Numerical Methods for Bayesian Inverse Problems

Lecture 3: Markov Chain Monte Carlo

Robert Scheichl

Institute for Applied Mathematics & IWR, Heidelberg University

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Autumn School – "Uncertainty Quantification for High-Dimensional Problems"
CWI Amsterdam, October 7-11, 2024

(Thanks to Björn Sprungk, TU Freiberg)

Bayesian Approach to Inverse Problems

Find parameter $u \in \mathcal{X}$ from noisy observations of the Bayesian statistical model

$$Y = \mathcal{G}(U) + \eta$$

- Y, U, η treated as random fields / variables
- Prior distribution $U \sim \pi_0$, Gaussian noise $\eta \sim \mathrm{N}(0, \Sigma)$
- Conditioning the prior on data $Y = y$ yields as the solution the posterior π^y.

Bayesian Approach to Inverse Problems

Find parameter $u \in \mathcal{X}$ from noisy observations of the Bayesian statistical model

$$Y = \mathcal{G}(U) + \eta$$

- Y, U, η treated as random fields / variables
- Prior distribution $U \sim \pi_0$, Gaussian noise $\eta \sim \mathrm{N}(0, \Sigma)$
- Conditioning the prior on data $Y = y$ yields as the solution the posterior π^y.

Provides

- a well-posed formulation of the inverse problem;
- point estimates for u (related to Tikhonov-regularized solution);
- a way to quantify the (remaining) uncertainty about u.

Bayesian Approach to Inverse Problems

Find parameter $u \in \mathcal{X}$ from noisy observations of the Bayesian statistical model

$$Y = \mathcal{G}(U) + \eta$$

- Y, U, η treated as random fields / variables
- Prior distribution $U \sim \pi_0$, Gaussian noise $\eta \sim \mathrm{N}(0, \Sigma)$
- Conditioning the prior on data $Y = y$ yields as the solution the posterior π^y.

Provides

- a well-posed formulation of the inverse problem;
- point estimates for u (related to Tikhonov-regularized solution);
- a way to quantify the (remaining) uncertainty about u.

But how to actually compute point estimates and quantify uncertainty?

Numerical Methods

Task

Generate (approximate) samples of $u \in \mathcal{X}$ according to the posterior measure

$$\pi_{u|y}(\mathrm{d}u) = \frac{1}{Z} \exp(-\Phi(u;y))\, \pi_0(\mathrm{d}u),$$

where $\mathcal{X}$ is high- or infinite-dimensional and the potential $\Phi(u;y)$ involves solving a complex infinite-dimensional physical (e.g., PDE) model.

Compute posterior expectations for quantities of interest $f : \mathcal{X} \to \mathcal{Z}$

$$\mathbb{E}_{\pi_{u|y}}[f] := \int_{\mathcal{X}} f(u)\, \pi_{u|y}(\mathrm{d}u),$$

such as posterior moments (e.g., $f(u) = u$), probabilities (e.g., $f(u) = \mathbf{1}_A(u)$),...

Task

Generate (approximate) samples of $u \in \mathcal{X}$ according to the posterior measure

$$\pi_{u|y}(\mathrm{d}u) = \frac{1}{Z} \exp(-\Phi(u;y))\, \pi_0(\mathrm{d}u),$$

where $\mathcal{X}$ is high- or infinite-dimensional and the potential $\Phi(u;y)$ involves solving a complex infinite-dimensional physical (e.g., PDE) model.

Compute posterior expectations for quantities of interest $f\colon \mathcal{X} \to \mathcal{Z}$

$$\mathbb{E}_{\pi_{u|y}}[f] := \int_{\mathcal{X}} f(u)\, \pi_{u|y}(\mathrm{d}u),$$

such as posterior moments (e.g., $f(u) = u$), probabilities (e.g., $f(u) = \mathbf{1}_A(u)$),...

Numerical Challenges:

1. Error due to num. approximation of Φ.
2. Error due to discretisation of $u \in \mathcal{X}$.
3. High-dimensional numerical quadrature.
4. Normalizing constant Z inaccessible.
5. The more data, the more $\pi_{u|y}$ concentrates ('needle in a haystack')

Addressing the challenges

- The discretization error due to approximating $\mathcal{G}$ by $\mathcal{G}_h$ (e.g., via FEMs) can be addressed using the well-posedness result from the previous lecture for $\pi^h_{u|y}(\mathrm{d}u) \propto \exp(-\Phi_h(u))\,\pi_0(\mathrm{d}u)$

$$\|\mathbb{E}_{\pi^h_{u|y}}[f] - \mathbb{E}_{\pi_{u|y}}[f]\| \leq c\|\mathcal{G} - \mathcal{G}_h\|_{L^2_{\pi_0}} \leq Ch^p.$$

Addressing the challenges

- The discretization error due to approximating $\mathcal{G}$ by $\mathcal{G}_h$ (e.g., via FEMs) can be addressed using the well-posedness result from the previous lecture for $\pi^h_{u|y}(\mathrm{d}u) \propto \exp(-\Phi_h(u))\,\pi_0(\mathrm{d}u)$

$$\|\mathbb{E}_{\pi^h_{u|y}}[f] - \mathbb{E}_{\pi_{u|y}}[f]\| \leq c\|\mathcal{G} - \mathcal{G}_h\|_{L^2_{\pi_0}} \leq Ch^p.$$

- Similar result can be shown for the discretization error due to approximating the posterior $\pi_{u|y}$ on the infinite-dimensional parameter domain $\mathcal{X}$ via a posterior $\pi^s_{u|y}$ on an s-dimensional one $\mathcal{X}_s$ (e.g., truncated Karhunen-Loève) [Hoang, Schwab, Stuart, 2013]:

$$\|\mathbb{E}_{\pi^s_{u|y}}[f] - \mathbb{E}_{\pi_{u|y}}[f]\| \leq Cs^{-q}.$$

Addressing the challenges

- The discretization error due to approximating $\mathcal{G}$ by $\mathcal{G}_h$ (e.g., via FEMs) can be addressed using the well-posedness result from the previous lecture for $\pi^h_{u|y}(\mathrm{d}u) \propto \exp(-\Phi_h(u))\,\pi_0(\mathrm{d}u)$

$$\|\mathbb{E}_{\pi^h_{u|y}}[f] - \mathbb{E}_{\pi_{u|y}}[f]\| \leq c\|\mathcal{G} - \mathcal{G}_h\|_{L^2_{\pi_0}} \leq Ch^p.$$

- Similar result can be shown for the discretization error due to approximating the posterior $\pi_{u|y}$ on the infinite-dimensional parameter domain $\mathcal{X}$ via a posterior $\pi^s_{u|y}$ on an s-dimensional one $\mathcal{X}_s$ (e.g., truncated Karhunen-Loève) [Hoang, Schwab, Stuart, 2013]:

$$\|\mathbb{E}_{\pi^s_{u|y}}[f] - \mathbb{E}_{\pi_{u|y}}[f]\| \leq Cs^{-q}.$$

Central Goal

Keeping those two errors small leads to **high cost per sample** (small FE mesh size h) and **high dimensional quadrature**! Thus, we need efficient (sampling-based) numerical integration methods that can deal with **high-dimensional, unnormalized densities** that **concentrate** in parts of the parameter domain.

Standard Monte Carlo Method

- Given a sequence $\{U_k\}$ of i.i.d. copies of a given random variable $U \sim \pi$, standard Monte Carlo simulation uses the estimator

$$\mathbb{E}_\pi[U] := \int_X u\,\pi(\mathrm{d}u) \approx \frac{S_M}{M}, \qquad S_M = U_1 + \cdots + U_M.$$

- *Strong Law of Large Numbers*: $\dfrac{S_M}{M} \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_\pi[U]$ (a.s. = almost surely)

Standard Monte Carlo Method

- Given a sequence $\{U_k\}$ of i.i.d. copies of a given random variable $U \sim \pi$, standard Monte Carlo simulation uses the estimator

$$\mathbb{E}_\pi[U] := \int_X u\,\pi(\mathrm{d}u) \approx \frac{S_M}{M}, \qquad S_M = U_1 + \cdots + U_M.$$

- *Strong Law of Large Numbers*: $\dfrac{S_M}{M} \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_\pi[U]$ (a.s. = almost surely)

- For any measurable function f also $\dfrac{1}{M}\displaystyle\sum_{k=1}^{M} f(U_k) \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_\pi[f(U)] =: \mathbb{E}_\pi[f]$

Standard Monte Carlo Method

- Given a sequence $\{U_k\}$ of i.i.d. copies of a given random variable $U \sim \pi$, standard Monte Carlo simulation uses the estimator

$$\mathbb{E}_\pi[U] := \int_X u\,\pi(\mathrm{d}u) \approx \frac{S_M}{M}, \qquad S_M = U_1 + \cdots + U_M.$$

- *Strong Law of Large Numbers*: $\displaystyle\frac{S_M}{M} \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_\pi[U]$ (a.s. = almost surely)

- For any measurable function f also $\displaystyle\frac{1}{M}\sum_{k=1}^{M} f(U_k) \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_\pi[f(U)] =: \mathbb{E}_\pi[f]$

- *Central Limit Theorem*: If $\mathbb{E}_\pi[U] = \mu$ and $\mathbb{V}_\pi[U] := \mathbb{E}_\pi\left[(U-\mu)^2\right] = \sigma^2$, then

$$\mathbb{E}_\pi[S_M] = M\mu, \;\; \mathbb{V}_\pi[S_M] = M\sigma^2 \text{ and } S_M^* = \frac{S_M - M\mu}{\sqrt{M}\sigma} \xrightarrow[M\to\infty]{\text{D}} \mathsf{N}(0,1)$$

i.e. the estimate is unbiased, its variance is σ^2/M and the distribution of the normalised RV S_M^* becomes Gaussian as $M \to \infty$.

Monte Carlo Convergence Statements

- Mean square convergence:

$$\mathbb{E}_\pi\left[\left(\frac{S_M}{M}-\mu\right)^2\right]=\mathbb{V}_\pi\left[\frac{S_M}{M}\right]=\frac{\sigma^2}{M}\xrightarrow[M\to\infty]{\text{a.s.}}0.$$

Monte Carlo Convergence Statements

- Mean square convergence:

$$\mathbb{E}_\pi\left[\left(\frac{S_M}{M}-\mu\right)^2\right]=\mathbb{V}_\pi\left[\frac{S_M}{M}\right]=\frac{\sigma^2}{M}\xrightarrow[M\to\infty]{\text{a.s.}}0.$$

- *Chebyshev's Inequality* implies, for any $\epsilon>0$:

$$\mathbb{P}\left(\left|\frac{S_M}{M}-\mu\right|>M^{-1/2+\epsilon}\right)\leq\frac{\sigma^2}{M^{2\epsilon}},$$

(i.e. the probability of the error being $> M^{-1/2+\epsilon}$ converges to zero as $M\to\infty$)

Monte Carlo Convergence Statements

- Mean square convergence:

$$\mathbb{E}_\pi\left[\left(\frac{S_M}{M}-\mu\right)^2\right]=\mathbb{V}_\pi\left[\frac{S_M}{M}\right]=\frac{\sigma^2}{M}\xrightarrow[M\to\infty]{\text{a.s.}}0.$$

- *Chebyshev's Inequality* implies, for any $\epsilon>0$:

$$\mathbb{P}\left(\left|\frac{S_M}{M}-\mu\right|>M^{-1/2+\epsilon}\right)\leq\frac{\sigma^2}{M^{2\epsilon}},$$

(i.e. the probability of the error being $> M^{-1/2+\epsilon}$ converges to zero as $M\to\infty$)

- If $\rho:=\mathbb{E}\left[|U-\mu|^3\right]<\infty$, then the *Berry-Esseen Inequality* gives

$$\left|\mathbb{P}(S_M^*\leq x)-\Phi(x)\right|\leq\frac{\rho}{2\sigma^3\sqrt{M}},$$

where Φ denotes *cumulative density function (CDF)* of $\mathsf{N}(0,1)$.

Monte Carlo Convergence Statements

- Mean square convergence:

$$\mathbb{E}_\pi\left[\left(\frac{S_M}{M}-\mu\right)^2\right]=\mathbb{V}_\pi\left[\frac{S_M}{M}\right]=\frac{\sigma^2}{M}\xrightarrow[M\to\infty]{\text{a.s.}}0.$$

- *Chebyshev's Inequality* implies, for any $\epsilon>0$:

$$\mathbb{P}\left(\left|\frac{S_M}{M}-\mu\right|>M^{-1/2+\epsilon}\right)\leq\frac{\sigma^2}{M^{2\epsilon}},$$

(i.e. the probability of the error being $> M^{-1/2+\epsilon}$ converges to zero as $M\to\infty$)

- If $\rho:=\mathbb{E}\left[|U-\mu|^3\right]<\infty$, then the *Berry-Esseen Inequality* gives

$$\left|\mathbb{P}(S_M^*\leq x)-\Phi(x)\right|\leq\frac{\rho}{2\sigma^3\sqrt{M}},$$

where Φ denotes *cumulative density function (CDF)* of $\mathsf{N}(0,1)$.

- Using Berry-Esseen, the asymptotic 95% confidence interval for S_M/M is

$$0.95-\frac{\rho}{\sigma^3\sqrt{M}}\leq\mathbb{P}\left(\mu\in\left[\frac{S_M}{M}-\frac{1.96\sigma}{\sqrt{M}},\frac{S_M}{M}+\frac{1.96\sigma}{\sqrt{M}}\right]\right)\leq 0.95+\frac{\rho}{\sigma^3\sqrt{M}}$$

Problems with plain Monte Carlo

Target measure

$$\pi(\mathrm{d}u) = \frac{1}{Z}\exp(-\Phi(\mathrm{d}u))\,\pi_0(\mathrm{d}u), \qquad \Phi(u) = \tfrac{1}{2}|y - \mathcal{G}(u)|^2_{\Sigma^{-1}}$$

- **Almost never possible** to draw samples u_i directly according to $\pi_{u|y}$ (particularly since normalizing constant Z is usually not known).

Problems with plain Monte Carlo

Target measure

$$\pi(\mathrm{d}u) = \frac{1}{Z}\exp(-\Phi(\mathrm{d}u))\,\pi_0(\mathrm{d}u), \qquad \Phi(u) = \tfrac{1}{2}|y - \mathcal{G}(u)|^2_{\Sigma^{-1}}$$

- **Almost never possible** to draw samples u_i directly according to $\pi_{u|y}$ (particularly since normalizing constant Z is usually not known).

$$\longrightarrow \quad \textbf{Essentially unusable!}$$

Problems with plain Monte Carlo and Alternatives

Target measure

$$\pi(\mathrm{d}u) = \frac{1}{Z}\exp(-\Phi(\mathrm{d}u))\,\pi_0(\mathrm{d}u), \qquad \Phi(u) = \tfrac{1}{2}|y - \mathcal{G}(u)|^2_{\Sigma^{-1}}$$

- **Almost never possible** to draw samples u_i directly according to $\pi_{u|y}$ (particularly since normalizing constant Z is usually not known).

$$\longrightarrow \quad \textbf{Essentially unusable!}$$

Alternatives:

- Rejection sampling
- Importance sampling and Quasi-Monte Carlo
- Markov chain Monte Carlo (MCMC) and multilevel variants
- Ensemble methods (Ensemble Kalman filter, sequential Monte Caro, ...)
- Simple approximations, surrogates, variational approaches, ...

Problems with plain Monte Carlo and Alternatives

Target measure

$$\pi(\mathrm{d}u) = \frac{1}{Z}\exp(-\Phi(\mathrm{d}u))\,\pi_0(\mathrm{d}u), \qquad \Phi(u) = \tfrac{1}{2}|y - \mathcal{G}(u)|^2_{\Sigma^{-1}}$$

- **Almost never possible** to draw samples u_i directly according to $\pi_{u|y}$ (particularly since normalizing constant Z is usually not known).

$$\longrightarrow \quad \textbf{Essentially unusable!}$$

Alternatives:

- Rejection sampling
- Importance sampling and Quasi-Monte Carlo
- Markov chain Monte Carlo (MCMC) and multilevel variants
- Ensemble methods (Ensemble Kalman filter, sequential Monte Caro, ...)
- Simple approximations, surrogates, variational approaches, ...

All have in common that they need a **good approximation** to $\pi_{u|y}$ as proposal, importance density, surrogate, ... (e.g., reduced order model, lower-dim. approx., ...)

Rejection sampling

Target measure

$$\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\, \pi_0(\mathrm{d}u)$$

Assuming we can sample directly from prior/reference distribtion π_0 on $\mathcal{X}$

Rejection sampling

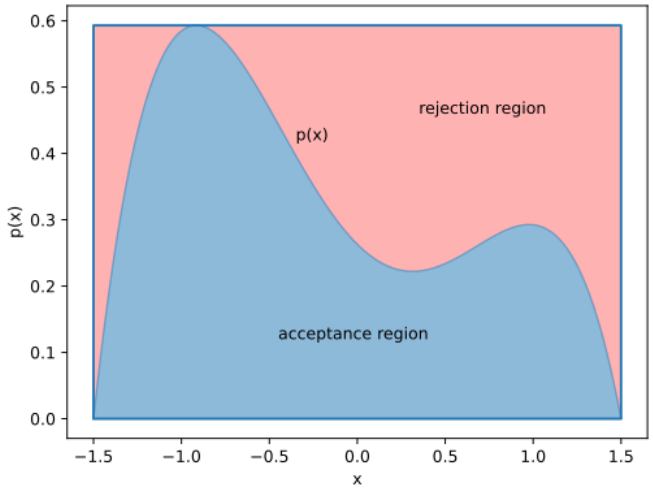
Target measure

$$\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\,\pi_0(\mathrm{d}u)$$

Assuming we can sample directly from prior/reference distribtion π_0 on $\mathcal{X}$

Rejection sampler

1. Draw a sample u according to π_0
2. Draw a sample a according to $\mathrm{U}[0,1]$
3. If
$$a \leq \exp(-\Phi(u)),$$
then accept u as sample, otherwise go back to step one.



Rejection sampling

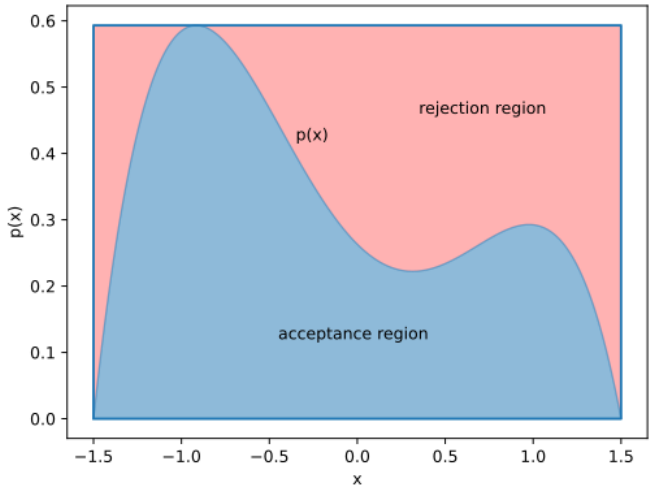
Target measure

$$\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\,\pi_0(\mathrm{d}u)$$

Assuming we can sample directly from prior/reference distribtion π_0 on $\mathcal{X}$

Rejection sampler

1. Draw a sample u according to π_0
2. Draw a sample a according to $\mathrm{U}[0,1]$
3. If
$$a \leq \exp(-\Phi(u)),$$
then accept u as sample, otherwise go back to step one.



The accepted samples then follow target π,

Rejection sampling

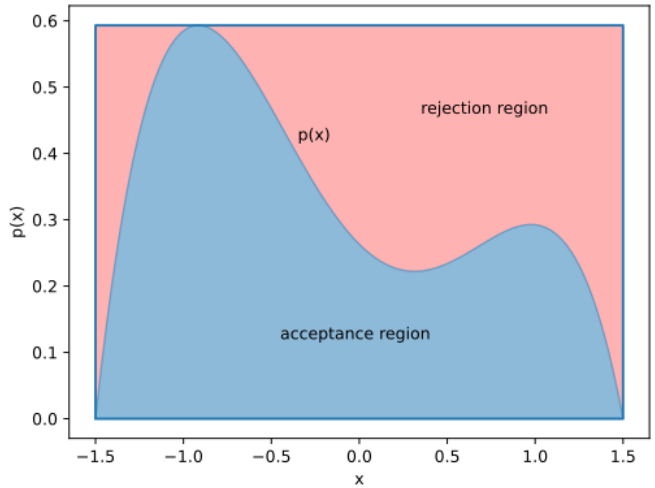
Target measure

$$\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\,\pi_0(\mathrm{d}u)$$

Assuming we can sample directly from prior/reference distribtion π_0 on $\mathcal{X}$

Rejection sampler

1. Draw a sample u according to π_0
2. Draw a sample a according to $\mathrm{U}[0,1]$
3. If
$$a \leq \exp(-\Phi(u)),$$
then accept u as sample, otherwise go back to step one.



The accepted samples then follow target π, **but**

$$\mathbb{E}\left[\#\mathsf{Tries}\right] = \tfrac{1}{Z} \quad (Z \text{ is a measure of how much } \pi_{u|y} \text{ concentrates.})$$

Importance sampling

Target measure

$$\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\, \pi_0(\mathrm{d}u)$$

Assuming we can sample directly from prior/reference distribtion π_0 on $\mathcal{X}$

Importance sampling

Target measure

$$\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\,\pi_0(\mathrm{d}u)$$

Assuming we can sample directly from prior/reference distribtion π_0 on $\mathcal{X}$

Given π dominating importance distribution μ with $\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(u) \propto p(u)$, we have

$$\int_{\mathcal{X}} f(u)\,\pi(\mathrm{d}u) = \int_{\mathcal{X}} f(u)\,\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(u)\,\mu(\mathrm{d}u) = \frac{\int_{\mathcal{X}} f(u)\,p(u)\,\mu(\mathrm{d}u)}{\int_{\mathcal{X}} p(u)\,\mu(\mathrm{d}u)}$$

Importance sampling

Target measure

$$\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\,\pi_0(\mathrm{d}u)$$

Assuming we can sample directly from prior/reference distribtion π_0 on $\mathcal{X}$

Given π dominating importance distribution μ with $\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(u) \propto p(u)$, we have

$$\int_{\mathcal{X}} f(u)\,\pi(\mathrm{d}u) = \int_{\mathcal{X}} f(u)\,\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(u)\,\mu(\mathrm{d}u) = \frac{\int_{\mathcal{X}} f(u)\,p(u)\,\mu(\mathrm{d}u)}{\int_{\mathcal{X}} p(u)\,\mu(\mathrm{d}u)}$$

Thus, we could choose, e.g., the prior π_0, then $p(u) = \exp(-\Phi(u))$

Importance sampling

Target measure

$$\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\,\pi_0(\mathrm{d}u)$$

Assuming we can sample directly from prior/reference distribtion π_0 on $\mathcal{X}$

Given π dominating importance distribution μ with $\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(u) \propto p(u)$, we have

$$\int_{\mathcal{X}} f(u)\,\pi(\mathrm{d}u) = \int_{\mathcal{X}} f(u)\,\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(u)\,\mu(\mathrm{d}u) = \frac{\int_{\mathcal{X}} f(u)\,p(u)\,\mu(\mathrm{d}u)}{\int_{\mathcal{X}} p(u)\,\mu(\mathrm{d}u)}$$

Thus, we could choose, e.g., the prior π_0, then $p(u) = \exp(-\Phi(u))$

Prior-based selfnormalizing importance sampling

Given i.i.d. samples $U_i \sim \pi_0$

$$\mathbb{E}_\pi[f] \approx IS_M := \frac{\sum_{i=1}^M f(U_i)\,\exp(-\Phi(U_i))}{\sum_{i=1}^M \exp(-\Phi(U_i))}.$$

Efficiency of importance sampling

Selfnormalizing importance sampling

Given i.i.d. samples $U_i \sim \mu$ with $\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(u) \propto p(u)$

$$\mathbb{E}_\pi[f] \approx IS_M := \frac{\sum_{i=1}^M f(U_i)\, p(U_i)}{\sum_{i=1}^M f(U_i)\, p(U_i)}.$$

Efficiency of importance sampling

Selfnormalizing importance sampling

Given i.i.d. samples $U_i \sim \mu$ with $\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(u) \propto p(u)$

$$\mathbb{E}_\pi[f] \approx IS_M := \frac{\sum_{i=1}^M f(U_i)\, p(U_i)}{\sum_{i=1}^M f(U_i)\, p(U_i)}.$$

- **Strong law of large numbers + Slutzky's theorem:**

$$IS_M \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_\pi[f]$$

Efficiency of importance sampling

Selfnormalizing importance sampling

Given i.i.d. samples $U_i \sim \mu$ with $\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(u) \propto p(u)$

$$\mathbb{E}_\pi[f] \approx IS_M := \frac{\sum_{i=1}^M f(U_i)\, p(U_i)}{\sum_{i=1}^M f(U_i)\, p(U_i)}.$$

- **Strong law of large numbers + Slutzky's theorem:**

$$IS_M \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_\pi[f]$$

- **Central limit theorem + Slutzky's theorem:** For $f : \mathcal{X} \to \mathbb{R}$

$$\sqrt{M}(IS_M - \mathbb{E}_\pi[f]) \xrightarrow[M\to\infty]{\mathcal{D}} \mathrm{N}(0, \sigma^2_{f,p}),$$

where

$$\sigma^2_{f,p} = \mathbb{E}_\pi\left[\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\,(f - \mathbb{E}_\pi[f])^2\right]$$

Efficiency of importance sampling

Prior-based selfnormalizing importance sampling

Given i.i.d. samples $U_i \sim \pi_0$

$$\mathbb{E}_\pi[f] \approx IS_M := \frac{\sum_{i=1}^M f(U_i)\,\exp(-\Phi(U_i))}{\sum_{i=1}^M \exp(-\Phi(U_i))}.$$

- **Strong law of large numbers + Slutzky's theorem:**

$$IS_M \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_\pi[f]$$

- **Central limit theorem + Slutzky's theorem:** For $f : \mathcal{X} \to \mathbb{R}$

$$\sqrt{M}\Big(IS_M - \mathbb{E}_\pi[f]\Big) \xrightarrow[M\to\infty]{\mathcal{D}} \mathrm{N}(0, \sigma_{f,p}^2),$$

where

$$\sigma_{f,p}^2 = \frac{1}{Z^2}\mathbb{E}_{\pi_0}\left[\exp(-2\Phi)\,(f - \mathbb{E}_\pi[f])^2\right]$$

Efficiency of importance sampling

Prior-based selfnormalizing importance sampling

Given i.i.d. samples $U_i \sim \pi_0$

$$\mathbb{E}_\pi[f] \approx IS_M := \frac{\sum_{i=1}^M f(U_i)\,\exp(-\Phi(U_i))}{\sum_{i=1}^M \exp(-\Phi(U_i))}.$$

- **Strong law of large numbers + Slutzky's theorem:**

$$IS_M \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_\pi[f]$$

- **Central limit theorem + Slutzky's theorem:** For $f : \mathcal{X} \to \mathbb{R}$

$$\sqrt{M}\Big(IS_M - \mathbb{E}_\pi[f]\Big) \xrightarrow[M\to\infty]{\mathcal{D}} \mathrm{N}(0, \sigma_{f,p}^2),$$

where

$$\sigma_{f,p}^2 = \frac{1}{Z^2}\mathbb{E}_{\pi_0}\left[\exp(-2\Phi)\,(f - \mathbb{E}_\pi[f])^2\right]$$

Again, this variance **grows very fast** when the **posterior concentrates!**

Efficiency of importance sampling (beyond prior-based)

- When posterior concentrates, prior π_0 not a good importance distribution. A better choice for μ is the Laplace approximation.
  (computable via numerical optimization if $\nabla^2\Phi$ is accessible).
- In that case, concentration even helps: Variance $\sigma^2_{f,p}$ of the selfnormalizing ratio estimator even reduces with $m \to \infty$ with a rate arbitrarily close to $1/2$ [Schillings, Sprungk, Wacker, 2020].

Efficiency of importance sampling (beyond prior-based)

- When posterior concentrates, prior π_0 not a good importance distribution. A better choice for μ is the Laplace approximation.
  (computable via numerical optimization if $\nabla^2\Phi$ is accessible).
- In that case, concentration even helps: Variance $\sigma^2_{f,p}$ of the selfnormalizing ratio estimator even reduces with $m \to \infty$ with a rate arbitrarily close to $1/2$ [Schillings, Sprungk, Wacker, 2020].
- In principle, any approximation to the (unnormalized) $\pi_{u|y}$ can be used, e.g., surrrogates based on low-rank tensor approximation, such as the TT-CD [Dolgov, Anaya-Izquierdo, Fox, **RS**, 2020] or DIRT sampler [Cui, Dolgov, 2022].

Efficiency of importance sampling (beyond prior-based)

- When posterior concentrates, prior π_0 not a good importance distribution. A better choice for μ is the Laplace approximation.
  (computable via numerical optimization if $\nabla^2\Phi$ is accessible).
- In that case, concentration even helps: Variance $\sigma^2_{f,p}$ of the selfnormalizing ratio estimator even reduces with $m \to \infty$ with a rate arbitrarily close to $1/2$ [Schillings, Sprungk, Wacker, 2020].
- In principle, any approximation to the (unnormalized) $\pi_{u|y}$ can be used, e.g., surrrogates based on low-rank tensor approximation, such as the TT-CD [Dolgov, Anaya-Izquierdo, Fox, **RS**, 2020] or DIRT sampler [Cui, Dolgov, 2022].
- The optimal importance density is in fact $p_{\text{opt}} \propto |f|\,\pi_{u|y}$. The DIRT sampler can in fact even obtain arbitrarily accurate approximations of p_{opt} at very low cost and in high dimensions [Cui, Dolgov, **RS**, 2024].

Efficiency of importance sampling (beyond prior-based)

- When posterior concentrates, prior π_0 not a good importance distribution. A better choice for μ is the Laplace approximation.
  (computable via numerical optimization if $\nabla^2\Phi$ is accessible).
- In that case, concentration even helps: Variance $\sigma^2_{f,p}$ of the selfnormalizing ratio estimator even reduces with $m \to \infty$ with a rate arbitrarily close to $1/2$ [Schillings, Sprungk, Wacker, 2020].
- In principle, any approximation to the (unnormalized) $\pi_{u|y}$ can be used, e.g., surrrogates based on low-rank tensor approximation, such as the TT-CD [Dolgov, Anaya-Izquierdo, Fox, **RS**, 2020] or DIRT sampler [Cui, Dolgov, 2022].
- The optimal importance density is in fact $p_{\text{opt}} \propto |f|\,\pi_{u|y}$. The DIRT sampler can in fact even obtain arbitrarily accurate approximations of p_{opt} at very low cost and in high dimensions [Cui, Dolgov, **RS**, 2024].
- Crucial advantage of importance sampling, compared to MCMC (see below): Other quadrature rules with faster asymptotic convergence can be used, such as **quasi-Monte Carlo** [**RS**, Stuart, Teckentrup, 2017], ...
  (i.e., $\mathcal{O}(M)$ w.r.t. number of samples M instead of $\mathcal{O}(M^{1/2})$ for MC)

Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC)

- Rejection sampling generates i.i.d. samples $U_i \sim \pi$, but very inefficient, particularly in high dimensions $\mathcal{X} = \mathbb{R}^n$, $n \gg 1$, and for concentrated π.

Markov chain Monte Carlo (MCMC)

- Rejection sampling generates i.i.d. samples $U_i \sim \pi$, but very inefficient, particularly in high dimensions $\mathcal{X} = \mathbb{R}^n$, $n \gg 1$, and for concentrated π.

- Importance sampling uses i.i.d. samples $U_i \sim \mu$ to approximate expectations w.r.t. π, efficiency highly depends on choice of μ and can be again be very low in high dimensions $n \gg 1$

Markov chain Monte Carlo (MCMC)

- Rejection sampling generates i.i.d. samples $U_i \sim \pi$, but very inefficient, particularly in high dimensions $\mathcal{X} = \mathbb{R}^n$, $n \gg 1$, and for concentrated π.

- Importance sampling uses i.i.d. samples $U_i \sim \mu$ to approximate expectations w.r.t. π, efficiency highly depends on choice of μ and can be again be very low in high dimensions $n \gg 1$

- We focus on a third sampling technique which generates correlated samples U_i which follow approximately π, i.e., $U_i \to \pi$ as $i \to \infty$

Markov chain Monte Carlo (MCMC)

- Rejection sampling generates i.i.d. samples $U_i \sim \pi$, but very inefficient, particularly in high dimensions $\mathcal{X} = \mathbb{R}^n$, $n \gg 1$, and for concentrated π.

- Importance sampling uses i.i.d. samples $U_i \sim \mu$ to approximate expectations w.r.t. π, efficiency highly depends on choice of μ and can be again be very low in high dimensions $n \gg 1$

- We focus on a third sampling technique which generates correlated samples U_i which follow approximately π, i.e., $U_i \to \pi$ as $i \to \infty$

- **Main tool:** Generate a Markov chain $(U_i)_{i \in \mathbb{N}}$ in $\mathcal{X}$ with π as its stationary or limit distribution

Markov chain Monte Carlo (MCMC)

- Rejection sampling generates i.i.d. samples $U_i \sim \pi$, but very inefficient, particularly in high dimensions $\mathcal{X} = \mathbb{R}^n$, $n \gg 1$, and for concentrated π.

- Importance sampling uses i.i.d. samples $U_i \sim \mu$ to approximate expectations w.r.t. π, efficiency highly depends on choice of μ and can be again be very low in high dimensions $n \gg 1$

- We focus on a third sampling technique which generates correlated samples U_i which follow approximately π, i.e., $U_i \to \pi$ as $i \to \infty$

- **Main tool:** Generate a Markov chain $(U_i)_{i \in \mathbb{N}}$ in $\mathcal{X}$ with π as its stationary or limit distribution

- Resulting Markov chain Monte Carlo estimator $\frac{1}{M} \sum_{i=1}^{M} f(U_i)$ for $\mathbb{E}_\pi[f]$ has, in general, a (mildly) dimension-dependent efficiency but also dimension independent MCMC possible.

Markov chain Basics

A sequence of random variables $(U_i)_{i \in \mathbb{N}_0}$ is a (time-homogeneous) **Markov chain** if there exists a stochastic kernel K such that for each $i \in \mathbb{N}$ and $A \subseteq \mathcal{X}$

$$\mathbb{P}(U_{i+1} \in A \mid U_i = u_i, \ldots, U_0 = u_0) \;=\; K(u_i, A)$$

The kernel K is called transition kernel.

Andrey A. Markow
(1856 – 1922)

Markov chain Basics

A sequence of random variables $(U_i)_{i\in\mathbb{N}_0}$ is a (time-homogeneous) **Markov chain** if there exists a stochastic kernel K such that for each $i \in \mathbb{N}$ and $A \subseteq \mathcal{X}$

$$\mathbb{P}(U_{i+1} \in A \mid U_i = u_i, \ldots, U_0 = u_0) \;=\; K(u_i, A)$$

The kernel K is called transition kernel.

Andrey A. Markow
(1856 – 1922)

Given an initial distribution ν for U_0 we have $U_i \sim \nu K^i$ where

$$\nu K^i(A) := \int_{\mathcal{X}} K^i(u, A)\, \nu(\mathrm{d}u), \qquad K^i(u, A) := \int_{\mathcal{X}} K(v, A)\, K^{i-1}(u, \mathrm{d}v),$$

i.e., $K^i(u, A) = \mathbb{P}(U_{k+i} \in A \mid U_k = u, U_{k-1} = u_{k-1}, \ldots, U_0 = u_0)$ denotes the i-step transition kernel.

Invariance and Reversibility

Definition. A probability measure $\pi \in \mathcal{P}(\mathcal{X})$ is an **invariant measure** of a Markov chain with transition kernel K, if $\pi = \pi K$.

Invariant measure means: $U_i \sim \pi \;\Rightarrow U_{i+1} \sim \pi$

Invariance and Reversibility

Definition. A probability measure $\pi \in \mathcal{P}(\mathcal{X})$ is an **invariant measure** of a Markov chain with transition kernel K, if $\pi = \pi K$.

Invariant measure means: $U_i \sim \pi \Rightarrow U_{i+1} \sim \pi$

Definition. A Markov chain is **stationary** if $U_0 \sim \pi$ and π is invariant.

Invariance and Reversibility

Definition. A probability measure $\pi \in \mathcal{P}(\mathcal{X})$ is an **invariant measure** of a Markov chain with transition kernel K, if $\pi = \pi K$.

Invariant measure means: $U_i \sim \pi \Rightarrow U_{i+1} \sim \pi$

Definition. A Markov chain is **stationary** if $U_0 \sim \pi$ and π is invariant.

Definition. A transition kernel K is **reversible** w.r.t. a probability measure $\pi \in \mathcal{P}(\mathcal{X})$ if for any $A, B \subseteq \mathcal{X}$

$$\int_A K(u, B)\, \pi(\mathrm{d}u) = \int_B K(u, A)\, \pi(\mathrm{d}u)$$

$\Rightarrow$ the dynamics of a stationary Markov chain is time-reversible: $(U_i, U_{i+1}) \sim (U_{i+1}, U_i)$

Invariance and Reversibility

Definition. A probability measure $\pi \in \mathcal{P}(\mathcal{X})$ is an **invariant measure** of a Markov chain with transition kernel K, if $\pi = \pi K$.

Invariant measure means: $U_i \sim \pi \Rightarrow U_{i+1} \sim \pi$

Definition. A Markov chain is **stationary** if $U_0 \sim \pi$ and π is invariant.

Definition. A transition kernel K is **reversible** w.r.t. a probability measure $\pi \in \mathcal{P}(\mathcal{X})$ if for any $A, B \subseteq \mathcal{X}$

$$\int_A K(u, B)\, \pi(\mathrm{d}u) = \int_B K(u, A)\, \pi(\mathrm{d}u)$$

$\Rightarrow$ the dynamics of a stationary Markov chain is time-reversible: $(U_i, U_{i+1}) \sim (U_{i+1}, U_i)$

Proposition. Reversibility of K w.r.t. π implies invariance of π.

Proof. $\pi(A) = \int_A K(u, \mathcal{X})\, \pi(\mathrm{d}u) = \int_{\mathcal{X}} K(u, A)\, \pi(\mathrm{d}u) = \pi K(A)$

Ergodicity

Definition. A Markov chain with transition kernel K and invariant measure π is called **ergodic**, if

$$\lim_{i \to \infty} d_{\mathsf{TV}}(K^i(x, \cdot), \pi) = 0 \qquad \forall x \in \mathcal{X}.$$

$\Rightarrow$ Ergodicity is **not a given!**

Ergodicity

Definition. A Markov chain with transition kernel K and invariant measure π is called **ergodic**, if

$$\lim_{i \to \infty} d_{\mathsf{TV}}(K^i(x, \cdot), \pi) = 0 \qquad \forall x \in \mathcal{X}.$$

$\Rightarrow$ Ergodicity is **not a given!**

Definition. A transition kernel K is π-**irreducible** if for any $A \subset \mathcal{X}$ with $\pi(A) > 0$ and any $x \in \mathcal{X}$ there exists an $i \in \mathbb{N}$ such that $K^i(x, A) > 0$

$\Rightarrow$ Markov chain does not get stuck in a subdomain.

Ergodicity

Definition. A Markov chain with transition kernel K and invariant measure π is called **ergodic**, if

$$\lim_{i\to\infty} d_{\mathsf{TV}}(K^i(x,\cdot),\pi) = 0 \qquad \forall x \in \mathcal{X}.$$

$\Rightarrow$ Ergodicity is **not a given!**

Definition. A transition kernel K is π-**irreducible** if for any $A \subset \mathcal{X}$ with $\pi(A) > 0$ and any $x \in \mathcal{X}$ there exists an $i \in \mathbb{N}$ such that $K^i(x, A) > 0$

$\Rightarrow$ Markov chain does not get stuck in a subdomain.

Definition. A transition kernel K is p-**periodic** if there exists $A_1, \ldots, A_p \subset \mathcal{X}$ such that for $j = 1, \ldots, p$

$$K(u, A_{j+1}) = 1 \qquad \forall u \in A_j$$

where $A_{p+1} := A_1$. If K is not p-periodic for any $p \in \mathbb{N}$, then K is **aperiodic**.

$\Rightarrow$ Markov chain does not "oscillate".

Theorem

Let the transition kernel K have invariant measure π and let $(U_i)_{i \in \mathbb{N}_0}$ be a Markov chain starting at $U_0 = u$ with transition kernel K.

Theorem

Let the transition kernel K have invariant measure π and let $(U_i)_{i\in\mathbb{N}_0}$ be a Markov chain starting at $U_0 = u$ with transition kernel K.

1. If K is irreducible, then for π-almost all $u \in \mathcal{X}$ we have a Strong Law of Large Numbers:

$$\frac{1}{M}\sum_{i=1}^{M} f(U_i) \xrightarrow[M\to\infty]{\text{a.s.}} \int_{\mathcal{X}} f(u)\,\pi(\mathrm{d}u), \qquad f \in L^1_\pi(\mathcal{X}).$$

Theorem

Let the transition kernel K have invariant measure π and let $(U_i)_{i\in\mathbb{N}_0}$ be a Markov chain starting at $U_0 = u$ with transition kernel K.

1. If K is irreducible, then for π-almost all $u \in \mathcal{X}$ we have a Strong Law of Large Numbers:

$$\frac{1}{M}\sum_{i=1}^{M} f(U_i) \xrightarrow[M\to\infty]{\text{a.s.}} \int_{\mathcal{X}} f(u)\,\pi(\mathrm{d}u), \qquad f \in L^1_\pi(\mathcal{X}).$$

2. If K is aperiodic and irreducible, then for π-almost all $u \in \mathcal{X}$ we have

$$\lim_{i\to\infty} d_{\mathsf{TV}}\left(K^i(x),\ \pi\right) = 0$$

Theorem

Let the transition kernel K have invariant measure π and let $(U_i)_{i\in\mathbb{N}_0}$ be a Markov chain starting at $U_0 = u$ with transition kernel K.

1. If K is irreducible, then for π-almost all $u \in \mathcal{X}$ we have a Strong Law of Large Numbers:

$$\frac{1}{M}\sum_{i=1}^{M} f(U_i) \xrightarrow[M\to\infty]{\text{a.s.}} \int_{\mathcal{X}} f(u)\,\pi(\mathrm{d}u), \qquad f \in L^1_\pi(\mathcal{X}).$$

2. If K is aperiodic and irreducible, then for π-almost all $u \in \mathcal{X}$ we have

$$\lim_{i\to\infty} d_{\mathsf{TV}}\left(K^i(x),\ \pi\right) = 0$$

3. If K is irreducible and for each $u \in \mathcal{X}$ und $A \subseteq \mathcal{X}$ with $\pi(A) > 0$

$$\mathbb{P}\left(U_i \in A \text{ infinitely often} \mid U_0 = u\right) = 1, \qquad \text{('Harris recurrence')}$$

then the above statements hold for each $u \in \mathcal{X}$.

Efficiency of MCMC

- Consider $f : \mathcal{X} \to \mathbb{R}$ and a stationary, reversible Markov chain $(U_i)_{i \in \mathbb{N}_0}$ with $U_0 \sim \pi$
- For the L^2-error of the Monte Carlo estimator $S_M = \frac{1}{M} \sum_{i=1}^{M} f(U_i)$ we have

Efficiency of MCMC

- Consider $f : \mathcal{X} \to \mathbb{R}$ and a stationary, reversible Markov chain $(U_i)_{i \in \mathbb{N}_0}$ with $U_0 \sim \pi$
- For the L^2-error of the Monte Carlo estimator $S_M = \frac{1}{M} \sum_{i=1}^{M} f(U_i)$ we have

$$\begin{aligned}
\mathbb{E}\left[|S_M - \mathbb{E}_\pi[f]|^2 \right] &= \mathbb{E}\left[\left(\frac{1}{M} \sum_{i=1}^{M} (f(U_i) - \mathbb{E}[f(U_i)]) \right)^2 \right] \\
&= \frac{1}{M^2} \sum_{i,j=1}^{M} \mathbb{C}\mathrm{ov}\, f(U_i) f(U_j) \\
&= \frac{1}{M} \mathrm{Var}_\pi f + \frac{2}{M^2} \sum_{i<j \leq M} \mathbb{C}\mathrm{ov}\, f(U_i) f(U_j)
\end{aligned}$$

Efficiency of MCMC

- Consider $f : \mathcal{X} \to \mathbb{R}$ and a stationary, reversible Markov chain $(U_i)_{i \in \mathbb{N}_0}$ with $U_0 \sim \pi$
- For the L^2-error of the Monte Carlo estimator $S_M = \frac{1}{M} \sum_{i=1}^{M} f(U_i)$ we have

$$\begin{aligned}
\mathbb{E}\left[\left| S_M - \mathbb{E}_\pi\left[f\right] \right|^2 \right] &= \mathbb{E}\left[\left(\frac{1}{M} \sum_{i=1}^{M} \left(f(U_i) - \mathbb{E}\left[f(U_i)\right] \right) \right)^2 \right] \\
&= \frac{1}{M^2} \sum_{i,j=1}^{M} \mathbb{C}\mathrm{ov}\, f(U_i) f(U_j) \\
&= \frac{1}{M} \mathrm{Var}_\pi\, f + \frac{2}{M^2} \sum_{i<j\leq M} \mathbb{C}\mathrm{ov}\, f(U_i) f(U_j)
\end{aligned}$$

- Thus, we get as asymptotic error (if finite)

$$\lim_{M\to\infty} M\mathbb{E}\left[\left| S_M - \mathbb{E}_\pi\left[f\right] \right|^2 \right] = \mathrm{Var}_\pi(f) \underbrace{\left[1 + 2 \sum_{j=0}^{\infty} \mathrm{Corr}\left(f(U_1), f(U_{1+j}) \right) \right]}_{\text{integrated autocorrelation time } \mathrm{IACT}_f}$$

Markov Chain Central Limit Theorem

For a reversible, stationary Markov chain $(U_i)_{i\in\mathbb{N}_0}$ with invariant measure π and an $f \in L^2_\pi(\mathcal{X})$ with finite

$$\mathrm{IACT}_f = 1 + 2\sum_{j=0}^{\infty} \mathrm{Corr}\left(f(U_1), f(U_{1+j})\right) < \infty$$

we have $$\sqrt{M}\Big(S_M - \mathbb{E}_\pi\left[f\right]\Big) \xrightarrow[M\to\infty]{\mathcal{D}} \mathrm{N}\Big(0, \mathrm{Var}_\pi(f)\,\mathrm{IACT}_f\Big)$$

[1] C. G. Geyer. Practical Markov Chain Monte Carlo. *Statist. Sci.* 7(4):473–483, 1992.

Markov Chain Central Limit Theorem

For a reversible, stationary Markov chain $(U_i)_{i\in\mathbb{N}_0}$ with invariant measure π and an $f \in L^2_\pi(\mathcal{X})$ with finite

$$\mathrm{IACT}_f = 1 + 2\sum_{j=0}^{\infty} \mathrm{Corr}\left(f(U_1), f(U_{1+j})\right) < \infty$$

we have $$\sqrt{M}\Big(S_M - \mathbb{E}_\pi\left[f\right]\Big) \xrightarrow[M\to\infty]{\mathcal{D}} \mathrm{N}\Big(0, \mathrm{Var}_\pi(f)\,\mathrm{IACT}_f\Big)$$

- Intrinsic accuracy estimate and allows for confidence intervals for $\mathbb{E}_\pi\left[f\right]$

[1] C. G. Geyer. Practical Markov Chain Monte Carlo. *Statist. Sci.* 7(4):473–483, 1992.

Markov Chain Central Limit Theorem

For a reversible, stationary Markov chain $(U_i)_{i\in\mathbb{N}_0}$ with invariant measure π and an $f \in L^2_\pi(\mathcal{X})$ with finite

$$\mathrm{IACT}_f = 1 + 2\sum_{j=0}^{\infty} \mathrm{Corr}\left(f(U_1), f(U_{1+j})\right) < \infty$$

we have $$\sqrt{M}\Big(S_M - \mathbb{E}_\pi\left[f\right]\Big) \xrightarrow[M\to\infty]{\mathcal{D}} \mathrm{N}\Big(0, \mathrm{Var}_\pi(f)\,\mathrm{IACT}_f\Big)$$

- Intrinsic accuracy estimate and allows for confidence intervals for $\mathbb{E}_\pi\left[f\right]$
- In practice, given a finite path $(u_i)_{i=1,\ldots,M}$ of the Markov chain, we estimate the autocorrelation empirically

$$\mathrm{Corr}\left(f(U_1), f(U_{1+j})\right) \approx \frac{\frac{1}{M-j}\sum_{i=1}^{M-j}\left(f(u_i) - s_M\right)\,\left(f(u_{i+j}) - s_M\right)}{\frac{1}{M-1}\sum_{i=1}^{M}\left(f(u_i) - s_M\right)^2}$$

and truncate the series above after sufficiently many terms[1]

[1] C. G. Geyer. Practical Markov Chain Monte Carlo. *Statist. Sci.* 7(4):473–483, 1992.

Geometric ergodicity

Definition. A Markov chain with transition kernel K and invariant measure π is **geometrically ergodic** if there exists a $C\colon \mathcal{X} \to [0,\infty)$ and $\beta < 1$ such that

$$d_{\mathsf{TV}}\left(K^i(u,\cdot),\pi\right) \leq C(u)\,\beta^i \qquad \forall u \in \mathcal{X}.$$

Geometric ergodicity

Definition. A Markov chain with transition kernel K and invariant measure π is **geometrically ergodic** if there exists a $C\colon \mathcal{X} \to [0,\infty)$ and $\beta < 1$ such that

$$d_{\mathsf{TV}}\left(K^i(u,\cdot),\pi\right) \leq C(u)\,\beta^i \qquad \forall u \in \mathcal{X}.$$

Theorem

If $(U_i)_{i\in\mathbb{N}_0}$ is a reversible and geometrically ergodic Markov chain with invariant measure π, then for any $f \in L^2_\pi(\mathcal{X})$ we have $\mathrm{IACT}_f < \infty$ and thus for any initial distribution $U_0 \sim \nu$ there holds

$$\sqrt{M}\Big(S_M - \mathbb{E}_\pi\left[f\right]\Big) \xrightarrow[M\to\infty]{\mathcal{D}} \mathrm{N}\Big(0, \mathrm{Var}_\pi(f)\,\mathrm{IACT}_f\Big)$$

Geometric ergodicity

Definition. A Markov chain with transition kernel K and invariant measure π is **geometrically ergodic** if there exists a $C\colon \mathcal{X} \to [0,\infty)$ and $\beta < 1$ such that

$$d_{\mathsf{TV}}\left(K^i(u,\cdot),\pi\right) \leq C(u)\,\beta^i \qquad \forall u \in \mathcal{X}.$$

Theorem

If $(U_i)_{i\in\mathbb{N}_0}$ is a reversible and geometrically ergodic Markov chain with invariant measure π, then for any $f \in L^2_\pi(\mathcal{X})$ we have $\mathrm{IACT}_f < \infty$ and thus for any initial distribution $U_0 \sim \nu$ there holds

$$\sqrt{M}\Big(S_M - \mathbb{E}_\pi\left[f\right]\Big) \xrightarrow[M\to\infty]{\mathcal{D}} \mathrm{N}\Big(0, \mathrm{Var}_\pi(f)\,\mathrm{IACT}_f\Big)$$

Geometric ergodicity is equivalent to a functional analytic property of transition kernels and their related transition operators.

Spectral gaps

Definition. Given a π-invariant transition kernel K on $\mathcal{X}$ the associated **transition operator** $\mathrm{K}\colon L^2_\pi \to L^2_\pi$ is defined by

$$\mathrm{K}f(u) := \int_{\mathcal{X}} f(v)\, K(u, \mathrm{d}v)$$

Spectral gaps

Definition. Given a π-invariant transition kernel K on $\mathcal{X}$ the associated **transition operator** $\mathrm{K}\colon L^2_\pi \to L^2_\pi$ is defined by

$$\mathrm{K}f(u) := \int_{\mathcal{X}} f(v)\, K(u, \mathrm{d}v)$$

If K is π-reversible, then K is self-adjoint. Moreover, by construction K^i is the transition operator associated to K^i.

Spectral gaps

Definition. Given a π-invariant transition kernel K on $\mathcal{X}$ the associated **transition operator** $\mathrm{K}\colon L^2_\pi \to L^2_\pi$ is defined by

$$\mathrm{K}f(u) := \int_{\mathcal{X}} f(v)\, K(u, \mathrm{d}v)$$

If K is π-reversible, then K is self-adjoint. Moreover, by construction K^i is the transition operator associated to K^i.

Definition. Given a π-reversible transition kernel K on $\mathcal{X}$ with associated transition operator $\mathrm{K}\colon L^2_\pi \to L^2_\pi$ we define the **spectral gap** by

$$\gamma(K) := 1 - \sup_{\|f\|_{L^2_\pi}=1} \|\mathrm{K}f - \mathbb{E}_\pi[f]\|_{L^2_\pi} \in [0, 1)$$

Spectral gaps

Definition. Given a π-invariant transition kernel K on $\mathcal{X}$ the associated **transition operator** $\mathrm{K}\colon L^2_\pi \to L^2_\pi$ is defined by

$$\mathrm{K}f(u) := \int_{\mathcal{X}} f(v)\, K(u, \mathrm{d}v)$$

If K is π-reversible, then K is self-adjoint. Moreover, by construction K^i is the transition operator associated to K^i.

Definition. Given a π-reversible transition kernel K on $\mathcal{X}$ with associated transition operator $\mathrm{K}\colon L^2_\pi \to L^2_\pi$ we define the **spectral gap** by

$$\gamma(K) := 1 - \sup_{\|f\|_{L^2_\pi}=1} \|\mathrm{K}f - \mathbb{E}_\pi[f]\|_{L^2_\pi} \in [0,1)$$

Theorem. If $\gamma(K) > 0$, then

$$\mathrm{IACT}_f \leq \frac{2}{\gamma(K)} \quad \text{and} \quad d_{\mathrm{TV}}\left(K^i(u,\cdot), \pi\right) \leq C(1-\gamma(K))^i.$$

Metropolis-Hastings Algorithm

The Metropolis–Hastings (MH) algorithm[2,3]

Let $\pi \in \mathcal{P}(\mathbb{R}^n)$ have unnormalized density $\pi : \mathbb{R}^n \to [0, \infty)$.

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21(6):1087–1092, 1953.

[3] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57(1):97–109, 1970.

The Metropolis–Hastings (MH) algorithm[2,3]

Let $\pi \in \mathcal{P}(\mathbb{R}^n)$ have unnormalized density $\pi\colon \mathbb{R}^n \to [0,\infty)$.

Transition mechanism

Given current state $U_i = u$, generate the next state as follows

1. Draw v according to chosen proposal kernel $P(u,\cdot)$ where P admits density p such that

$$P(u,A) = \int_A p(u,v)\,\mathrm{d}v.$$

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21(6):1087–1092, 1953.

[3] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57(1):97–109, 1970.

The Metropolis–Hastings (MH) algorithm[2,3]

Let $\pi \in \mathcal{P}(\mathbb{R}^n)$ have unnormalized density $\pi\colon \mathbb{R}^n \to [0,\infty)$.

Transition mechanism

Given current state $U_i = u$, generate the next state as follows

1. Draw v according to chosen proposal kernel $P(u,\cdot)$ where P admits density p such that
$$P(u,A) = \int_A p(u,v)\,\mathrm{d}v.$$
2. Draw a according to $\mathrm{U}[0,1]$

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21(6):1087–1092, 1953.

[3] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57(1):97–109, 1970.

The Metropolis–Hastings (MH) algorithm[2,3]

Let $\pi \in \mathcal{P}(\mathbb{R}^n)$ have unnormalized density $\pi\colon \mathbb{R}^n \to [0,\infty)$.

Transition mechanism

Given current state $U_i = u$, generate the next state as follows

1. Draw v according to chosen proposal kernel $P(u,\cdot)$ where P admits density p such that
$$P(u,A) = \int_A p(u,v)\,\mathrm{d}v.$$
2. Draw a according to $\mathrm{U}[0,1]$
3. If
$$a \leq \alpha(u,v) := \min\left\{1, \frac{\pi(v)\ p(v,u)}{\pi(u)\ p(u,v)}\right\}$$
then set $U_{i+1} = v$ (accept) otherwise set $U_{i+1} = u$ (reject).

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21(6):1087–1092, 1953.

[3] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57(1):97–109, 1970.

Remarks

- By construction MH algorithm generates paths of a π-reversible Markov chain.

Remarks

- By construction MH algorithm generates paths of a π-reversible Markov chain.
- The corresponding transition kernel is given by

$$K(u, A) = \int_A \alpha(u, v)\ P(u, \mathrm{d}v) + \Bigg(\underbrace{1 - \int_{\mathcal{X}} \alpha(u, v)\ P(u, \mathrm{d}v)}_{\text{rejection probability } r(u)} \Bigg)\ \mathbf{1}_A(u)$$

Remarks

- By construction MH algorithm generates paths of a π-reversible Markov chain.
- The corresponding transition kernel is given by

$$K(u,A) = \int_A \alpha(u,v)\ P(u,\mathrm{d}v) + \Bigg(\underbrace{1 - \int_{\mathcal{X}} \alpha(u,v)\ P(u,\mathrm{d}v)}_{\text{rejection probability } r(u)} \Bigg)\ \mathbf{1}_A(u)$$

- MH algorithm basically implements a Markov chain rejection sampling and corrects "wrong" proposal chain with kernel P to achieve π-invariance.

Remarks

- By construction MH algorithm generates paths of a π-reversible Markov chain.
- The corresponding transition kernel is given by

$$K(u, A) = \int_A \alpha(u, v)\ P(u, \mathrm{d}v) + \Bigg(\underbrace{1 - \int_{\mathcal{X}} \alpha(u, v)\ P(u, \mathrm{d}v)}_{\text{rejection probability } r(u)} \Bigg)\ \mathbf{1}_A(u)$$

- MH algorithm basically implements a Markov chain rejection sampling and corrects "wrong" proposal chain with kernel P to achieve π-invariance.
- Efficiency of the MH algorithm depends on good choice of proposal kernel P.

Remarks

- By construction MH algorithm generates paths of a π-reversible Markov chain.
- The corresponding transition kernel is given by

$$K(u, A) = \int_A \alpha(u, v)\ P(u, \mathrm{d}v) + \Bigg(\underbrace{1 - \int_{\mathcal{X}} \alpha(u, v)\ P(u, \mathrm{d}v)}_{\text{rejection probability } r(u)} \Bigg)\ \mathbf{1}_A(u)$$

- MH algorithm basically implements a Markov chain rejection sampling and corrects "wrong" proposal chain with kernel P to achieve π-invariance.
- Efficiency of the MH algorithm depends on good choice of proposal kernel P.
- Consider first classical proposal kernel P: (Gaussian) random walk proposal

$$P(u) = \mathrm{N}(u, s^2 C) \qquad \Longrightarrow \qquad \alpha(u, v) = \min\left\{1, \frac{\pi(v)\ p(v, u)}{\pi(u)\ p(u, v)}\right\}$$

with tunable stepsize $s > 0$ and proposal covariance C.

Remarks

- By construction MH algorithm generates paths of a π-reversible Markov chain.
- The corresponding transition kernel is given by

$$K(u, A) = \int_A \alpha(u, v)\ P(u, \mathrm{d}v) + \Bigg(\underbrace{1 - \int_{\mathcal{X}} \alpha(u, v)\ P(u, \mathrm{d}v)}_{\text{rejection probability } r(u)} \Bigg)\ \mathbf{1}_A(u)$$

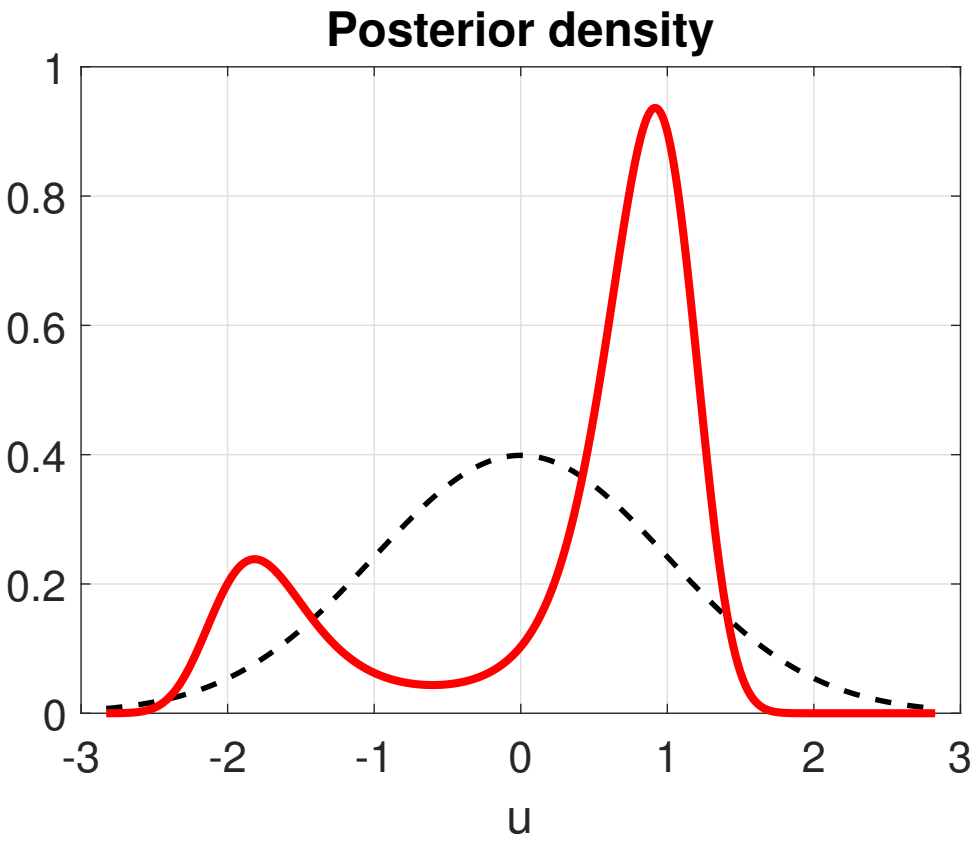
- MH algorithm basically implements a Markov chain rejection sampling and corrects "wrong" proposal chain with kernel P to achieve π-invariance.
- Efficiency of the MH algorithm depends on good choice of proposal kernel P.
- Consider first classical proposal kernel P: (Gaussian) random walk proposal

$$P(u) = \mathrm{N}(u, s^2 C) \qquad \Longrightarrow \qquad \alpha(u, v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$$

with tunable stepsize $s > 0$ and proposal covariance C.

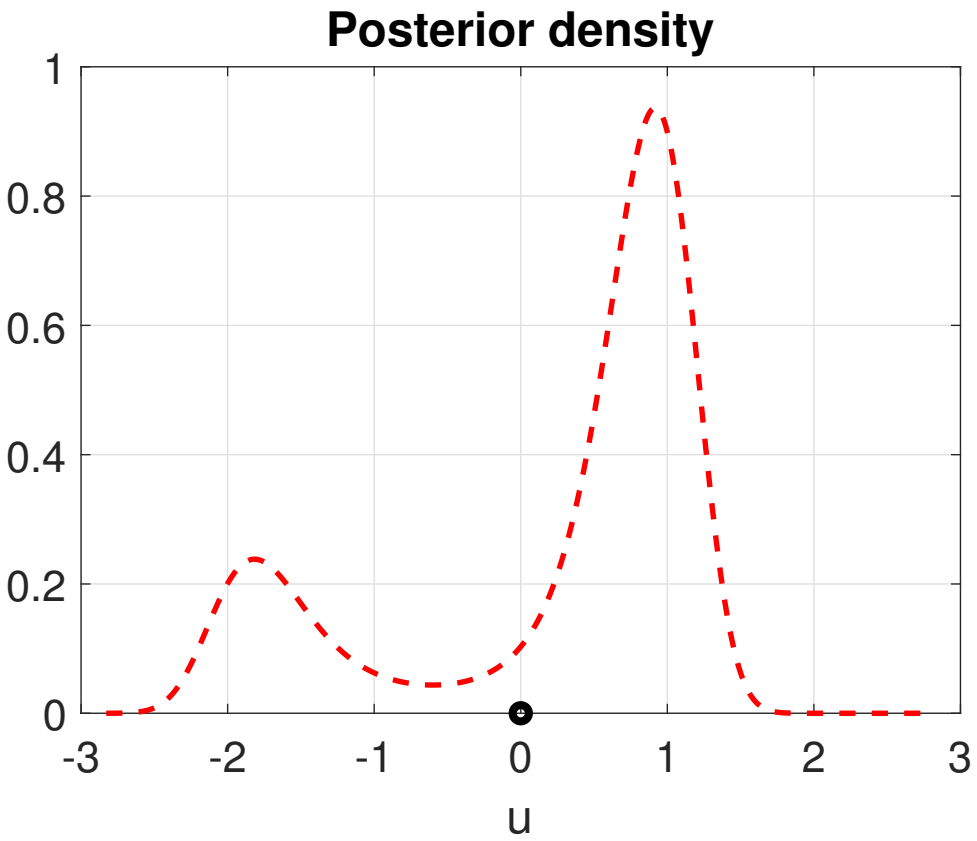
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



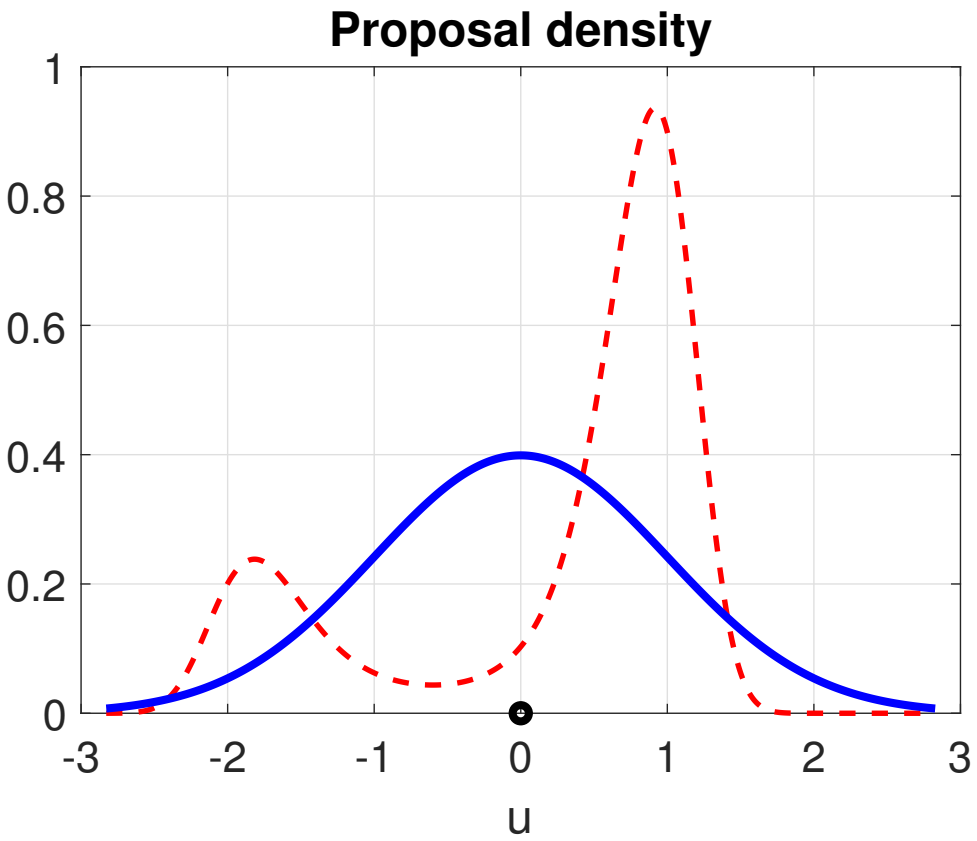
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



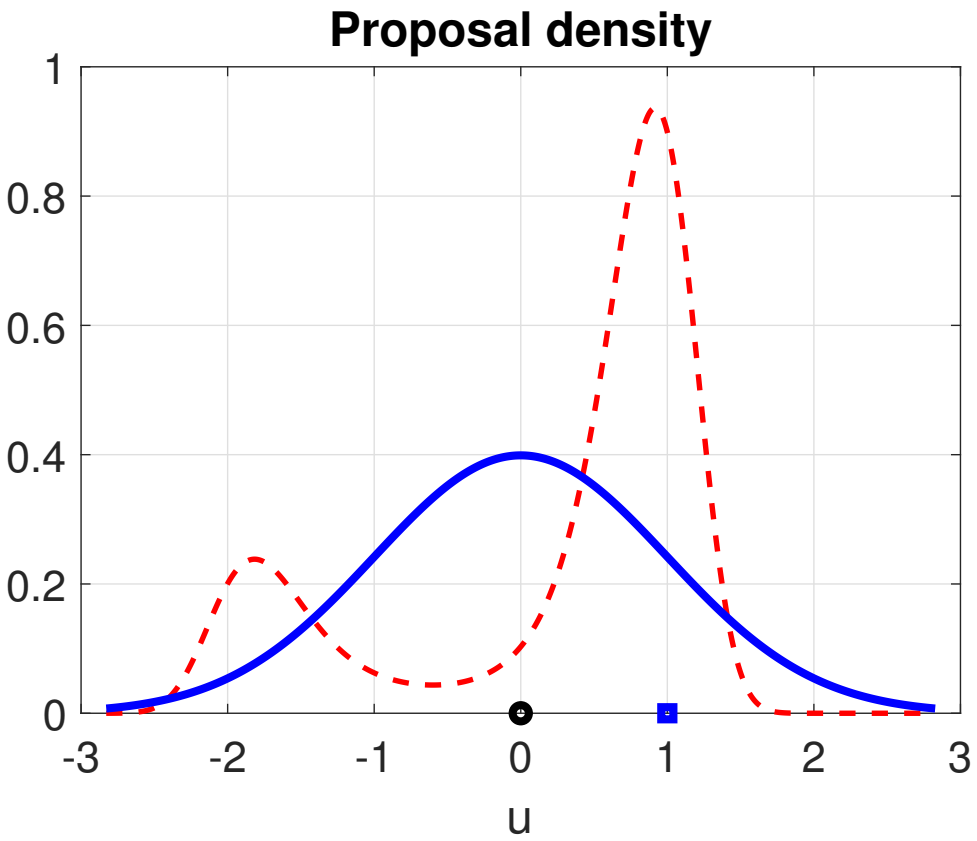
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



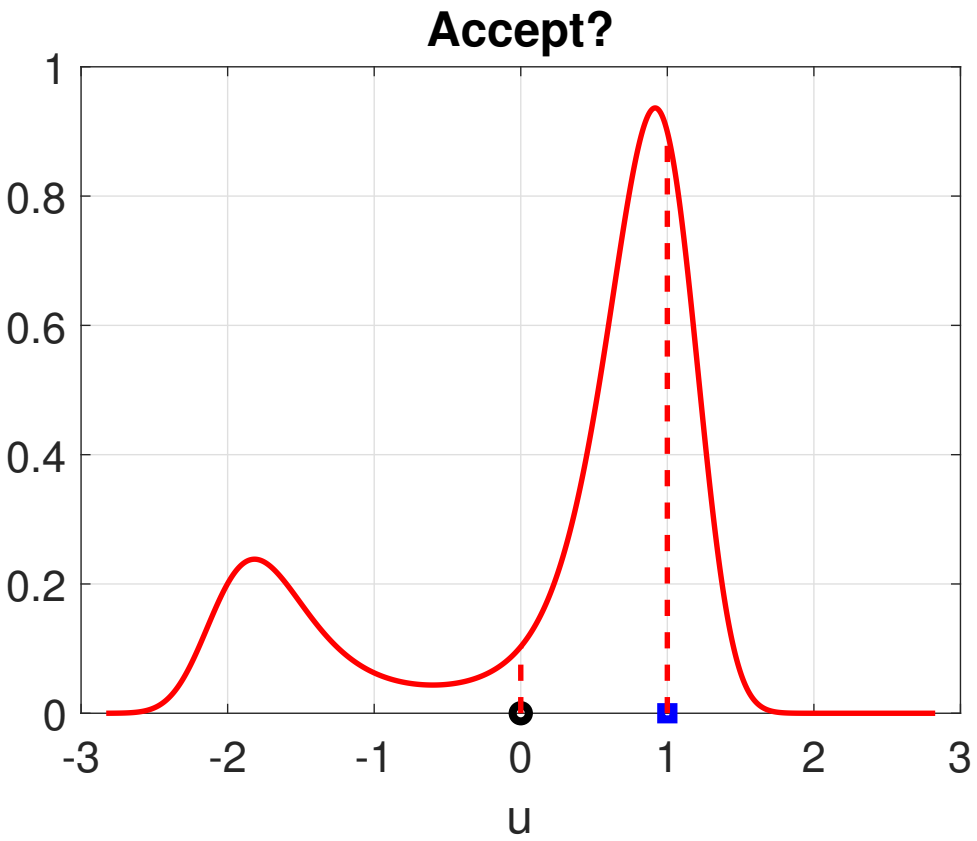
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



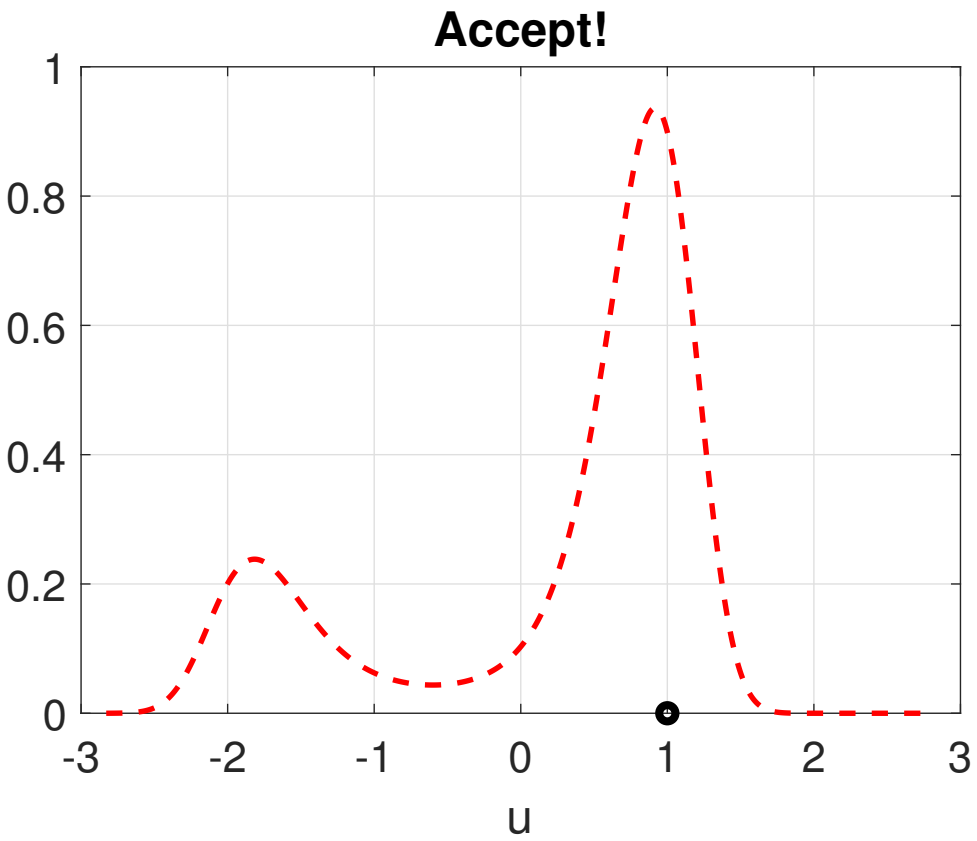
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



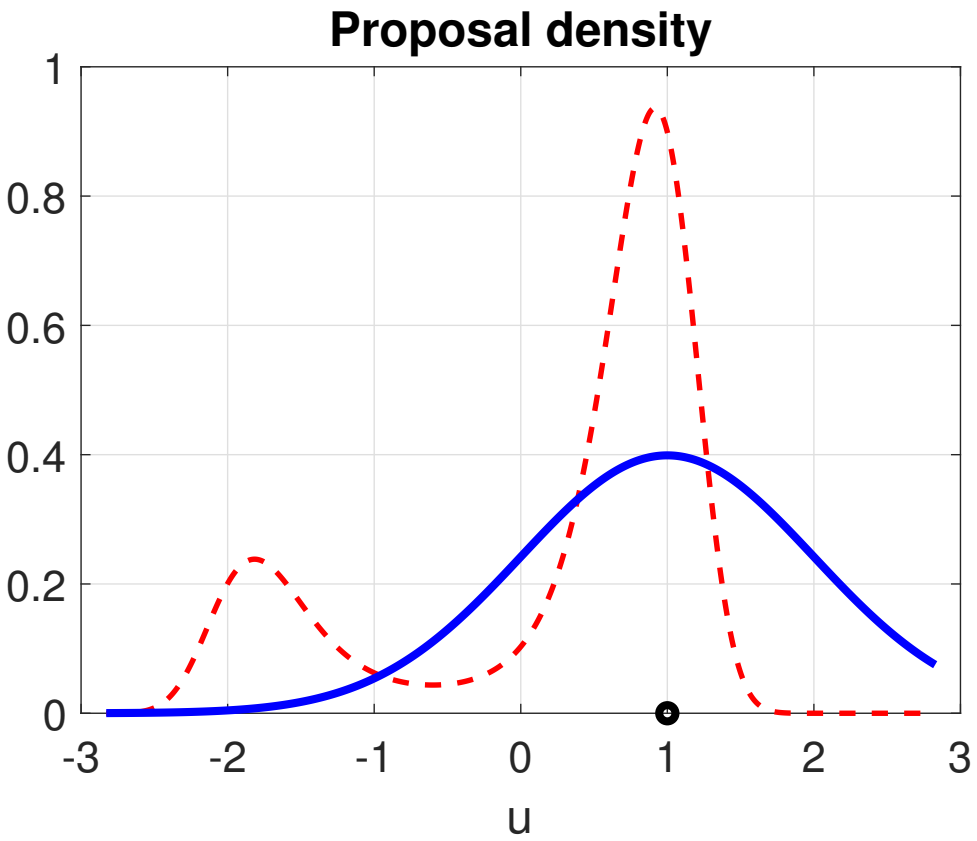
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



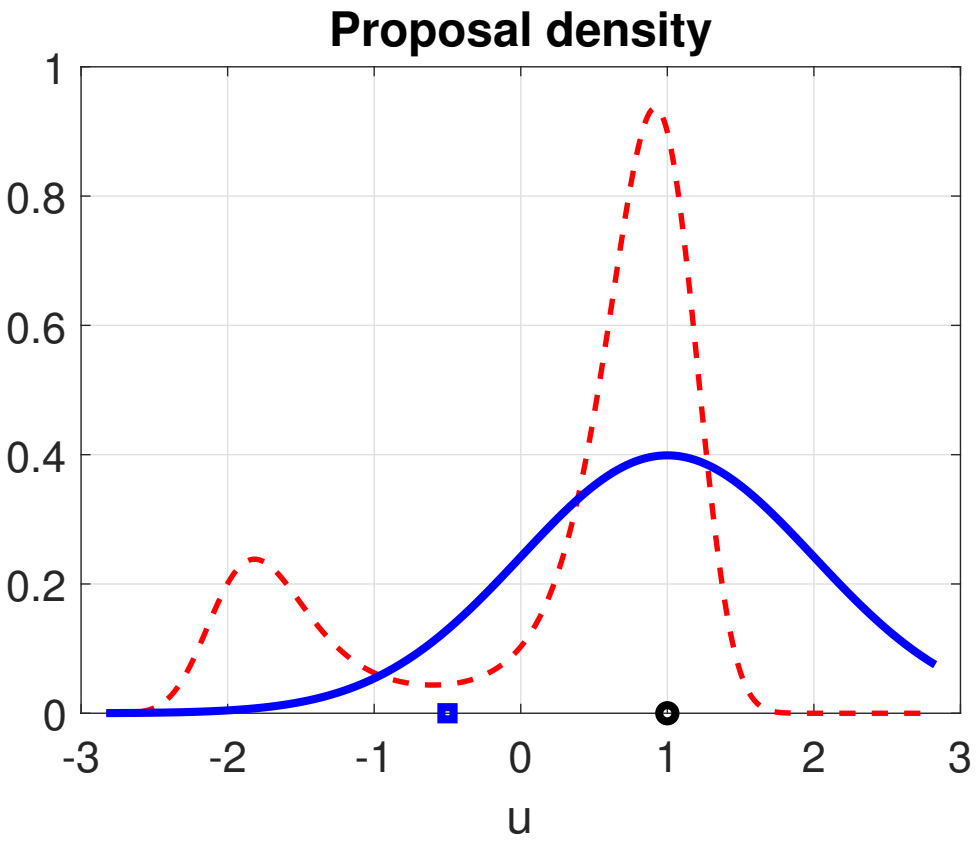
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



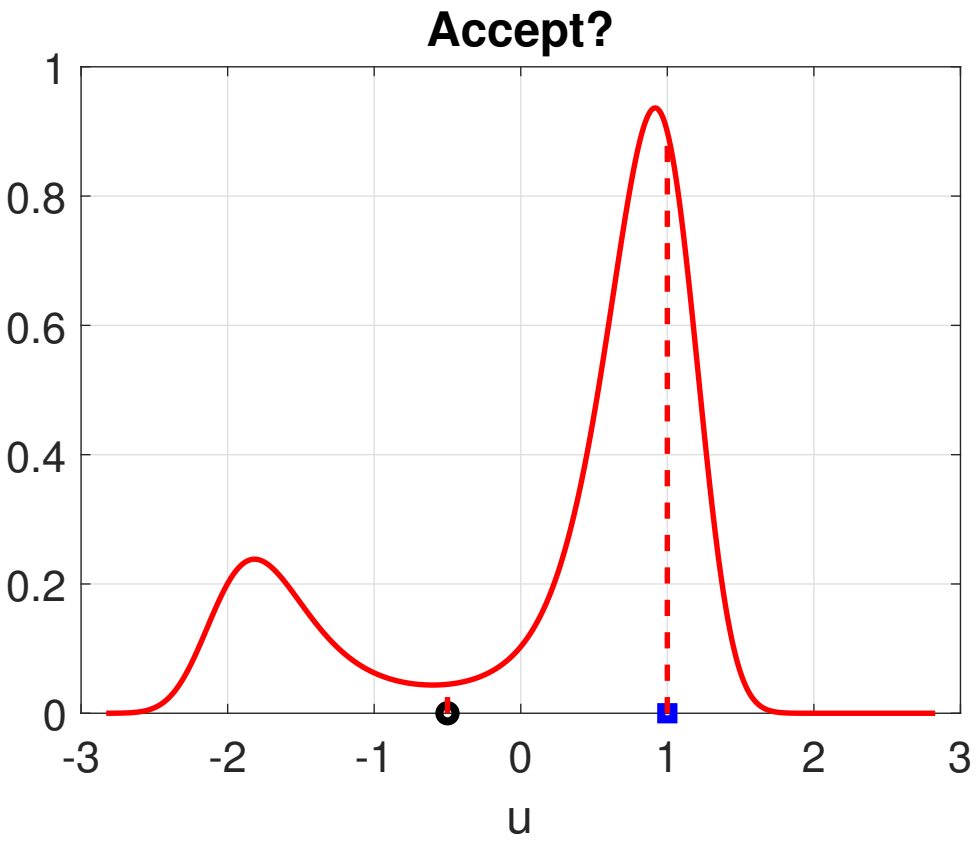
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



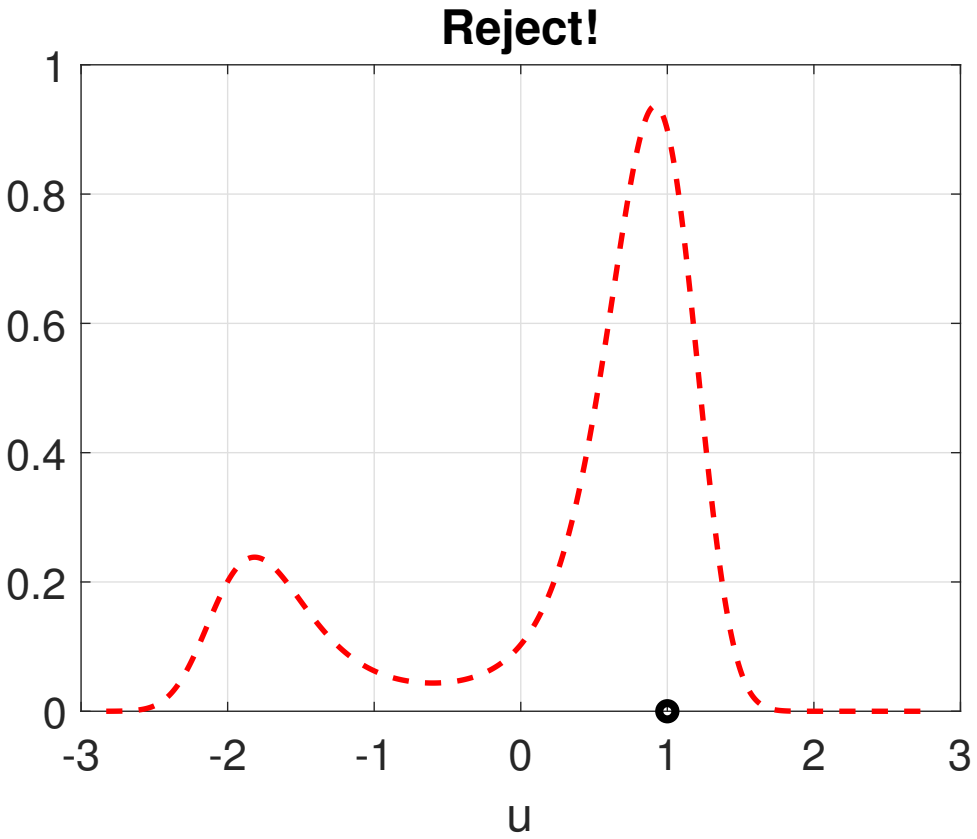
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



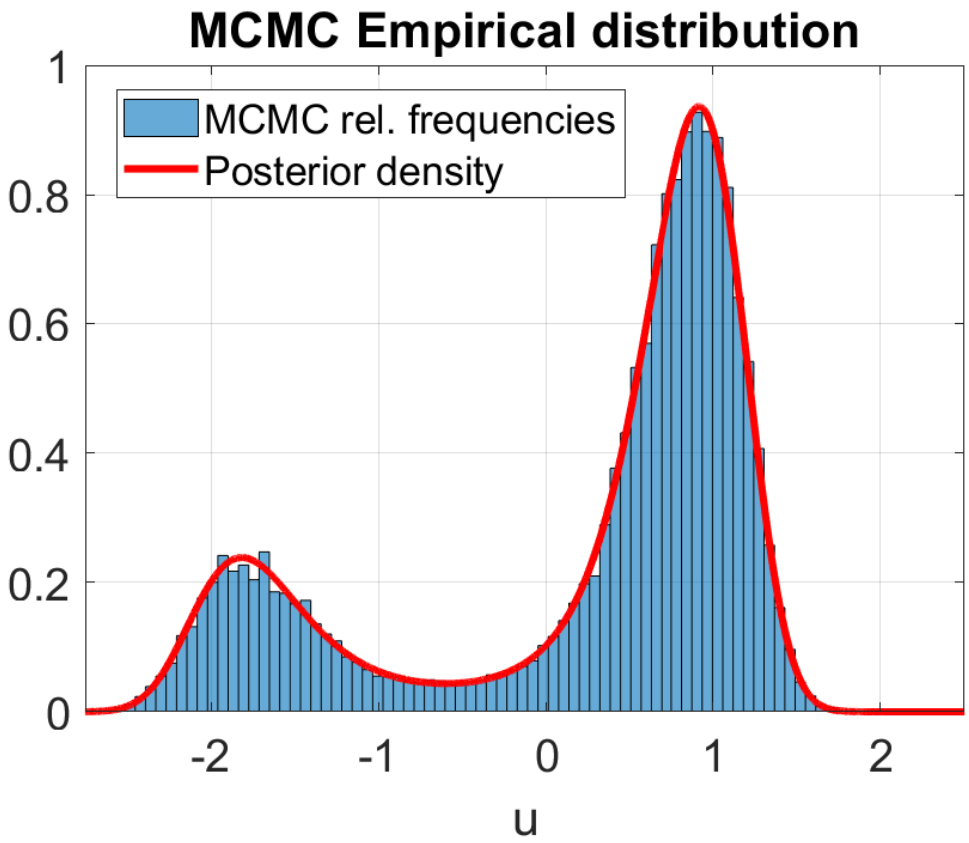
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



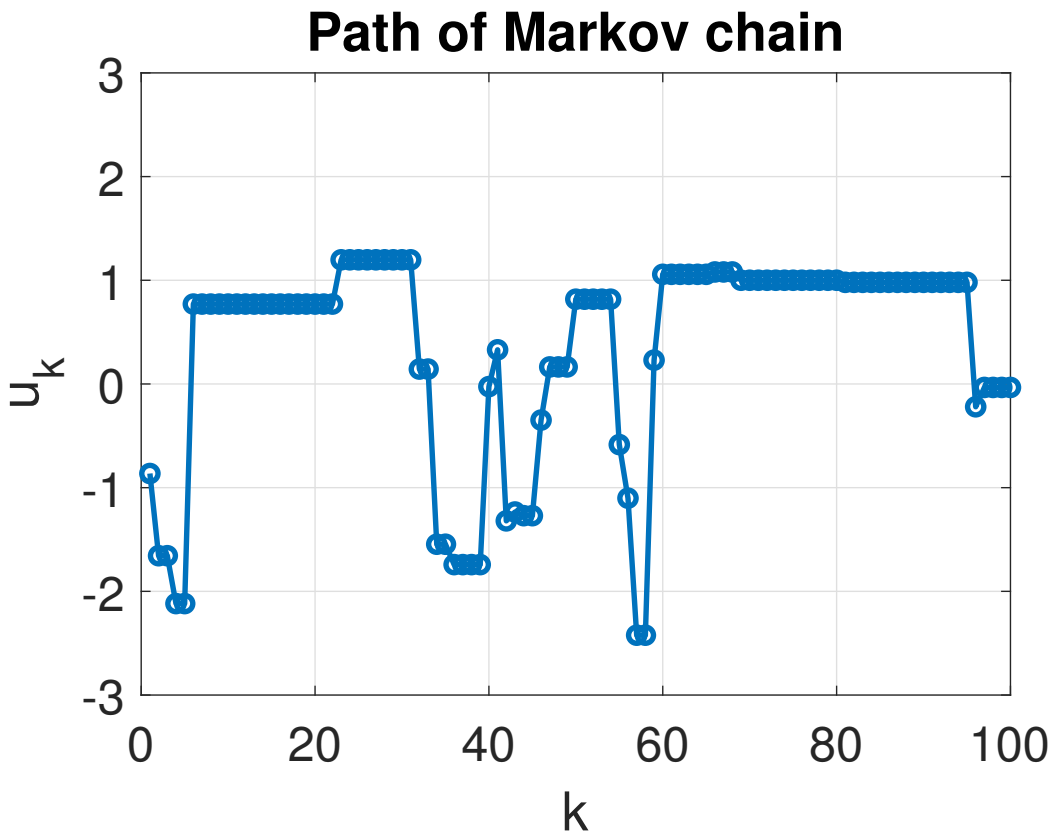
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u, 1)$ and $\alpha(u, v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



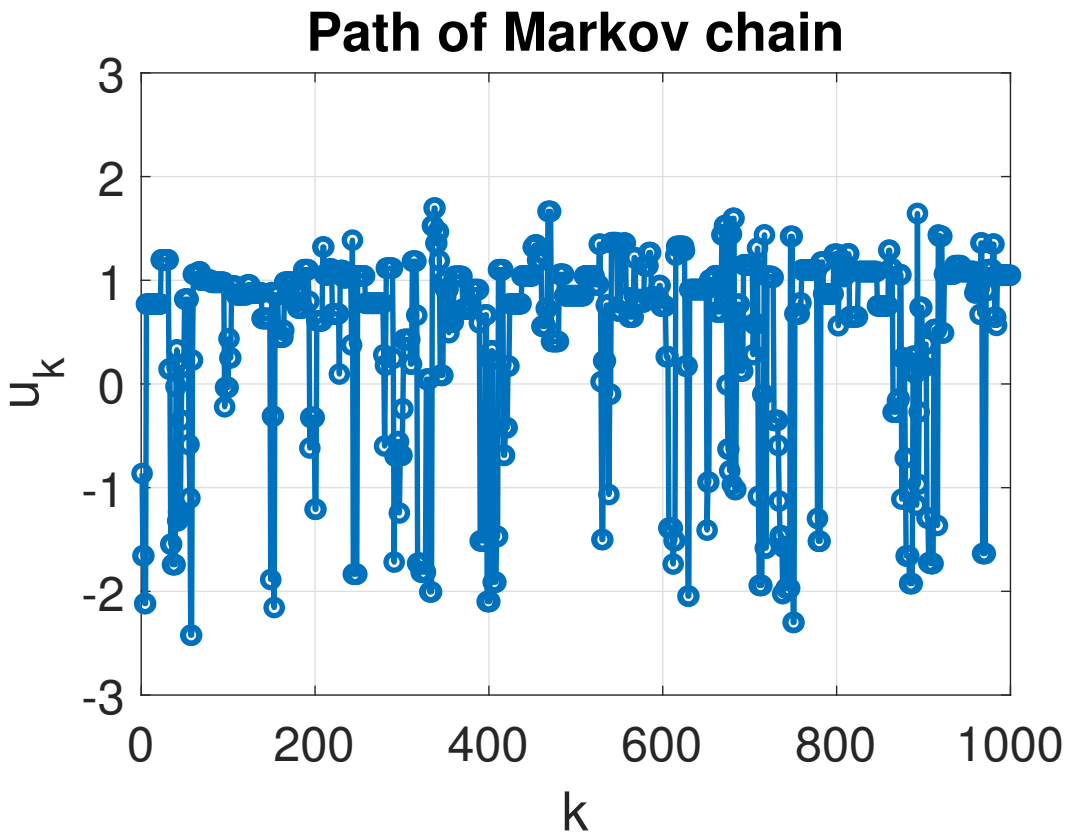
Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u,1)$ and $\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



Example

- Recall: Condition $u \sim \pi_0 = N(0,1)$ on $y = 2$, $\mathcal{G}(u) = u^2 + u$, $\eta \sim N(0,1)$
- **MH algorithm:** Proposal $P(u) = \mathrm{N}(u, 1)$ and $\alpha(u, v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$



Optimal scaling of the stepsize

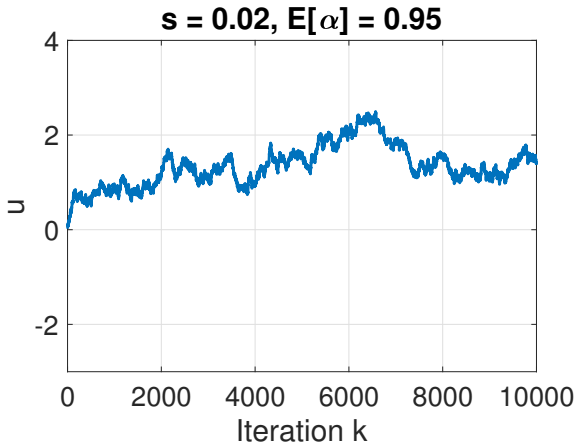
$$P(u) = N(u, s^2 C), \qquad s > 0$$

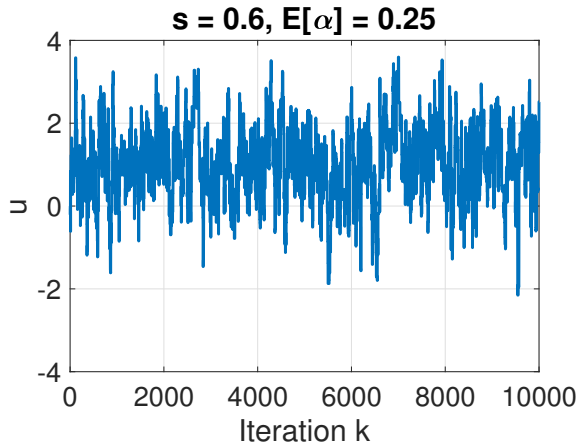
- How to choose the stepsize s? Is it better to propose larger moves or samller and obtain high acceptance probability?

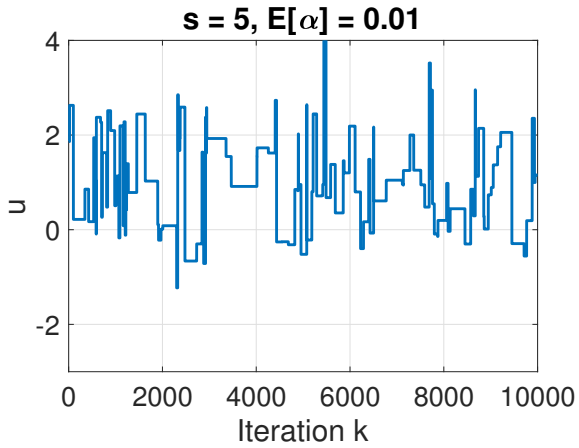
Optimal scaling of the stepsize

$$P(u) = N(u, s^2 C), \qquad s > 0$$

- How to choose the stepsize s? Is it better to propose larger moves or samller and obtain high acceptance probability?



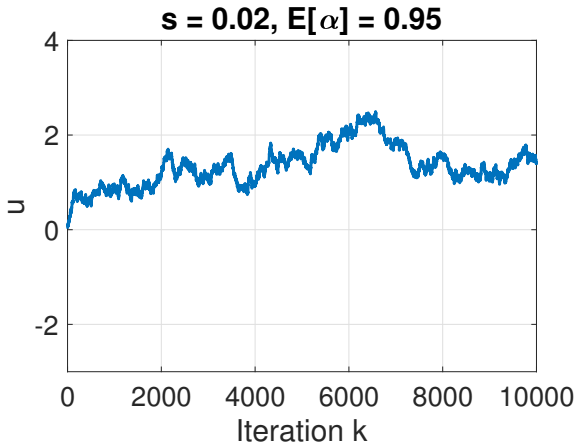


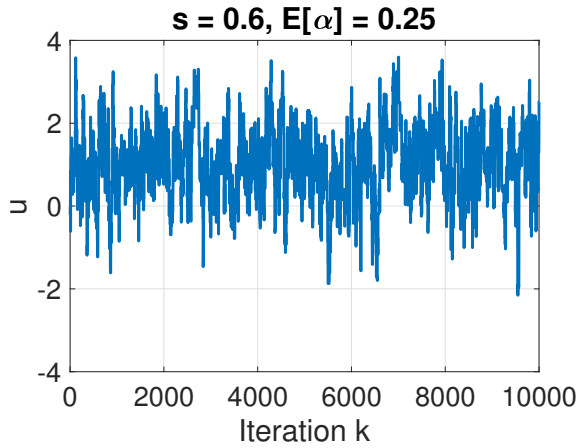


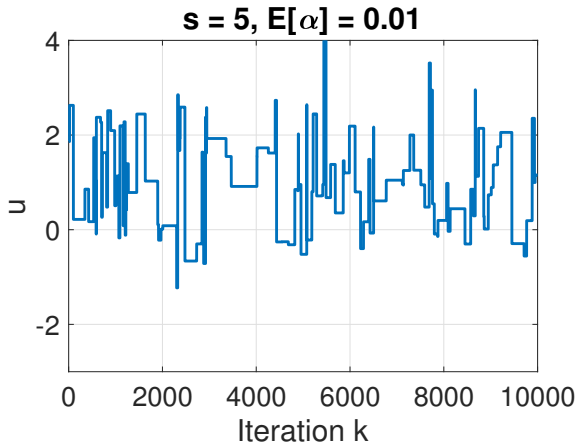
Optimal scaling of the stepsize

$$P(u) = N(u, s^2 C), \qquad s > 0$$

- How to choose the stepsize s? Is it better to propose larger moves or samller and obtain high acceptance probability?







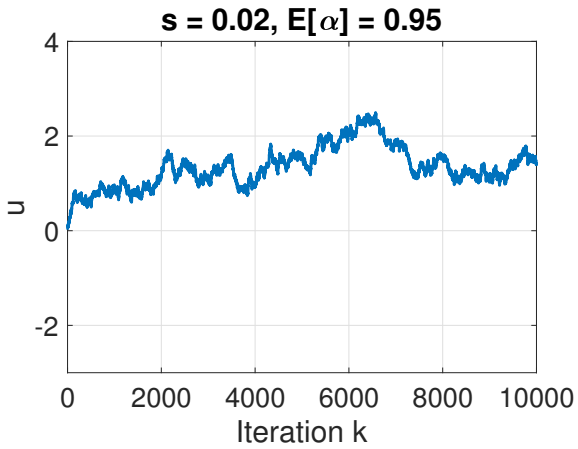
- Apparently, there is a sweet spot: "rule of thumb" [Roberts, Rosenthal, 2001]

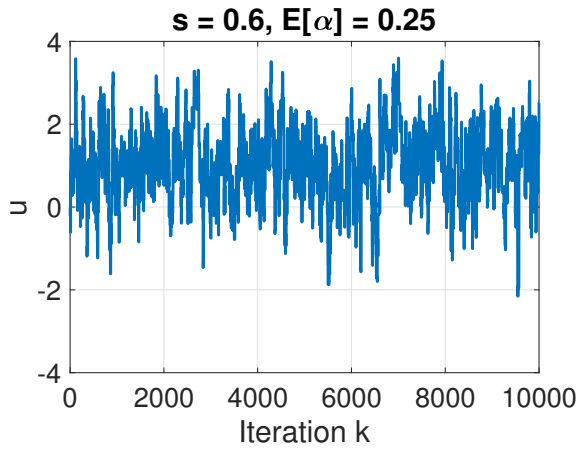
Choose s such that $\mathbb{E}\left[\alpha(U, V)\right] = 0.234$

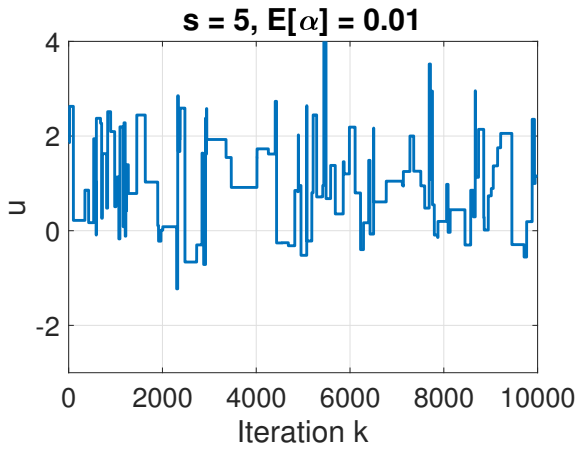
Optimal scaling of the stepsize

$$P(u) = N(u, s^2 C), \qquad s > 0$$

- How to choose the stepsize s? Is it better to propose larger moves or samller and obtain high acceptance probability?







- Apparently, there is a sweet spot: "rule of thumb" [Roberts, Rosenthal, 2001]

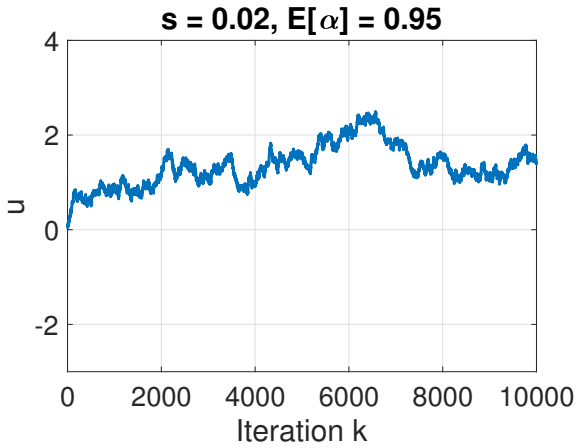
$$\text{Choose } s \text{ such that} \qquad \mathbb{E}\left[\alpha(U, V)\right] = 0.234$$

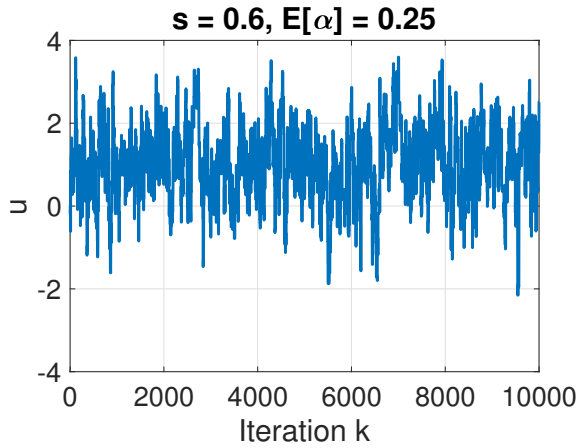
- **Problem:** dimension-dependence $s \sim \frac{1}{\dim(\mathcal{X})}$

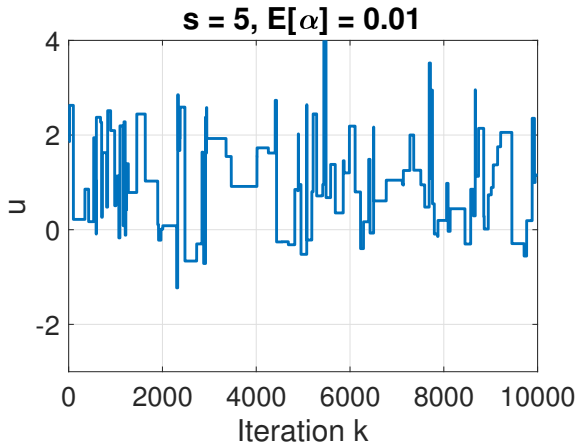
Optimal scaling of the stepsize

$$P(u) = N(u, s^2 C), \qquad s > 0$$

- How to choose the stepsize s? Is it better to propose larger moves or samller and obtain high acceptance probability?







- Apparently, there is a sweet spot: "rule of thumb" [Roberts, Rosenthal, 2001]

$$\text{Choose } s \text{ such that} \qquad \mathbb{E}\left[\alpha(U, V)\right] = 0.234$$

- **Problem:** dimension-dependence $s \sim \frac{1}{\dim(\mathcal{X})}$ as well as $\mathrm{IACT}_f \sim \dim(\mathcal{X})$

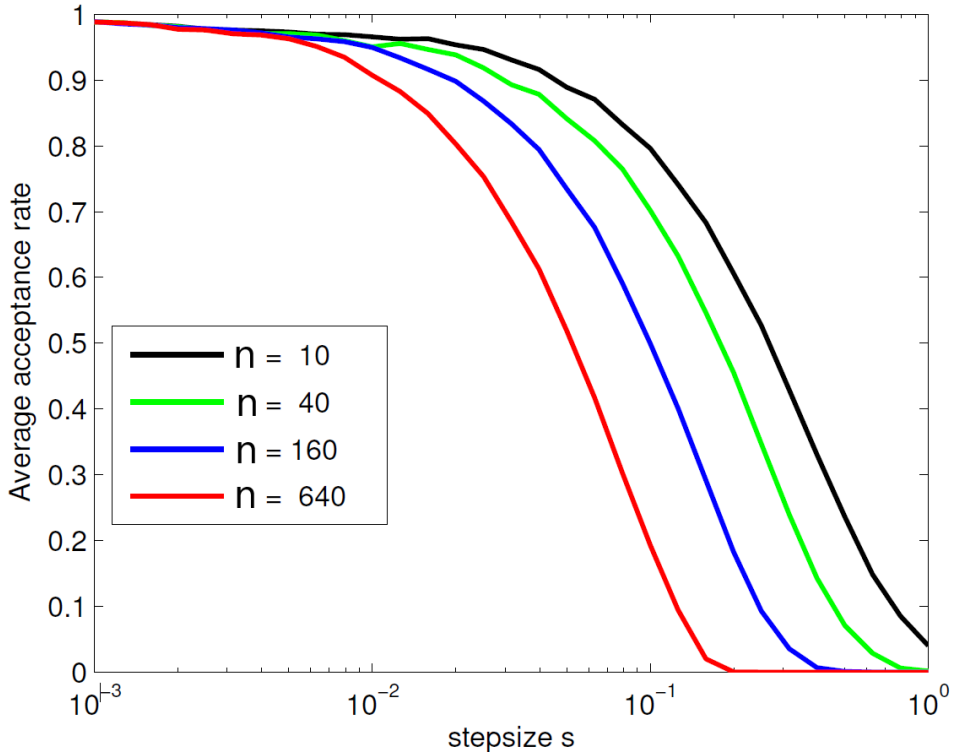
Numerical experiment

Problem: Bayesian inference in 2D groundwater flow model (elliptic diffusion problem)

Acceptance rate vs. stepsize s for different dimensions n of $u \in \mathbb{R}^n$

(n is the truncation length of the KL expansion of the random field)

Random walk-proposal $P(u) = N(u, s^2 C)$



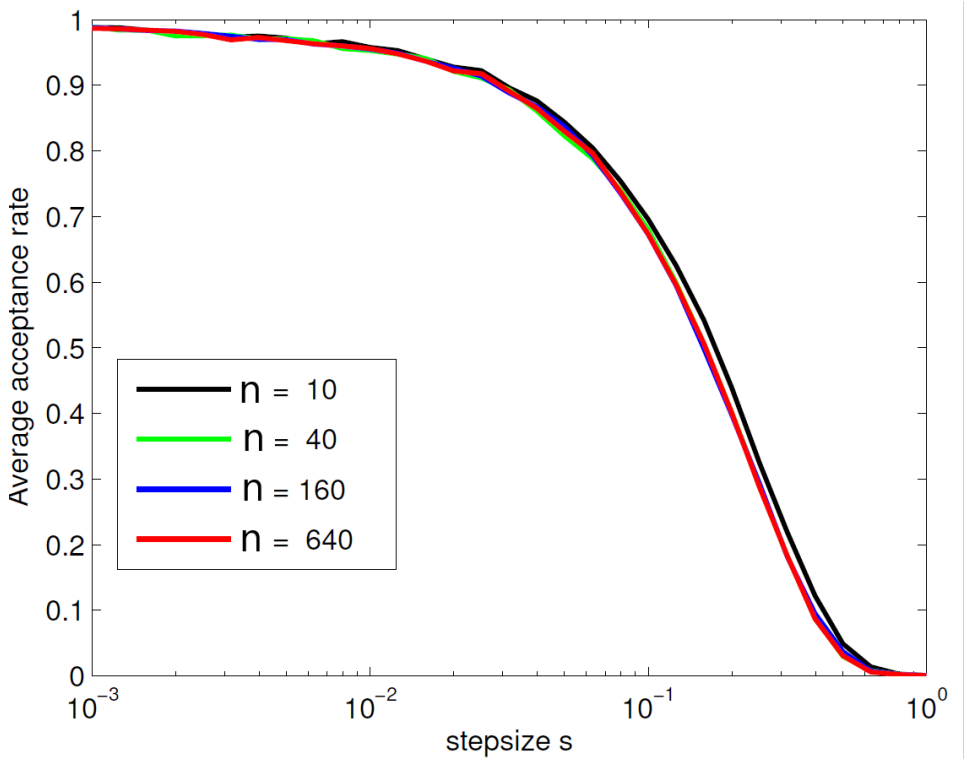
Numerical experiment

Problem: Bayesian inference in 2D groundwater flow model (elliptic diffusion problem)

Acceptance rate vs. stepsize s for different dimensions n of $u \in \mathbb{R}^n$

(n is the truncation length of the KL expansion of the random field)

pCN-proposal $P(u) = N(\sqrt{1-s^2}u, s^2 C)$ (details below)



Theoretical Analysis of MH algorithm

Proposition

The MH algorithm generates a π-reversible Markov chain.

Theoretical Analysis of MH algorithm

Proposition

The MH algorithm generates a π-reversible Markov chain.

Proof. For ease we consider disjoint $A, B \subseteq \mathcal{X}$. Then, we require that

$$\int_A K(u, B)\, \pi(\mathrm{d}u) = \int_B K(u, A)\, \pi(\mathrm{d}u)$$

Theoretical Analysis of MH algorithm

Proposition

The MH algorithm generates a π-reversible Markov chain.

Proof. For ease we consider disjoint $A, B \subseteq \mathcal{X}$. Then, we require that

$$\int_A \int_B \alpha(u,v)\, p(u,v)\, \pi(u)\, \mathrm{d}u\mathrm{d}v = \int_B \int_A \alpha(u,v)\, p(u,v)\, \pi(u)\, \mathrm{d}u\mathrm{d}v$$

Theoretical Analysis of MH algorithm

Proposition

The MH algorithm generates a π-reversible Markov chain.

Proof. For ease we consider disjoint $A, B \subseteq \mathcal{X}$. Then, we require that

$$\int_A \int_B \alpha(u,v)\, p(u,v)\, \pi(u)\, \mathrm{d}u\mathrm{d}v = \int_B \int_A \alpha(u,v)\, p(u,v)\, \pi(u)\, \mathrm{d}u\mathrm{d}v$$

This follows by Fubini and

$$\begin{aligned}
\alpha(u,v)\ p(u,v)\ \pi(u) &= \min\left\{1, \frac{p(v,u)\ \pi(v)}{p(u,v)\ \pi(u)}\right\} p(u,v)\ \pi(u) \\
&= \min\left\{p(u,v)\ \pi(x), p(v,u)\ \pi(v)\right\} \\
&= \min\left\{\frac{p(u,v)\ \pi(x)}{p(v,u)\ \pi(v)}, 1\right\} p(v,u)\ \pi(v) = \alpha(v,u)\ p(v,u)\ \pi(v).
\end{aligned}$$

Theoretical Analysis of MH algorithm

Theorem

Let K be a π-reversible MH transition kernel with proposal kernel P.

1. If K is irreducible, then it is also Harris recurrent.

Theoretical Analysis of MH algorithm

Theorem

Let K be a π-reversible MH transition kernel with proposal kernel P.

1. If K is irreducible, then it is also Harris recurrent.
2. If K is irreducible and there exists $u \in A$ with $\mu(A) > 0$ such that $K(u, \{u\}) > 0$, then K is also aperiodic.

Theoretical Analysis of MH algorithm

Theorem

Let K be a π-reversible MH transition kernel with proposal kernel P.

1. If K is irreducible, then it is also Harris recurrent.
2. If K is irreducible and there exists $u \in A$ with $\mu(A) > 0$ such that $K(u, \{u\}) > 0$, then K is also aperiodic.
3. If for each $u \in \mathcal{X}$ the proposal distribution $P(u, \cdot)$ has a continuous density $p(u, \cdot) \colon \mathcal{X} \to \mathbb{R}$ and $p(u, v) > 0$ for all $u, v \in \mathcal{X} \subseteq \mathbb{R}^n$, then K is irreducible.

Theoretical Analysis of MH algorithm

Theorem

Let K be a π-reversible MH transition kernel with proposal kernel P.

1. If K is irreducible, then it is also Harris recurrent.
2. If K is irreducible and there exists $u \in A$ with $\mu(A) > 0$ such that $K(u, \{u\}) > 0$, then K is also aperiodic.
3. If for each $u \in \mathcal{X}$ the proposal distribution $P(u, \cdot)$ has a continuous density $p(u, \cdot) \colon \mathcal{X} \to \mathbb{R}$ and $p(u, v) > 0$ for all $u, v \in \mathcal{X} \subseteq \mathbb{R}^n$, then K is irreducible.

Corollary for Random Walk in $\mathcal{X} = \mathbb{R}^n$

Given target measure π with Lebesgue density $\pi(u) \propto \exp(-\Phi(u))\, \pi_0(u)$ the MH algorithm with proposal kernel $P(u) = \mathrm{N}(u, s^2 C)$ generates

1. an irreducible, aperiodic, Harris recurrent, and, thus, ergodic, Markov chain
2. which is, moreover, geometrically ergodic (under suitable conditions on π).

Metropolis–Hastings in Hilbert spaces

- Now $\pi \in \mathcal{P}(\mathcal{X})$ and $\mathcal{X}$ is a separable Hilbert space.
- Then, the MH algorithm targeting π and using "any" proposal kernel $P\colon \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ is **not always well-defined**.

Metropolis–Hastings in Hilbert spaces

- Now $\pi \in \mathcal{P}(\mathcal{X})$ and $\mathcal{X}$ is a separable Hilbert space.
- Then, the MH algorithm targeting π and using "any" proposal kernel $P\colon \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ is **not always well-defined**.
- In particular, for Gaussian priors $\pi_0 = \mathrm{N}(0, C)$

Metropolis–Hastings in Hilbert spaces

- Now $\pi \in \mathcal{P}(\mathcal{X})$ and $\mathcal{X}$ is a separable Hilbert space.
- Then, the MH algorithm targeting π and using "any" proposal kernel $P\colon \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ is **not always well-defined**.
- In particular, for Gaussian priors $\pi_0 = \mathrm{N}(0, C)$ and random walk proposal $P = \mathrm{N}(u, s^2 C)$, the MH algorithm is not well-defined in infinite dimensions.
  (which also explains the scaling $s \sim \frac{1}{\dim(\mathcal{X})}$ to achieve a certain acceptance rate)

Metropolis–Hastings in Hilbert spaces

- Now $\pi \in \mathcal{P}(\mathcal{X})$ and $\mathcal{X}$ is a separable Hilbert space.
- Then, the MH algorithm targeting π and using "any" proposal kernel $P \colon \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ is **not always well-defined**.
- In particular, for Gaussian priors $\pi_0 = \mathrm{N}(0, C)$ and random walk proposal $P = \mathrm{N}(u, s^2 C)$, the MH algorithm is not well-defined in infinite dimensions.
  (which also explains the scaling $s \sim \frac{1}{\dim(\mathcal{X})}$ to achieve a certain acceptance rate)
- However, if $\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\, \pi_0(\mathrm{d}u)$ then it suffices to construct a prior π_0-reversible proposal kernel P. In that case:

$$\alpha(u, v) := \min \left\{ 1, \frac{\exp(-\Phi(v))}{\exp(-\Phi(u))} \right\}$$

Metropolis–Hastings in Hilbert spaces

- Now $\pi \in \mathcal{P}(\mathcal{X})$ and $\mathcal{X}$ is a separable Hilbert space.
- Then, the MH algorithm targeting π and using "any" proposal kernel $P\colon \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ is **not always well-defined**.
- In particular, for Gaussian priors $\pi_0 = \mathrm{N}(0, C)$ and random walk proposal $P = \mathrm{N}(u, s^2 C)$, the MH algorithm is not well-defined in infinite dimensions.
  (which also explains the scaling $s \sim \frac{1}{\dim(\mathcal{X})}$ to achieve a certain acceptance rate)
- However, if $\pi(\mathrm{d}u) \propto \exp(-\Phi(u))\, \pi_0(\mathrm{d}u)$ then it suffices to construct a prior π_0-reversible proposal kernel P. In that case:

$$\alpha(u,v) := \min\left\{1, \frac{\exp(-\Phi(v))}{\exp(-\Phi(u))}\right\}$$

- The slightly modified pCN-proposal kernel $P = \mathrm{N}(\sqrt{1-s^2}u, s^2 C)$ on the other hand is π_0-reversible and yields a **well-defined MH algorithm on infinite-dimensional Hilbert spaces**.

Remarks

- pCN stands for "preconditioned Crank-Nicolson"

[4] A. Beskos, G. Roberts, A. Stuart, J. Voss. MCMC Methods for diffusion bridges. *Stoch. Dynam.* 8, 2008.

[5] R. M. Neal. Regression and Classification using Gaussian Process Priors. In: *Bayesian Statistics* 6, 1999.

[6] M. Hairer, A. Stuart, S. Vollmer. Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* 24(6):2455-2490, 2014

Remarks

- pCN stands for "preconditioned Crank-Nicolson"
- It was derived by a corresponding numerical discretiziation of a Langevin SDE with π_0 as invariant measure[4]

[4] A. Beskos, G. Roberts, A. Stuart, J. Voss. MCMC Methods for diffusion bridges. *Stoch. Dynam.* 8, 2008.

[5] R. M. Neal. Regression and Classification using Gaussian Process Priors. In: *Bayesian Statistics* 6, 1999.

[6] M. Hairer, A. Stuart, S. Vollmer. Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* 24(6):2455-2490, 2014

Remarks

- pCN stands for "preconditioned Crank-Nicolson"
- It was derived by a corresponding numerical discretiziation of a Langevin SDE with π_0 as invariant measure[4]
- Although, it was proposed earlier[5] but without derivation or analysis

[4] A. Beskos, G. Roberts, A. Stuart, J. Voss. MCMC Methods for diffusion bridges. *Stoch. Dynam.* 8, 2008.

[5] R. M. Neal. Regression and Classification using Gaussian Process Priors. In: *Bayesian Statistics* 6, 1999.

[6] M. Hairer, A. Stuart, S. Vollmer. Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* 24(6):2455-2490, 2014

Remarks

- pCN stands for "preconditioned Crank-Nicolson"
- It was derived by a corresponding numerical discretiziation of a Langevin SDE with π_0 as invariant measure[4]
- Although, it was proposed earlier[5] but without derivation or analysis
- It requires Gaussian priors $\pi_0 = \mathrm{N}(m, C)$ but can be generalized to use other covariances than C, e.g., the covariance of the Laplace approximation

[4] A. Beskos, G. Roberts, A. Stuart, J. Voss. MCMC Methods for diffusion bridges. *Stoch. Dynam.* 8, 2008.

[5] R. M. Neal. Regression and Classification using Gaussian Process Priors. In: *Bayesian Statistics* 6, 1999.

[6] M. Hairer, A. Stuart, S. Vollmer. Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* 24(6):2455-2490, 2014

Remarks

- pCN stands for "preconditioned Crank-Nicolson"
- It was derived by a corresponding numerical discretiziation of a Langevin SDE with π_0 as invariant measure[4]
- Although, it was proposed earlier[5] but without derivation or analysis
- It requires Gaussian priors $\pi_0 = \mathrm{N}(m, C)$ but can be generalized to use other covariances than C, e.g., the covariance of the Laplace approximation
- It was shown under suitable assumptions on the potential Φ[6] that the corresponding transition kernel has a dimension-independent spectral gap

[4] A. Beskos, G. Roberts, A. Stuart, J. Voss. MCMC Methods for diffusion bridges. *Stoch. Dynam.* 8, 2008.

[5] R. M. Neal. Regression and Classification using Gaussian Process Priors. In: *Bayesian Statistics* 6, 1999.

[6] M. Hairer, A. Stuart, S. Vollmer. Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* 24(6):2455-2490, 2014

Remarks

- pCN stands for "preconditioned Crank-Nicolson"
- It was derived by a corresponding numerical discretiziation of a Langevin SDE with π_0 as invariant measure[4]
- Although, it was proposed earlier[5] but without derivation or analysis
- It requires Gaussian priors $\pi_0 = \mathrm{N}(m, C)$ but can be generalized to use other covariances than C, e.g., the covariance of the Laplace approximation
- It was shown under suitable assumptions on the potential Φ[6] that the corresponding transition kernel has a dimension-independent spectral gap
- In practice, also require 'burn-in' to overcome pre-asymptotic phase for $U_0 \not\sim \pi$

[4] A. Beskos, G. Roberts, A. Stuart, J. Voss. MCMC Methods for diffusion bridges. *Stoch. Dynam.* 8, 2008.

[5] R. M. Neal. Regression and Classification using Gaussian Process Priors. In: *Bayesian Statistics* 6, 1999.

[6] M. Hairer, A. Stuart, S. Vollmer. Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* 24(6):2455-2490, 2014

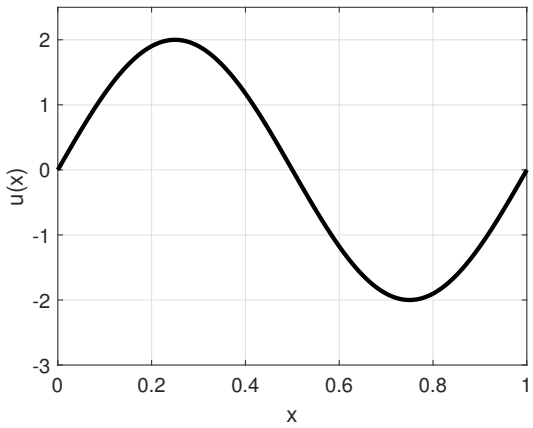
Numerical Experiment

Task: Infer unknown $u\colon [0,1] \to \mathbb{R}$ based on 4 noisy observations of p

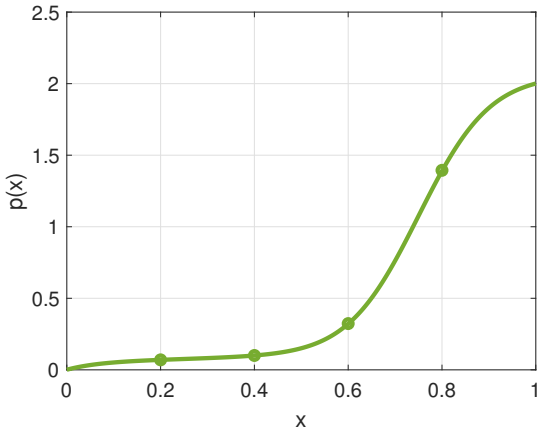
$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{e}^{u(x)}\frac{\mathrm{d}p}{\mathrm{d}x}(x)\right) = 0, \qquad p(0) = 0,\ p(1) = 2.$$

Prior: $u \sim \pi_0 = \mathrm{N}(0, (-\Delta)^{-1})$ **Noise:** $\eta \sim \mathrm{N}(0, \sigma^2 I_4)$

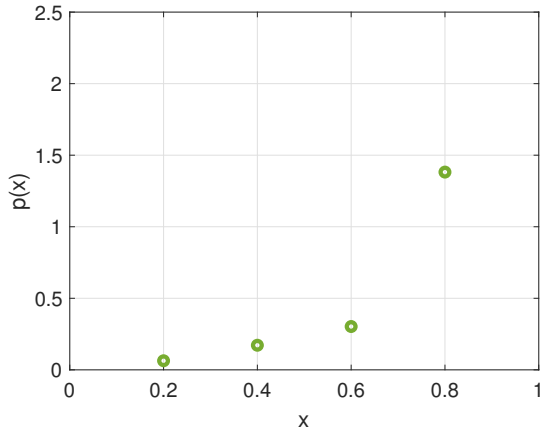
True u



True p



Noisy data $\boldsymbol{\delta}$



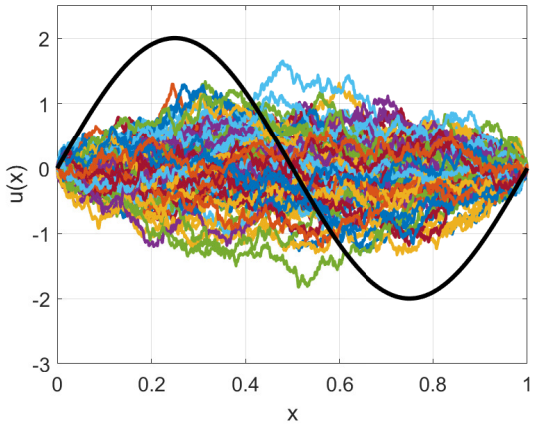
Numerical Experiment

Task: Infer unknown $u\colon [0,1] \to \mathbb{R}$ based on 4 noisy observations of p

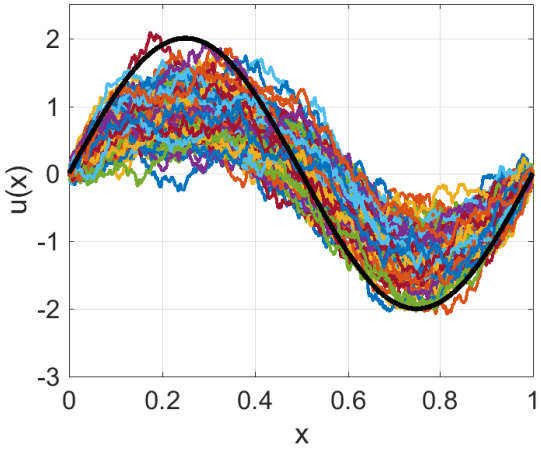
$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{e}^{u(x)}\frac{\mathrm{d}p}{\mathrm{d}x}(x)\right) = 0, \qquad p(0) = 0,\ p(1) = 2.$$

Prior: $u \sim \pi_0 = \mathrm{N}(0, (-\Delta)^{-1})$ **Noise:** $\eta \sim \mathrm{N}(0, \sigma^2 I_4)$

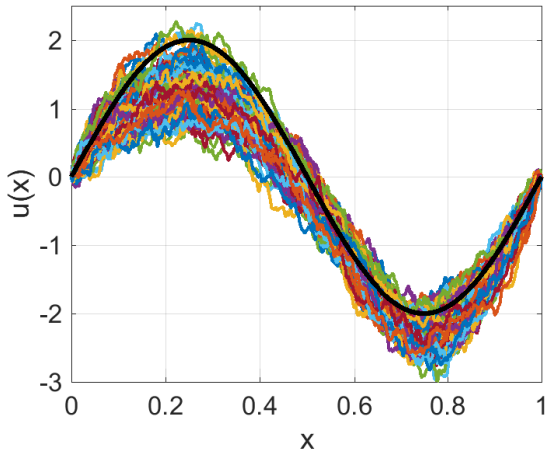
Prior π_0



π^y for $\sigma^2 = 0.01$



π^y for $\sigma^2 = 0.01^2$



Numerical Experiment

Task: Infer unknown $u\colon [0,1] \to \mathbb{R}$ based on 4 noisy observations of p

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{e}^{u(x)}\frac{\mathrm{d}p}{\mathrm{d}x}(x)\right) = 0, \qquad p(0) = 0,\ p(1) = 2.$$

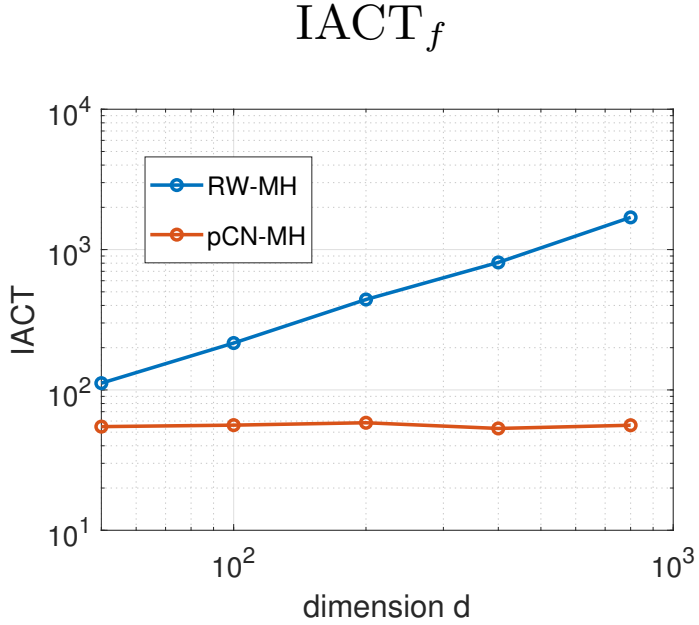
Prior: $u \sim \pi_0 = \mathrm{N}(0, (-\Delta)^{-1})$ **Noise:** $\eta \sim \mathrm{N}(0, \sigma^2 I_4)$

Performance:

RW $P(u) = \mathrm{N}(u, s^2(-\Delta)^{-1})$

pCN $P(u) = \mathrm{N}(\sqrt{1-s^2}u, s^2(-\Delta)^{-1})$

for increasing dimension, i.e., truncation length in KL expansion for u.



$\Longrightarrow$ **Dimension-independent efficiency of pCN-MH algorithm for sampling posterior in Bayesian inverse problem**

Summary – Part I

- Several Monte Carlo methods available to sample or integrate (approximately) w.r.t. posterior measure $\pi_{u|y}$.
- Markov chain Monte Carlo yields correlated samples U_i that are asymptotically distributed according to $\pi_{u|y}$ as $i \to \infty$
- Nonetheless preferable due to mild dimension-(in)dependence and usually robust performance
- Broad class of Metropolis–Hastings algorithms easy to implement
- pCN-Metropolis algorithm well-defined in Hilbert spaces for Gaussian priors/reference measures with dimension-independent spectral gap

Summary – Part I

- Several Monte Carlo methods available to sample or integrate (approximately) w.r.t. posterior measure $\pi_{u|y}$.
- Markov chain Monte Carlo yields correlated samples U_i that are asymptotically distributed according to $\pi_{u|y}$ as $i \to \infty$
- Nonetheless preferable due to mild dimension-(in)dependence and usually robust performance
- Broad class of Metropolis–Hastings algorithms easy to implement
- pCN-Metropolis algorithm well-defined in Hilbert spaces for Gaussian priors/reference measures with dimension-independent spectral gap

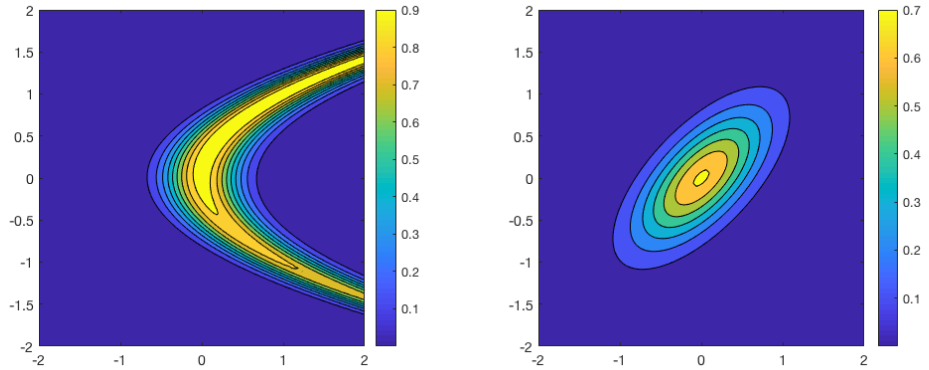
However, for concentrating posteriors in high dimensions the spectral gap of the pCN-Metropolis algorithm is very close to 1 and it **converges very slowly!**

(like Jacobi method for linear systems)

Better Proposals

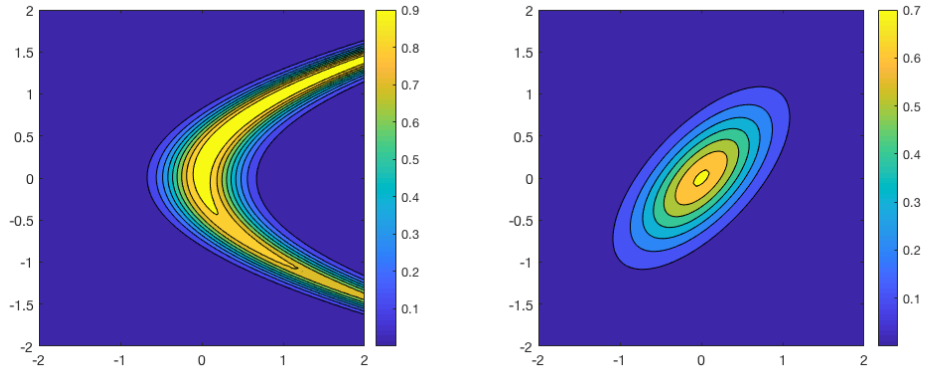
How to incorporate gradient information?

- The proposals we have seen so far are agnostic about which parts of state space are more probable.
- Ideally we would like proposals that take this into account ($\Rightarrow$ make it more probable to move to areas where π is large).



How to incorporate gradient information?

- The proposals we have seen so far are agnostic about which parts of state space are more probable.
- Ideally we would like proposals that take this into account ($\Rightarrow$ make it more probable to move to areas where π is large).



- Connecting to optimisation, a possible way to do this is to use gradient information and propose the next move in the following way

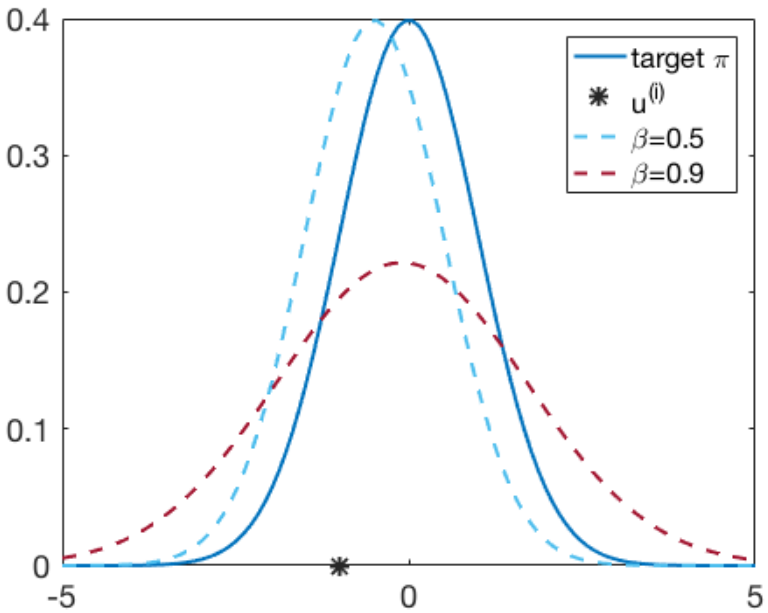
$$u' = u^{(i)} + \beta \nabla \pi(u^{(i)})$$

1. This is a deterministic move (we are losing randomness, and the ability to explore the state space, as we would converge to a local maximum).
2. How can we do this properly?

Metropolis adjusted Langevin algorithm (MALA)

- MALA [Pillai, Stuart, Thiery, 2012]: $P(u) = \mathrm{N}(u + \beta \nabla \log \pi(u), 2\beta \mathrm{I}),\ \beta > 0.$
- For optimal efficiency, step size β tuned s.t. average acceptance rate ≈ 0.574.

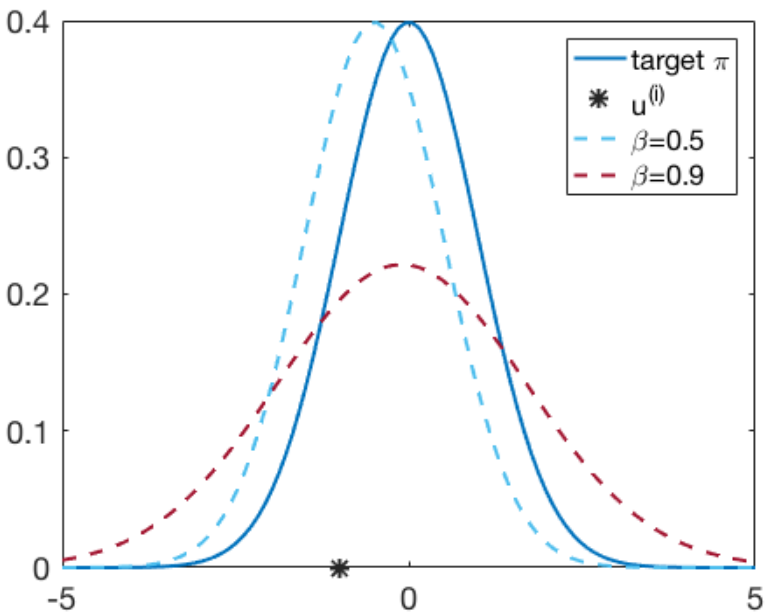
$$\Rightarrow \quad \beta \sim \dim(\mathcal{X})^{-1/3}$$



Metropolis adjusted Langevin algorithm (MALA)

- MALA [Pillai, Stuart, Thiery, 2012]: $P(u) = \mathrm{N}(u + \beta \nabla \log \pi(u), 2\beta \mathrm{I}),\ \beta > 0.$
- For optimal efficiency, step size β tuned s.t. average acceptance rate ≈ 0.574.

$$\Rightarrow \quad \beta \sim \dim(\mathcal{X})^{-1/3}$$



- Note that this is one Euler-Maruyama step applied to the Langevin SDE

$$dX = \nabla \log \pi(X) dt + \sqrt{2} dW$$

$$X_{n+1} = X_n + \beta \nabla \log \pi(X_n) + \sqrt{2\beta}\xi_n$$

with $\xi_n \sim \mathcal{N}(0, \mathrm{I})$.

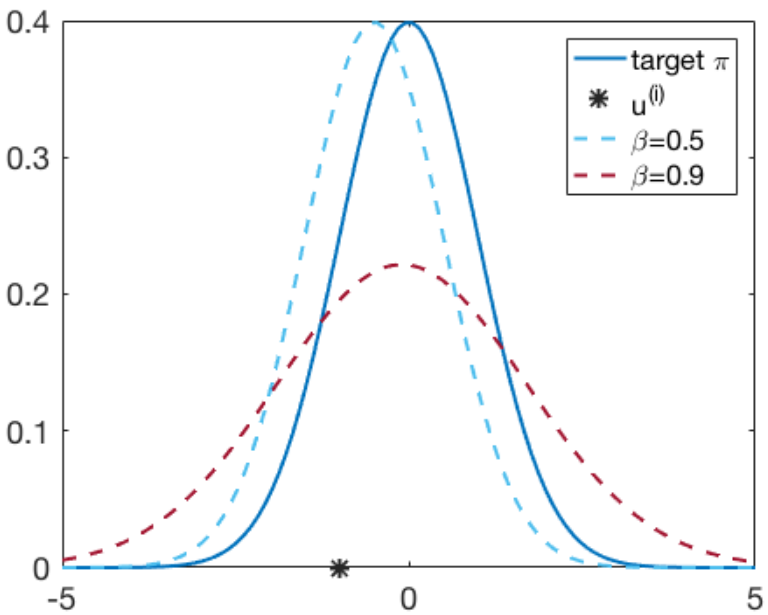
- The stationary distribution of the Langevin SDE is π.

(if $u \sim \pi$ and Euler-Maruyama exact $\Rightarrow v \sim \pi$)

Metropolis adjusted Langevin algorithm (MALA)

- MALA [Pillai, Stuart, Thiery, 2012]: $P(u) = \mathrm{N}(u + \beta \nabla \log \pi(u), 2\beta \mathrm{I}),\ \beta > 0.$
- For optimal efficiency, step size β tuned s.t. average acceptance rate ≈ 0.574.

$$\Rightarrow \quad \beta \sim \dim(\mathcal{X})^{-1/3}$$



- Note that this is one Euler-Maruyama step applied to the Langevin SDE

$$dX = \nabla \log \pi(X) dt + \sqrt{2} dW$$

$$X_{n+1} = X_n + \beta \nabla \log \pi(X_n) + \sqrt{2\beta}\xi_n$$

with $\xi_n \sim \mathcal{N}(0, \mathrm{I})$.

- The stationary distribution of the Langevin SDE is π.

(if $u \sim \pi$ and Euler-Maruyama exact $\Rightarrow v \sim \pi$)

- Can also include **Hessian** (2^{nd}-order) information. [Girolami, Calderhead, 2011], [Cui , Law, Marzouk, 2016], [Rudolf, Sprungk, 2018], ...

Surrogate transition method [Liu, 2001]

- This method uses a surrogate posterior $\pi^*(u)$ to pre-screen proposals.

Surrogate transition method

- At state u, sample a proposal v^* from proposal density $p^*(u, \cdot)$.
- Set $v = v^*$ with probability

$$\alpha_1(u, v^*) = \min\left(1, \frac{\pi^*(v^*)\, p^*(u, v^*)}{\pi^*(u)\, p^*(v^*, u)}\right),$$

otherwise $v = u$. Denote proposal density associated with this draw of v by $p(v, u)$.

Surrogate transition method [Liu, 2001]

- This method uses a surrogate posterior $\pi^*(u)$ to pre-screen proposals.

Surrogate transition method

- At state u, sample a proposal v^* from proposal density $p^*(u, \cdot)$.
- Set $v = v^*$ with probability

$$\alpha_1(u, v^*) = \min\left(1, \frac{\pi^*(v^*)\, p^*(u, v^*)}{\pi^*(u)\, p^*(v^*, u)}\right),$$

  otherwise $v = u$. Denote proposal density associated with this draw of v by $p(v, u)$.
- Accept v with probability

$$\alpha_2(u, v) = \min\left(1, \frac{\pi(v)\, p(u, v)}{\pi(u)\, p(v, u)}\right) = \min\left(1, \frac{\pi(v)\, \pi^*(u)}{\pi(u)\, \pi^*(v)}\right),$$

  i.e. $U^{(i+1)} = v$ with probability $\alpha_2(u, v)$; otherwise stay at $U^{(i+1)} = u$.

Surrogate transition method [Liu, 2001]

- This method uses a surrogate posterior $\pi^*(u)$ to pre-screen proposals.

Surrogate transition method

- At state u, sample a proposal v^* from proposal density $p^*(u, \cdot)$.
- Set $v = v^*$ with probability

$$\alpha_1(u, v^*) = \min\left(1, \frac{\pi^*(v^*)\, p^*(u, v^*)}{\pi^*(u)\, p^*(v^*, u)}\right),$$

  otherwise $v = u$. Denote proposal density associated with this draw of v by $p(v, u)$.
- Accept v with probability

$$\alpha_2(u, v) = \min\left(1, \frac{\pi(v)\, p(u, v)}{\pi(u)\, p(v, u)}\right) = \min\left(1, \frac{\pi(v)\, \pi^*(u)}{\pi(u)\, \pi^*(v)}\right),$$

  i.e. $U^{(i+1)} = v$ with probability $\alpha_2(u, v)$; otherwise stay at $U^{(i+1)} = u$.

- **Important.** $\Phi(v)$ only evaluated for proposals that were accepted for π^*.

Surrogate transition method [Liu, 2001]

- This method uses a surrogate posterior $\pi^*(u)$ to pre-screen proposals.

Surrogate transition method

- At state u, sample a proposal v^* from proposal density $p^*(u, \cdot)$.
- Set $v = v^*$ with probability

$$\alpha_1(u, v^*) = \min\left(1, \frac{\pi^*(v^*)\, p^*(u, v^*)}{\pi^*(u)\, p^*(v^*, u)}\right),$$

  otherwise $v = u$. Denote proposal density associated with this draw of v by $p(v, u)$.
- Accept v with probability

$$\alpha_2(u, v) = \min\left(1, \frac{\pi(v)\, p(u, v)}{\pi(u)\, p(v, u)}\right) = \min\left(1, \frac{\pi(v)\, \pi^*(u)}{\pi(u)\, \pi^*(v)}\right),$$

  i.e. $U^{(i+1)} = v$ with probability $\alpha_2(u, v)$; otherwise stay at $U^{(i+1)} = u$.

- **Important.** $\Phi(v)$ only evaluated for proposals that were accepted for π^*.
- The surrogate π^* can be, e.g., the posterior associated with a coarser discretisation ($h^* > h$ and/or $s^* < s$)

$\Rightarrow$ Multilevel MCMC [Dodwell, Ketelsen **RS**, Teckentrup, 2015], [Lykkegaard et al, 2023]

Summary – Part II

- MH algorithm with random walk or pCN proposals agnostic about the data and the likelihood.
- Better proposals are available that incorporate gradient or Hessian information abot the negative log-likelihood: **MALA, HMC, DILI, ...**
- To reduce cost we can 'prescreen' proposals with a cheaper surrogate using the **surrogate transition method** [Liu, 2001].
- Two particularly efficient methods of this type are the **multilevel MCMC method** [Dodwell, Ketelsen **RS**, Teckentrup, 2015] and the **multilevel delayed-acceptance algorithm** [Lykkegaard, Dodwell, Fox, Mingas, **RS**, 2023].
- However, this is a very active field of research and many new methods are being developped at the moment – **Watch this space!**

Summary – Part II

- MH algorithm with random walk or pCN proposals agnostic about the data and the likelihood.
- Better proposals are available that incorporate gradient or Hessian information abot the negative log-likelihood: **MALA, HMC, DILI, ...**
- To reduce cost we can 'prescreen' proposals with a cheaper surrogate using the **surrogate transition method** [Liu, 2001].
- Two particularly efficient methods of this type are the **multilevel MCMC method** [Dodwell, Ketelsen **RS**, Teckentrup, 2015] and the **multilevel delayed-acceptance algorithm** [Lykkegaard, Dodwell, Fox, Mingas, **RS**, 2023].
- However, this is a very active field of research and many new methods are being developped at the moment – **Watch this space!**

Thank you very much!

Main References and other Useful Literature

R Scheichl & J Zech, *Numerical Methods for Bayesian Inverse Problems*, Lecture Notes, Heidelberg University, 2021. Available in the git repository and at `https://katana.iwr.uni-heidelberg.de/pdfs/NM4BIP21_notes.pdf`

- M Dashti and & A Stuart, The Bayesian Approach To Inverse Problems, *Handbook of Uncertainty Quantification* (Ghanem, Higdon, Owhadi, Eds.), Springer, 2015.
- HW Engl, M Hanke & A Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, 2000.
- J Kaipio & E Somersalo, *Statistical and Computational Inverse Problems*, Springer, 2004.
- A Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, 2nd edition, Springer, 2011.
- J Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, 2001.
- CP Robert & G Casella, *Monte Carlo Statistical Methods*, 2nd edit., Springer, 2004.
- AM Stuart, Inverse problems: A Bayesian perspective, *Acta Numerica*, **19**, 2010.

Some recent more advanced publications

- T Cui, G Detommaso & R Scheichl, Multilevel dimension-independent likelihood- Informed MCMC for large-scale inverse problems, *Inverse Prob* **40**, 2024
- T Cui, S Dolgov, R Scheichl, Deep importance sampling using tensor-trains with application to a priori and a posteriori rare event estimation, *SIAM J Sci Comput*, **46**, 2024
- T Cui, KJH Law & YM Marzouk, Dimension-independent likelihood-informed MCMC, *J Comput Phys*, **304**, 2016
- TJ Dodwell, C Ketelsen, R Scheichl & AL Teckentrup, A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow, *SIAM/ASA J Uncertain Quant*, **3**, 2015
- S Dolgov, K Anaya-Izquierdo, C Fox & R Scheichl, Approximation and sampling of multivariate probability distributions in tensor train decomposition, *Stat Comput*, **30**, 2020
- VH Hoang, C Schwab & AM Stuart, Complexity analysis of accelerated MCMC methods for Bayesian inversion, *Inverse Prob*, **29**, 2013
- MB Lykkegaard, TJ Dodwell, C Fox, G Mingas & R Scheichl, Multilevel Delayed Acceptance MCMC, *SIAM/ASA J Uncertain Quant*, **11**, 2023
- D Rudolf & B Sprungk, On a generalization of the preconditioned Crank- Nicholson Metropolis algorithm, *Found Comput Math*, **18**, 2018
- R Scheichl, AM Stuart & AL Teckentrup, Quasi-MC and MLMC methods for computing posterior expectations in elliptic inverse problems, *SIAM J Uncertain Quantif*, **5**, 2017
- C Schillings, B Sprungk & P Wacker, On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems, *Numer Math*, **145**, 2020