{(k)} = \frac{1}{M}\sum_{i=1}^{M} q(\cdot;\theta_i^{(k)},\beta)$$

- $q(\cdot;\cdot,\beta)$ a transition kernel, e.g. Gaussian, MALA proposal, etc
- Resampling step; ensemble transform method (or for large $M$, greedy approximation)
- If $C_{\text{overheads}} \ll C_{\text{likelihood}}$, big parallelisation payoff
- Error scales superlinearly with $M^{-1/2}$

---

C. Cotter, SLC, P. Russell, “Parallel adaptive importance sampling”, submitted to SIAM JUQ.

S. Reich, “A non-parametric ensemble transform method for Bayesian inference”, SISC 2013.

# Parallel Adaptive Importance Sampling

- An ensemble importance sampling method
- Proposal distribution in $k$th iteration informed by $M$ ensemble members

$$\chi^{(k)} = \frac{1}{M}\sum_{i=1}^{M} q(\cdot; \theta_i^{(k)}, \beta)$$

- $q(\cdot;\cdot,\beta)$ a transition kernel, e.g. Gaussian, MALA proposal, etc
- Resampling step; ensemble transform method (or for large $M$, greedy approximation)
- If $C_{\text{overheads}} \ll C_{\text{likelihood}}$, big parallelisation payoff
- Error scales superlinearly with $M^{-1/2}$

---

C. Cotter, SLC, P. Russell, "Parallel adaptive importance sampling", submitted to SIAM JUQ.

S. Reich, "A non-parametric ensemble transform method for Bayesian inference", SISC 2013.

- An ensemble importance sampling method
- Proposal distribution in $k$th iteration informed by $M$ ensemble members

$$\chi^{(k)} = \frac{1}{M}\sum_{i=1}^{M} q(\cdot;\theta_i^{(k)},\beta)$$

- $q(\cdot;\cdot,\beta)$ a transition kernel, e.g. Gaussian, MALA proposal, etc
- Resampling step; ensemble transform method (or for large $M$, greedy approximation)
- If $C_{\text{overheads}} \ll C_{\text{likelihood}}$, big parallelisation payoff
- Error scales superlinearly with $M^{-1/2}$

---

C. Cotter, SLC, P. Russell, "Parallel adaptive importance sampling", submitted to SIAM JUQ.

S. Reich, "A non-parametric ensemble transform method for Bayesian inference", SISC 2013.

# Parallel Adaptive Importance Sampling

- An ensemble importance sampling method
- Proposal distribution in $k$th iteration informed by $M$ ensemble members

$$\chi^{(k)} = \frac{1}{M}\sum_{i=1}^{M} q(\cdot;\theta_i^{(k)},\beta)$$

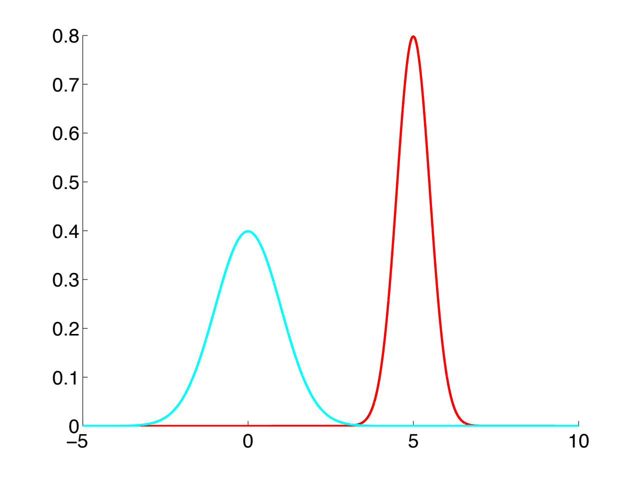
- $q(\cdot;\cdot,\beta)$ a transition kernel, e.g. Gaussian, MALA proposal, etc
- Resampling step; ensemble transform method (or for large $M$, greedy approximation)
- If $C_{\text{overheads}} \ll C_{\text{likelihood}}$, big parallelisation payoff
- Error scales superlinearly with $M^{-1/2}$

---

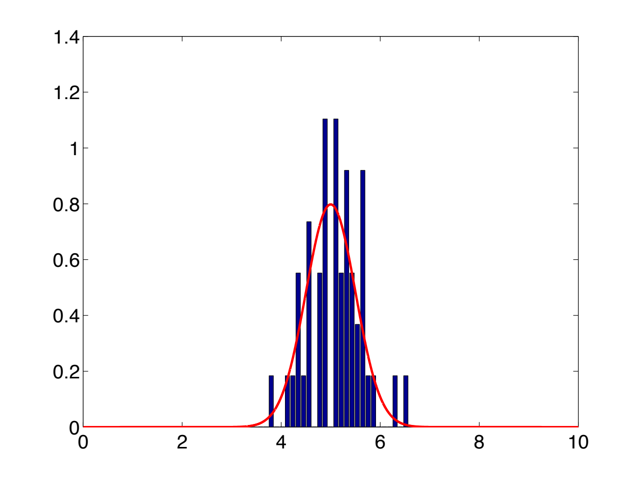
C. Cotter, SLC, P. Russell, "Parallel adaptive importance sampling", submitted to SIAM JUQ.

S. Reich, "A non-parametric ensemble transform method for Bayesian inference", SISC 2013.

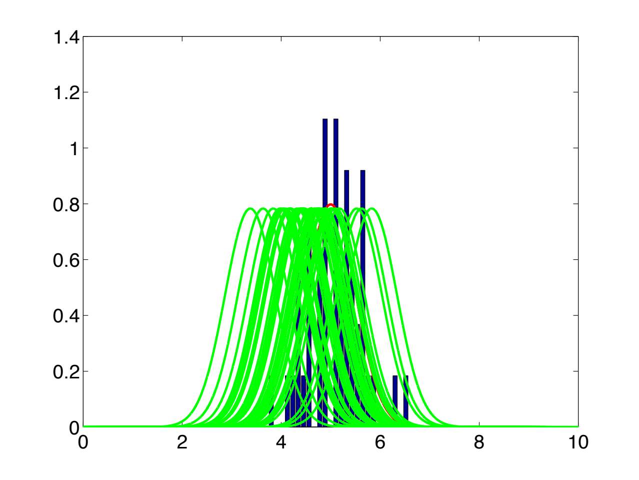
# Parallel Adaptive Importance Sampling: Prior and Posterior



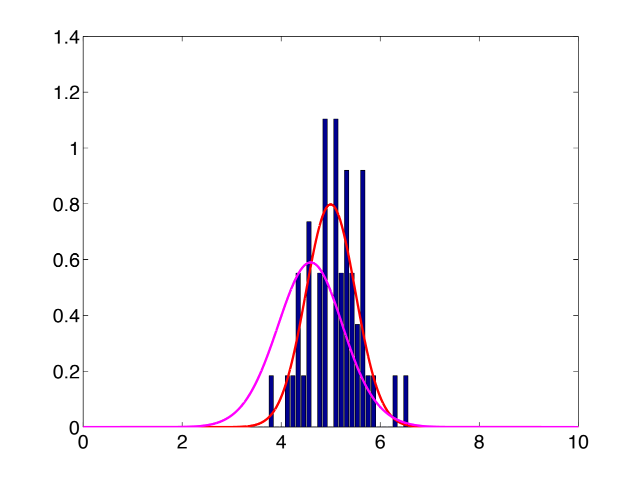
# Parallel Adaptive Importance Sampling: Current State $\mathbf{X}_i$



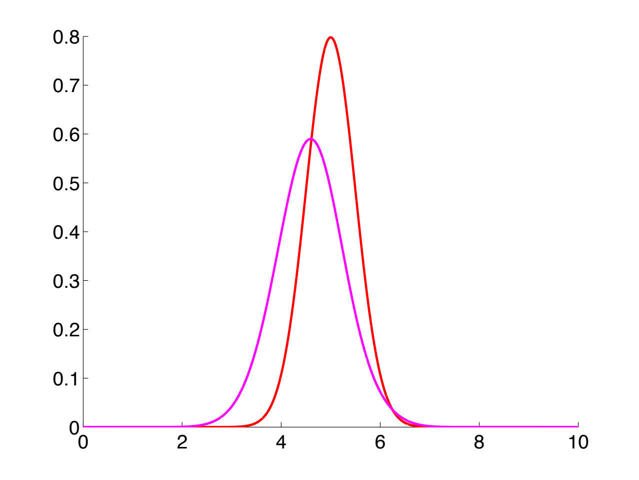
# Parallel Adaptive Importance Sampling: MALA Proposals



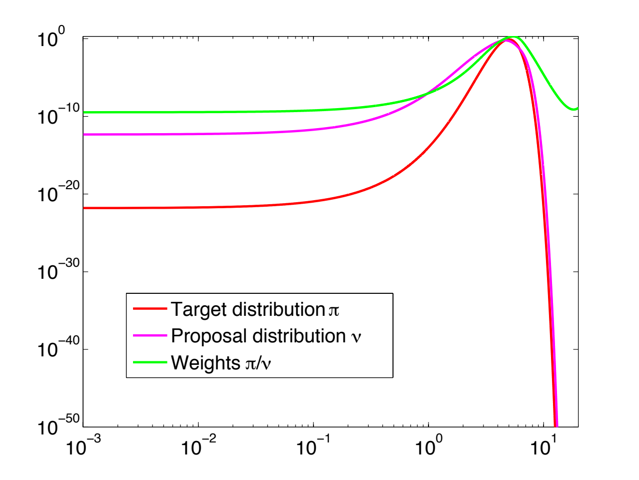
# Parallel Adaptive Importance Sampling: Aggregate Proposal



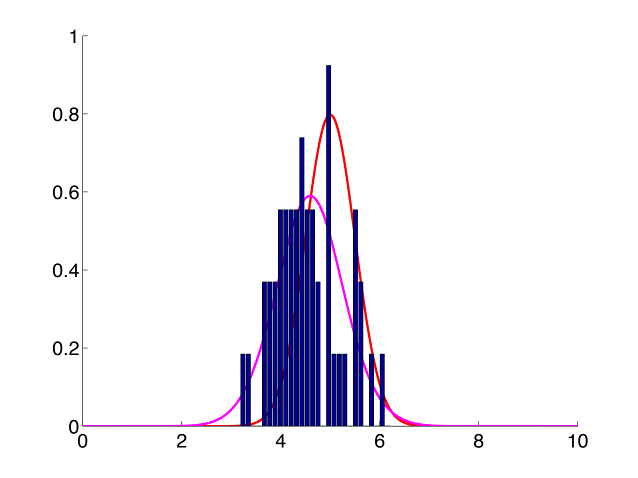
# Parallel Adaptive Importance Sampling: Aggregate Proposal



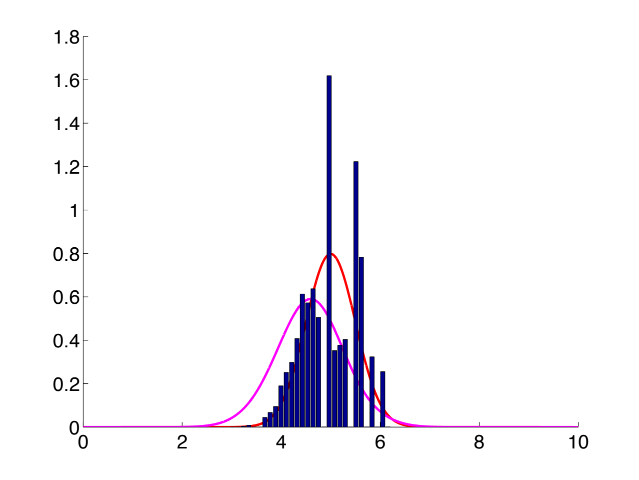
# Parallel Adaptive Importance Sampling: Aggregate Proposal and Weight Function



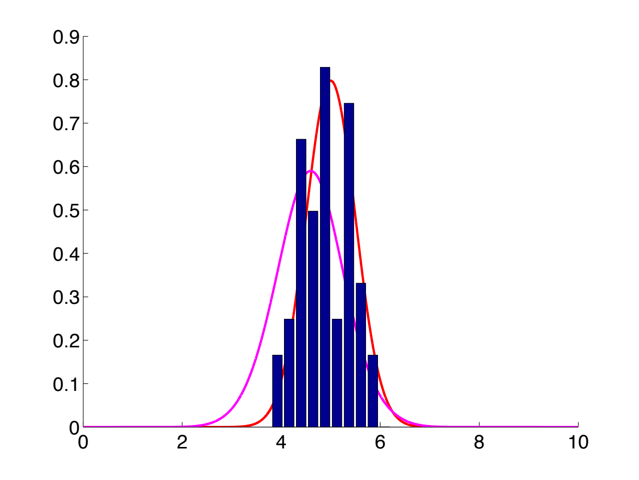
# Parallel Adaptive Importance Sampling: Samples from Proposal



# Parallel Adaptive Importance Sampling: Sample Weights



# Parallel Adaptive Importance Sampling: Resampled States



# PAIS - pros and cons

**PROS:**

- Possible big speed-ups with parallelisation
- Well-informed proposals
- Reduces variance of importance weights
- Adaptive to global differences in scales of parameters

**CONS:**

- Posterior concentrated on lower dimensional manifold:
  - Stability issues
  - Slow convergence
  - Requires large ensemble size (expensive)
- Particle transition kernel $q$ needs to “know” about the manifold

# PAIS - pros and cons

**PROS:**

- Possible big speed-ups with parallelisation
- Well-informed proposals
- Reduces variance of importance weights
- Adaptive to global differences in scales of parameters

**CONS:**

- Posterior concentrated on lower dimensional manifold:
  - Stability issues
  - Slow convergence
  - Requires large ensemble size (expensive)
- Particle transition kernel $q$ needs to “know” about the manifold

# PAIS - pros and cons

**PROS:**

- Possible big speed-ups with parallelisation
- Well-informed proposals
- Reduces variance of importance weights
- Adaptive to global differences in scales of parameters

**CONS:**

- Posterior concentrated on lower dimensional manifold:
  - Stability issues
  - Slow convergence
  - Requires large ensemble size (expensive)
- Particle transition kernel $q$ needs to “know” about the manifold

# PAIS - pros and cons

**PROS:**

- Possible big speed-ups with parallelisation
- Well-informed proposals
- Reduces variance of importance weights
- Adaptive to global differences in scales of parameters

**CONS:**

- Posterior concentrated on lower dimensional manifold:
  - Stability issues
  - Slow convergence
  - Requires large ensemble size (expensive)
- Particle transition kernel $q$ needs to “know” about the manifold

# PAIS - pros and cons

**PROS:**

- Possible big speed-ups with parallelisation
- Well-informed proposals
- Reduces variance of importance weights
- Adaptive to global differences in scales of parameters

**CONS:**

- Posterior concentrated on lower dimensional manifold:
  - Stability issues
  - Slow convergence
  - Requires large ensemble size (expensive)
- Particle transition kernel $q$ needs to “know” about the manifold

# PAIS - pros and cons

**PROS:**

- Possible big speed-ups with parallelisation
- Well-informed proposals
- Reduces variance of importance weights
- Adaptive to global differences in scales of parameters

**CONS:**

- Posterior concentrated on lower dimensional manifold:
  - Stability issues
  - Slow convergence
  - Requires large ensemble size (expensive)
- Particle transition kernel $q$ needs to "know" about the manifold

# PAIS - pros and cons

**PROS:**

- Possible big speed-ups with parallelisation
- Well-informed proposals
- Reduces variance of importance weights
- Adaptive to global differences in scales of parameters

**CONS:**

- Posterior concentrated on lower dimensional manifold:
  - Stability issues
  - Slow convergence
  - Requires large ensemble size (expensive)
- Particle transition kernel $q$ needs to “know” about the manifold

# PAIS - pros and cons

**PROS:**

- Possible big speed-ups with parallelisation
- Well-informed proposals
- Reduces variance of importance weights
- Adaptive to global differences in scales of parameters

**CONS:**

- Posterior concentrated on lower dimensional manifold:
  - Stability issues
  - Slow convergence
  - Requires large ensemble size (expensive)
- Particle transition kernel $q$ needs to “know” about the manifold

# PAIS - pros and cons

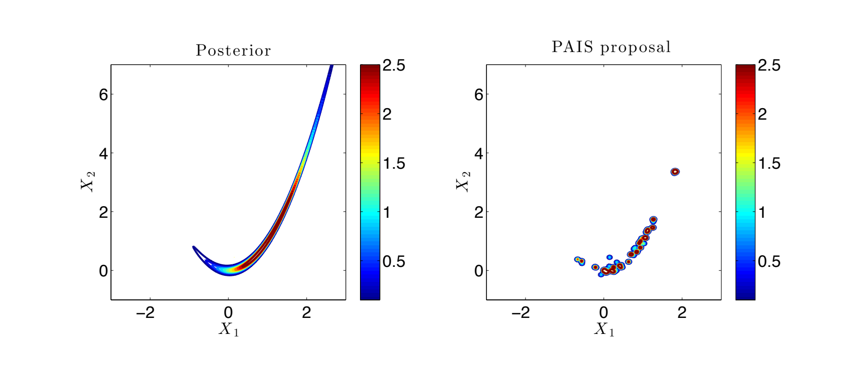
**PROS:**

- Possible big speed-ups with parallelisation
- Well-informed proposals
- Reduces variance of importance weights
- Adaptive to global differences in scales of parameters

**CONS:**

- Posterior concentrated on lower dimensional manifold:
  - Stability issues
  - Slow convergence
  - Requires large ensemble size (expensive)
- Particle transition kernel $q$ needs to “know” about the manifold

# Motivation



# Outline


1 Motivating example: multiscale stochastic reaction networks

2 Parallel Adaptive Importance Sampling

3 Transport maps

4 Numerical Results

# Transport maps

- Posteriors concentrated on lower dimensional manifolds lead to poor mixing
- Transport maps simplify the problem
- Find homeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$ which maps target measure $\pi$ to an easily explored reference measure $\pi_r$

$$\mu(T^{-1}(A)) = \mu_r(A)$$

- Simple proposal densities on $\pi_r$ map to complex informed densities on $\pi$ via $T^{-1}$

$$v \sim T^{-1}(q(\cdot, u; \beta))$$

# Transport maps

- Posteriors concentrated on lower dimensional manifolds lead to poor mixing
- Transport maps simplify the problem
- Find homeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$ which maps target measure $\pi$ to an easily explored reference measure $\pi_r$

$$\mu(T^{-1}(A)) = \mu_r(A)$$

- Simple proposal densities on $\pi_r$ map to complex informed densities on $\pi$ via $T^{-1}$

$$v \sim T^{-1}(q(\cdot, u; \beta))$$

# Transport maps

- Posteriors concentrated on lower dimensional manifolds lead to poor mixing
- Transport maps simplify the problem
- Find homeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$ which maps target measure $\pi$ to an easily explored reference measure $\pi_r$

$$\mu(T^{-1}(A)) = \mu_r(A)$$

- Simple proposal densities on $\pi_r$ map to complex informed densities on $\pi$ via $T^{-1}$

$$v \sim T^{-1}(q(\cdot, u; \beta))$$

# Transport maps

- Posteriors concentrated on lower dimensional manifolds lead to poor mixing
- Transport maps simplify the problem
- Find homeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$ which maps target measure $\pi$ to an easily explored reference measure $\pi_r$

$$\mu(T^{-1}(A)) = \mu_r(A)$$

- Simple proposal densities on $\pi_r$ map to complex informed densities on $\pi$ via $T^{-1}$

$$v \sim T^{-1}(q(\cdot, u; \beta))$$

# Transport maps

- Exists subject to conditions, but not necessarily invertible
- Find invertible map $T$ which minimises KL divergence between $\pi$ and $|J_{\tilde{T}}(\theta)|\pi_r \circ \tilde{T} = \tilde{\pi}$ where $\pi_r = \mathcal{N}(0, I)$
- In practice, find finite dimensional monotonic map $T$ which minimises the Monte Carlo approximation of KL divergence from samples from $\pi$

$$
\begin{aligned}
D_{\mathrm{KL}}(\pi\|\tilde{\pi}) &= \mathbb{E}_{\pi}\left[\log\left(\frac{\pi(\theta)}{\tilde{\pi}(\theta)}\right)\right] \\
&= \mathbb{E}_{\pi}\left[\log \pi(\theta) - \log \pi_r(\tilde{T}(\theta)) - \log |J_{\tilde{T}}(\theta)|\right]
\end{aligned}
$$

M. Parno, Y. Marzouk, "Transport Map Accelerated Markov Chain Monte Carlo", SIAM journal on uncertainty quantification, 2018.

# Transport maps

- Exists subject to conditions, but not necessarily invertible
- Find invertible map $T$ which minimises KL divergence between $\pi$ and $|J_{\tilde{T}}(\theta)|\pi_r \circ \tilde{T} = \tilde{\pi}$ where $\pi_r = \mathcal{N}(0, I)$
- In practice, find finite dimensional monotonic map $T$ which minimises the Monte Carlo approximation of KL divergence from samples from $\pi$

$$
\begin{aligned}
D_{\text{KL}}(\pi \| \tilde{\pi}) &= \mathbb{E}_\pi \left[ \log \left( \frac{\pi(\theta)}{\tilde{\pi}(\theta)} \right) \right] \\
&= \mathbb{E}_\pi \left[ \log \pi(\theta) - \log \pi_r(\tilde{T}(\theta)) - \log |J_{\tilde{T}}(\theta)| \right]
\end{aligned}
$$

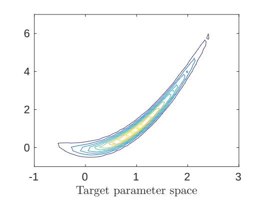
M. Parno, Y. Marzouk, “Transport Map Accelerated Markov Chain Monte Carlo”, SIAM journal on uncertainty quantification, 2018.

- Exists subject to conditions, but not necessarily invertible
- Find invertible map $T$ which minimises KL divergence between $\pi$ and $|J_{\tilde{T}}(\theta)|\pi_r \circ \tilde{T} = \tilde{\pi}$ where $\pi_r = \mathcal{N}(0, I)$
- In practice, find finite dimensional monotonic map $T$ which minimises the Monte Carlo approximation of KL divergence from samples from $\pi$

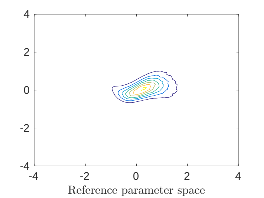
$$
\begin{aligned}
D_{\text{KL}}(\pi \| \tilde{\pi}) &= \mathbb{E}_\pi \left[ \log \left( \frac{\pi(\theta)}{\tilde{\pi}(\theta)} \right) \right] \\
&= \mathbb{E}_\pi \left[ \log \pi(\theta) - \log \pi_r(\tilde{T}(\theta)) - \log |J_{\tilde{T}}(\theta)| \right]
\end{aligned}
$$

M. Parno, Y. Marzouk, “Transport Map Accelerated Markov Chain Monte Carlo”, SIAM journal on uncertainty quantification, 2018.

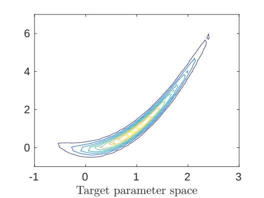
# Transport map simplification of Rosenbrock




(a) Original sample $\theta$ from MH-RW algorithm.




(b) Push forward of $\theta$ onto reference space.




(c) Pull back of reference sample onto target space.

Figure: The effect of the approximate transport map $\tilde{T}$ on a sample from the Rosenbrock target density.

# Outline of approach

- Run standard PAIS with transport map equal to the identity
- Periodically train the transport map on the current importance-weighted sample
- Proposal distribution becomes sum of pullback of Gaussians through the transport map
- Learns local correlations and structure
- Allows complex targets to be described more accurately by sum of fewer kernels

# Outline of approach

- Run standard PAIS with transport map equal to the identity
- Periodically train the transport map on the current importance-weighted sample
- Proposal distribution becomes sum of pullback of Gaussians through the transport map
- Learns local correlations and structure
- Allows complex targets to be described more accurately by sum of fewer kernels

# Outline of approach

- Run standard PAIS with transport map equal to the identity
- Periodically train the transport map on the current importance-weighted sample
- Proposal distribution becomes sum of pullback of Gaussians through the transport map
- Learns local correlations and structure
- Allows complex targets to be described more accurately by sum of fewer kernels

# Outline of approach

- Run standard PAIS with transport map equal to the identity
- Periodically train the transport map on the current importance-weighted sample
- Proposal distribution becomes sum of pullback of Gaussians through the transport map
- Learns local correlations and structure
- Allows complex targets to be described more accurately by sum of fewer kernels

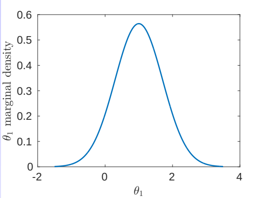
# Outline of approach

- Run standard PAIS with transport map equal to the identity
- Periodically train the transport map on the current importance-weighted sample
- Proposal distribution becomes sum of pullback of Gaussians through the transport map
- Learns local correlations and structure
- Allows complex targets to be described more accurately by sum of fewer kernels

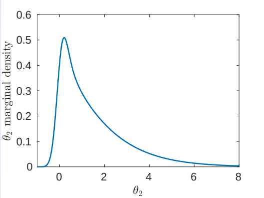
# Outline

1. Motivating example: multiscale stochastic reaction networks
2. Parallel Adaptive Importance Sampling
3. Transport maps
4. Numerical Results

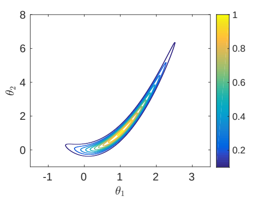
# Rosenbrock density




(a) Marginal density function for $\theta_1$.



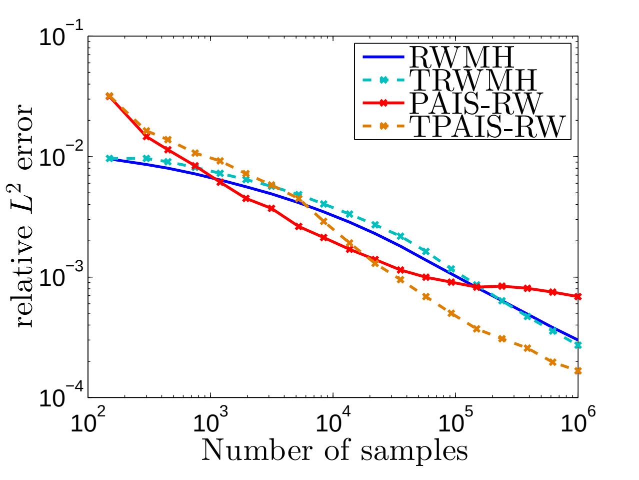

(b) Marginal density function for $\theta_2$.




(c) Contour plot for Rosenbrock density.

Figure: Visualisation of the Rosenbrock density.

# Rosenbrock density



# Multiscale stochastic reaction network example

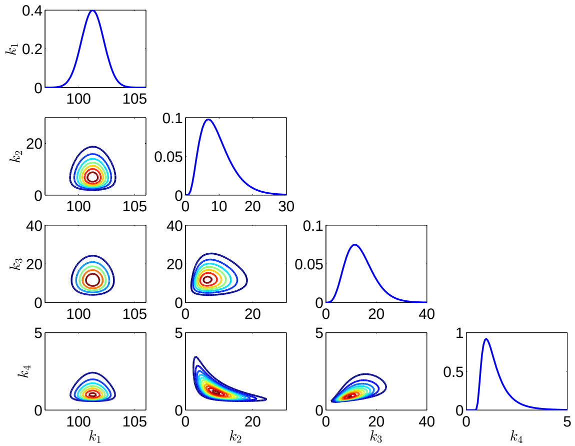



Figure: CMA approximation of the posterior arising from observations of the slow variable $S = X_1 + X_2$, concentrated around a manifold $\frac{k_1(k_2+k_3+k_4)}{k_2 k_4} = C$, i.e. more challenging than this suggests.

# Multiscale stochastic reaction network example

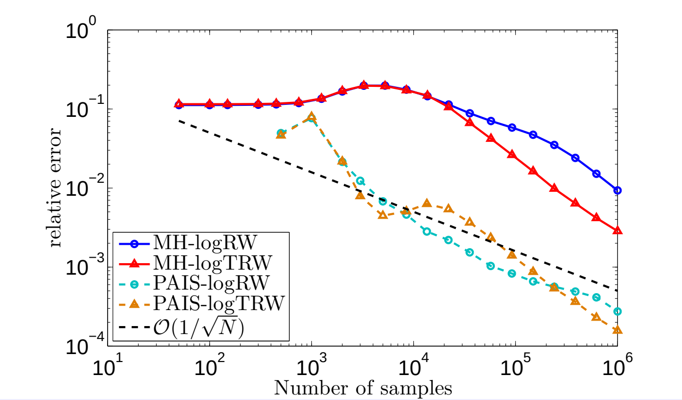
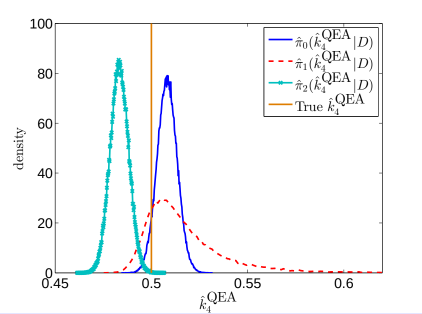
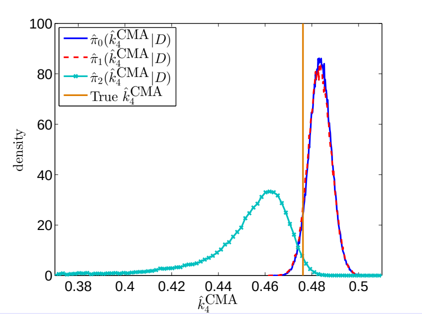



Figure: Sampling algorithms with a log preconditioner for $\tilde{T}$.

# Multiscale stochastic reaction network example




(a) Marginal for $\hat{k}_4^{\mathrm{QEA}}$.




(b) Marginal for $\hat{k}_4^{\mathrm{CMA}}$.

Figure: Comparison of the approximate marginal densities for the quantities $\hat{k}_4^{\mathrm{QEA}} = \frac{k_2 k_4}{k_2+k_3}$ and $\hat{k}_4^{\mathrm{CMA}} = \frac{k_2 k_4}{k_2+k_3+k_4}$ for the posteriors arising from (i) fast and slow data (blue), and slow data using (ii) constrained (red) and (iii) QSSA (cyan) multiscale approximations.

# Conclusions

- Noisily observed multiscale systems often result in inverse problems with density concentrated near a manifold
- Transport maps can accelerate sampling of complex probability distributions
- Importantly for importance sampling schemes, they can improve stability significantly, reduce number of required particles
- The map requires a good initial sample from the posterior
- Numerical result appears to validate constrained multiscale approximation method
- Methodology also works very well for multimodal targets

# Conclusions

- Noisily observed multiscale systems often result in inverse problems with density concentrated near a manifold
- Transport maps can accelerate sampling of complex probability distributions
- Importantly for importance sampling schemes, they can improve stability significantly, reduce number of required particles
- The map requires a good initial sample from the posterior
- Numerical result appears to validate constrained multiscale approximation method
- Methodology also works very well for multimodal targets

# Conclusions

- Noisily observed multiscale systems often result in inverse problems with density concentrated near a manifold
- Transport maps can accelerate sampling of complex probability distributions
- Importantly for importance sampling schemes, they can improve stability significantly, reduce number of required particles
- The map requires a good initial sample from the posterior
- Numerical result appears to validate constrained multiscale approximation method
- Methodology also works very well for multimodal targets

# Conclusions

- Noisily observed multiscale systems often result in inverse problems with density concentrated near a manifold
- Transport maps can accelerate sampling of complex probability distributions
- Importantly for importance sampling schemes, they can improve stability significantly, reduce number of required particles
- The map requires a good initial sample from the posterior
- Numerical result appears to validate constrained multiscale approximation method
- Methodology also works very well for multimodal targets

# Conclusions

- Noisily observed multiscale systems often result in inverse problems with density concentrated near a manifold
- Transport maps can accelerate sampling of complex probability distributions
- Importantly for importance sampling schemes, they can improve stability significantly, reduce number of required particles
- The map requires a good initial sample from the posterior
- Numerical result appears to validate constrained multiscale approximation method
- Methodology also works very well for multimodal targets

# Conclusions

- Noisily observed multiscale systems often result in inverse problems with density concentrated near a manifold
- Transport maps can accelerate sampling of complex probability distributions
- Importantly for importance sampling schemes, they can improve stability significantly, reduce number of required particles
- The map requires a good initial sample from the posterior
- Numerical result appears to validate constrained multiscale approximation method
- Methodology also works very well for multimodal targets

# References


- SLC, I. Kevrekidis, P. Russell, “Transport map accelerated adaptive importance sampling, and application to inverse problems arising from multiscale stochastic reaction networks”, appearing on arxiv shortly.
- M. Parno, Y. Marzouk, “Transport Map Accelerated Markov Chain Monte Carlo”, SIAM journal on uncertainty quantification, 2018.
- C. Cotter, SLC, P. Russell, “Parallel adaptive importance sampling”, submitted to SIAM JUQ.
- S. Reich, “A non-parametric ensemble transform method for Bayesian inference”, SISC 2013.
- SLC, “Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling”, Journal of Computational Physics, 2016