The Convergence of the Laplace Approximation and Noise-Level-Robust Monte Carlo Methods for Bayesian Inverse Problems

Daniel Rudolf, Claudia Schillings, Björn Sprungk, Philipp Wacker

Institute of Mathematical Stochastics, University of Göttingen

Workshop "Optimization and Inversion under Uncertainty"
RICAM Linz, November 15th, 2019

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Bayesian Inverse Problems

- Infer unknown $x \in \mathbb{R}^d$ given noisy observations of forward map $G \colon \mathbb{R}^d \to \mathbb{R}^J$

$$y = G(x) + \varepsilon, \qquad \varepsilon \sim N(0, n^{-1}\Sigma), \qquad n \in \mathbb{N},$$

- Given prior measure μ_0 for x, here $\mu_0 = N(0, C_0)$, we obtain a posterior

$$\mu_n(\mathrm{d}x) = \frac{1}{Z_n} \exp(-n\Phi(x))\, \mu_0(\mathrm{d}x), \qquad \Phi(x) = \frac{1}{2} \left|y - G(x)\right|^2_{\Sigma^{-1}},$$

where $Z_n := \int_{\mathbb{R}^d} \mathrm{e}^{-n\Phi(x)} \, \mathrm{d}\mu_0$

- **Objective:** Sample (approximately) from μ_n and compute

$$\mathbb{E}_{\mu_n}[f] = \int_{\mathbb{R}^d} f(x)\, \mu_n(\mathrm{d}x), \qquad f \in L^1_{\mu_0}(\mathbb{R})$$

- In this talk we are interested in the case of increasing precision $n \to \infty$

Computational Bayesian Inference

- Computational **methods** for approximate sampling or integrating w.r.t. μ:
  - Markov chain Monte Carlo,
  - Importance sampling
  - Sequential Monte Carlo and particle filters,
  - Quasi-Monte Carlo and numerical quadrature, ...
- Common computational **challenges**:

Computational Bayesian Inference

- Computational **methods** for approximate sampling or integrating w.r.t. μ:
  - Markov chain Monte Carlo,
  - Importance sampling
  - Sequential Monte Carlo and particle filters,
  - Quasi-Monte Carlo and numerical quadrature, ...
- Common computational **challenges**:
  1. Expensive evaluation of forward model G
     $\rightarrow$ Multilevel or surrogate methods

Computational Bayesian Inference

- Computational **methods** for approximate sampling or integrating w.r.t. μ:
  - Markov chain Monte Carlo,
  - Importance sampling
  - Sequential Monte Carlo and particle filters,
  - Quasi-Monte Carlo and numerical quadrature, ...
- Common computational **challenges**:
  1. Expensive evaluation of forward model G
     $\rightarrow$ Multilevel or surrogate methods
  2. High-dimensional or even infinite-dimensional state space, e.g., function spaces
     $\rightarrow$ Intense research in recent years for all mentioned methods

Computational Bayesian Inference

- Computational **methods** for approximate sampling or integrating w.r.t. μ:
  - Markov chain Monte Carlo,
  - Importance sampling
  - Sequential Monte Carlo and particle filters,
  - Quasi-Monte Carlo and numerical quadrature, ...
- Common computational **challenges**:
  1. Expensive evaluation of forward model G
     $\rightarrow$ Multilevel or surrogate methods
  2. High-dimensional or even infinite-dimensional state space, e.g., function spaces
     $\rightarrow$ Intense research in recent years for all mentioned methods
  3. Concentrated μ_n due to informative data, i.e., $n \gg 1$ or $J \gg 1$
     $\rightarrow$ Analyzed so far in [Beskos et al., 2018] and [Schillings & Schwab, 2016]

Computational Bayesian Inference

- Computational **methods** for approximate sampling or integrating w.r.t. μ:
  - Markov chain Monte Carlo,
  - Importance sampling
  - Sequential Monte Carlo and particle filters,
  - Quasi-Monte Carlo and numerical quadrature, ...

- Common computational **challenges**:
  1. Expensive evaluation of forward model G
     $\rightarrow$ Multilevel or surrogate methods
  2. High-dimensional or even infinite-dimensional state space, e.g., function spaces
     $\rightarrow$ Intense research in recent years for all mentioned methods
  3. Concentrated μ_n due to informative data, i.e., $n \gg 1$ or $J \gg 1$
     $\rightarrow$ Analyzed so far in [Beskos et al., 2018] and [Schillings & Schwab, 2016]

Outline

Next

1. Laplace Approximation
2. Markov Chain Monte Carlo
3. Importance Sampling
4. Quasi Monte Carlo

General Approach For Noise-Level Robust Sampling

- Prior-based sampling or integration will suffer from the increasing difference between μ_n and μ_0 as $n \to \infty$, i.e.,

$$\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu_0} \propto \mathrm{e}^{-n\Phi} \to \delta_{\operatorname{argmin} \Phi} \quad \text{and} \quad d_{\mathrm{TV}}(\mu_n, \mu_0) \to 1$$

- **Idea:** Base sampling methods on a suitable (simple) reference measure mimicking the (increasing) concentration of μ_n

General Approach For Noise-Level Robust Sampling

- Prior-based sampling or integration will suffer from the increasing difference between μ_n and μ_0 as $n \to \infty$, i.e.,

$$\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu_0} \propto \mathrm{e}^{-n\Phi} \to \delta_{\operatorname{argmin} \Phi} \quad \text{and} \quad d_{\mathrm{TV}}(\mu_n, \mu_0) \to 1$$

- **Idea:** Base sampling methods on a suitable (simple) reference measure mimicking the (increasing) concentration of μ_n

- Here, **Laplace approximation** of μ_n: $\quad \mathscr{L}_{\mu_n} := N(x_n, C_n)$,

$$x_n := \operatorname*{argmin}_x n\Phi(x) + \frac{1}{2}\|C_0^{-1/2}x\|^2, \qquad C_n := \left(n\nabla^2\Phi(x_n) + C_0^{-1}\right)^{-1}$$

General Approach For Noise-Level Robust Sampling

- Prior-based sampling or integration will suffer from the increasing difference between μ_n and μ_0 as $n \to \infty$, i.e.,

$$\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu_0} \propto \mathrm{e}^{-n\Phi} \to \delta_{\operatorname{argmin}\Phi} \quad \text{and} \quad d_{TV}(\mu_n, \mu_0) \to 1$$

- **Idea:** Base sampling methods on a suitable (simple) reference measure mimicking the (increasing) concentration of μ_n

- Here, **Laplace approximation** of μ_n: $\mathcal{L}_{\mu_n} := N(x_n, C_n)$,

$$x_n := \operatorname*{argmin}_x n\Phi(x) + \frac{1}{2}\|C_0^{-1/2}x\|^2, \qquad C_n := \left(n\nabla^2\Phi(x_n) + C_0^{-1}\right)^{-1}$$

- Very common approximation in Bayesian statistics and OED ([Long et al., 2013], [Alexanderian et al., 2016], [Chen & Ghattas, 2017] ...)

Laplace's Method for Asymptotics of Integrals [Laplace, 1774]

- [Wong, 2001]: Considering integrals

$$J(n) := \int_D f(x) \exp(-n\Phi(x))\, \mathrm{d}x, \qquad D \subseteq \mathbb{R}^d$$

with sufficiently smooth f and Φ, we have, under suitable conditions, as $n \to \infty$

$$J(n) \;=\; \mathrm{e}^{-n\Phi(x_\star)}\; n^{-d/2} \left(\frac{f(x_\star)}{\sqrt{\det(2\pi H_\star)}} + \mathscr{O}(n^{-1}) \right)$$

where $x_\star := \operatorname{argmin}_{x \in \mathbb{R}^d} \Phi \in D$ and $H_\star := \nabla^2 \Phi(x_\star) > 0$

Laplace's Method for Asymptotics of Integrals [Laplace, 1774]

- [Wong, 2001]: Considering integrals

$$J(n) := \int_D f(x) \exp(-n\Phi(x)) \, \mathrm{d}x, \qquad D \subseteq \mathbb{R}^d$$

with sufficiently smooth f and Φ, we have, under suitable conditions, as $n \to \infty$

$$J(n) \; = \; \mathrm{e}^{-n\Phi(x_\star)} \; n^{-d/2} \left(\frac{f(x_\star)}{\sqrt{\det(2\pi H_\star)}} + \mathscr{O}(n^{-1}) \right)$$

where $x_\star := \operatorname{argmin}_{x \in \mathbb{R}^d} \Phi \in D$ and $H_\star := \nabla^2 \Phi(x_\star) > 0$

- **Yields:** Given smooth Lebesgue density of μ_0, then for suitable f

$$\left| \int_{\mathbb{R}^d} f \, \mathrm{d}\mu_n - \int_{\mathbb{R}^d} f \, \mathrm{d}N(x_\star, (nH_\star)^{-1}) \right| \in \mathscr{O}(n^{-1})$$

Convergence of Laplace Approximation

Theorem ([Schillings, S., Wacker, 2019])

Given that

- $\Phi \in C^3(\mathbb{R}^d)$, *unique* x_n *and* $C_n > 0$ *for sufficiently large* $n > 0$,
- *a unique minimizer* $x_\star := \operatorname{argmin}_{x\in\mathbb{R}^d} \Phi(x)$ *exists with* $\nabla^2\Phi(x_\star) > 0$ *and*

$$\inf_{\|x-x_\star\|>r} \Phi(x) \geq \Phi(x_\star) + \delta_r, \qquad \delta_r > 0,$$

- $\lim_{n\to\infty} x_n = x_\star$.

Then

$$d_H(\mu_n, \mathscr{L}_{\mu_n}) \in \mathscr{O}(n^{-1/2}).$$

Convergence of Laplace Approximation

Theorem ([Schillings, S., Wacker, 2019])

Given that

- $\Phi \in C^3(\mathbb{R}^d)$, *unique* x_n *and* $C_n > 0$ *for sufficiently large* $n > 0$,
- *a unique minimizer* $x_\star := \operatorname{argmin}_{x \in \mathbb{R}^d} \Phi(x)$ *exists with* $\nabla^2 \Phi(x_\star) > 0$ *and*

$$\inf_{\|x - x_\star\| > r} \Phi(x) \geq \Phi(x_\star) + \delta_r, \qquad \delta_r > 0,$$

- $\lim_{n \to \infty} x_n = x_\star$.

Then

$$d_H(\mu_n, \mathscr{L}_{\mu_n}) \in \mathcal{O}(n^{-1/2}).$$

Closely related to the Bernstein–von Mises theorem but:

- Covariance of $\mathscr{L}_{\mu_n}$ depends on given data (BvM: Fisher information)
- Misspecification ("ground truth" not in prior support) not important
- Density $\mathrm{d}\mu_n / \mathrm{d}\mathscr{L}_{\mu_n}$ also exists in Hilbert spaces (for Gaussian μ_0)

Remarks

The convergence theorem can be extended under suitable assumptions to

1. any prior μ_0 which is absolutely continuous w.r.t. Lebesgue measure,

2. sequences of Φ_n, e.g.,

$$\Phi_n(x) = \frac{1}{2n} \sum_{i=1}^{n} \|y_i - G(x)\|^2$$

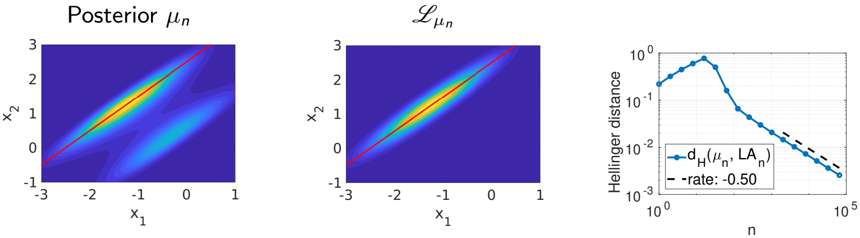
3. the underdetermined case $G: \mathbb{R}^d \to \mathbb{R}^J$, $J < d$, **iff** μ_0 is Gaussian and G acts only on linear active subspace $\mathscr{M}$ with $\dim(\mathscr{M}) \leq J$:

$$G(x + m) = G(x), \qquad \forall x \in \mathbb{R}^M \ \forall m \in \mathscr{M}^\perp$$

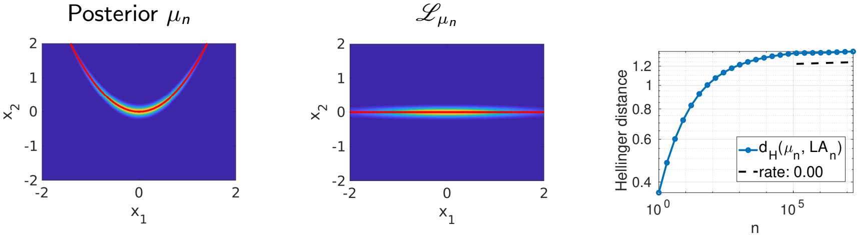
4. Approximations $\widetilde{x}_n, \widetilde{C}_n$ of x_n, C_n such that $\|x_n - \widetilde{x}_n\|, \|C_n - \widetilde{C}_n\| \in \mathscr{O}(n^{-1})$

Examples

- $\mu_0 = N(0, I_2)$, $\Phi(x) = \frac{1}{2}\|y - G(x)\|^2$, $G(x) = [\exp(\frac{1}{5}(x_2 - x_1)), \sin(x_2 - x_1)]^\top$



- $\mu_0 = N(0, I_2)$ and $\Phi(x) = \frac{1}{2}\|0 - G(x)\|^2$ with $G(x) = x_2 - x_1^2$



Next

1. Laplace Approximation
2. Markov Chain Monte Carlo
3. Importance Sampling
4. Quasi Monte Carlo

Markov Chain Monte Carlo (MCMC)

- Construct Markov chain $(X_m)_{m\in\mathbb{N}}$ with invariant measure μ_n, i.e.,

$$X_m \sim \mu_n \quad \Rightarrow \quad X_{m+1} \sim \mu_n$$

- Given suitable conditions, we have $X_m \xrightarrow[m\to\infty]{\mathscr{D}} \mu_n$ and for $f \in L^2_{\mu_0}(\mathbb{R})$

$$S_M(f) := \frac{1}{M}\sum_{m=1}^{M} f(X_m) \quad \xrightarrow[M\to\infty]{\text{a.s.}} \quad \mathbb{E}_{\mu_n}[f]$$

- Autocorrelation of Markov chain effects efficiency:

$$M\,\mathbb{E}\left[\left|S_M(f) - \mathbb{E}_{\mu_n}[f]\right|^2\right] \xrightarrow[M\to\infty]{} \operatorname{Var}_{\mu_n}(f)\underbrace{\left[1 + 2\sum_{m=0}^{\infty} \operatorname{Corr}\left(f(X_1), f(X_{1+m})\right)\right]}_{\text{integrated autocorrelation time (IACT)}}$$

Metropolis-Hastings (MH) algorithm [Metropolis et al., 1953]

Given current state $X_m = x$,

1. draw new state y according to proposal kernel $P(x, \cdot)$: $\quad Y_m \sim P(x)$
2. accept proposed y with acceptance probability $\alpha(x, y)$, i.e., set

$$X_{m+1} = \begin{cases} y, & \text{with probability } \alpha(x, y), \\ x, & \text{with probability } 1 - \alpha(x, y). \end{cases}$$

Metropolis-Hastings (MH) algorithm [Metropolis et al., 1953]

Given current state $X_m = x$,

1. draw new state y according to proposal kernel $P(x, \cdot)$: $\quad Y_m \sim P(x)$
2. accept proposed y with acceptance probability $\alpha(x, y)$, i.e., set

$$X_{m+1} = \begin{cases} y, & \text{with probability } \alpha(x, y), \\ x, & \text{with probability } 1 - \alpha(x, y). \end{cases}$$

- Correct $\alpha = \alpha_n$ for μ_n-invariance well-known
- Efficiency of MH algorithm depends entirely on "good" choice of proposal P

Metropolis-Hastings (MH) algorithm [Metropolis et al., 1953]

Given current state $X_m = x$,

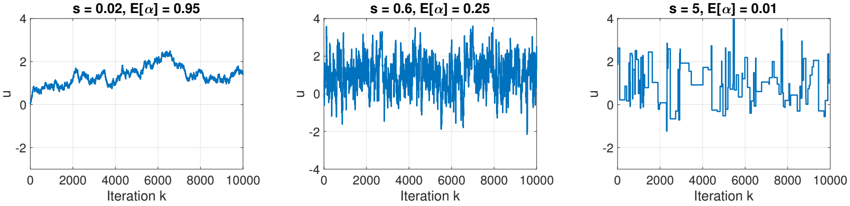
1. draw new state y according to proposal kernel $P(x, \cdot)$: $\quad Y_m \sim P(x)$
2. accept proposed y with acceptance probability $\alpha(x, y)$, i.e., set

$$X_{m+1} = \begin{cases} y, & \text{with probability } \alpha(x, y), \\ x, & \text{with probability } 1 - \alpha(x, y). \end{cases}$$

- Correct $\alpha = \alpha_n$ for μ_n-invariance well-known
- Efficiency of MH algorithm depends entirely on “good” choice of proposal P
- Construct proposals s. th. efficiency/autocorrelation is robust w.r.t. $n \to \infty$

Gaussian Random Walk-MH

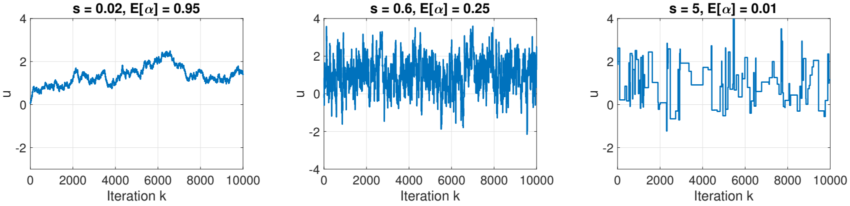
- Proposal kernel: $P(x) = N(x, s^2 C_0)$ with tunable stepsize $s > 0$:



- If $\pi_n \colon \mathbb{R}^d \to (0, \infty)$ denotes density of μ_n: $\alpha_n(x, y) = \min\left\{1, \frac{\pi_n(y)}{\pi_n(x)}\right\}$

Gaussian Random Walk-MH

- Proposal kernel: $P(x) = N(x, s^2 C_0)$ with tunable stepsize $s > 0$:



- If $\pi_n : \mathbb{R}^d \to (0, \infty)$ denotes density of μ_n: $\alpha_n(x, y) = \min\left\{1, \frac{\pi_n(y)}{\pi_n(x)}\right\}$

- **Dimension-robust** version: pCN-proposal [Beskos et al., 2008]

$$P(x) = N(\sqrt{1 - s^2}x, s^2 C_0), \qquad s \in (0, 1],$$

is μ_0-reversible which yields $\alpha_n(x, y) = \min\left\{1, \left(\frac{\exp(-\Phi(y))}{\exp(-\Phi(x))}\right)^n\right\}$

Idea for Noise-Level Robust MH Algorithms

- Inform proposal P about (increasing) concentration of μ_n by using (an approximation of) posterior covariance for proposing

  (cf. [Tierney, 1994], [Haario et al., 2001], [Martin et al., 2012]...)

Idea for Noise-Level Robust MH Algorithms

- Inform proposal P about (increasing) concentration of μ_n by using (an approximation of) posterior covariance for proposing
  (cf. [Tierney, 1994], [Haario et al., 2001], [Martin et al., 2012]...)

- Here, we use the covariance C_n of the Laplace approximation $\mathscr{L}_{\mu_n}$

- [Rudolf & S., 2018]: Candidates for noise lebel-robust RW- & pCN-variants

$$\textbf{H-RW:} \qquad P_n(x) = N(x, s^2 C_n),$$
$$\textbf{generalized pCN (gpCN):} \qquad P_n(x) = N(A_{s,n}x, s^2 C_n)$$

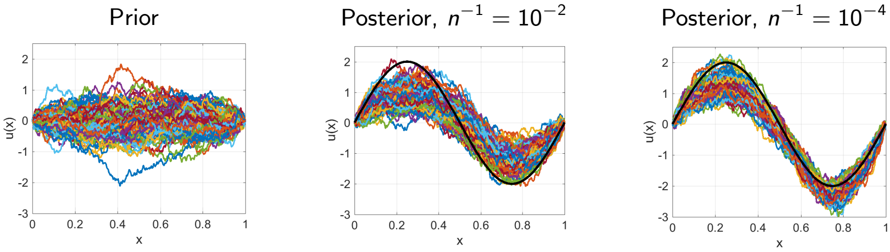
  where bounded linear operator $A_{s,n}$ ensures μ_0-reversibility
  (cf. operator weighted proposals [Law, 2013] and [Cui et al., 2016])

Numerical Experiment

- **Problem:** Infer coefficient in 1D BVP by observing solution at 4 points
- **Proposals:**

$$\text{RW: } P_0(x) = N(x, s^2 C_0), \qquad \text{pCN: } P_0(x) = N(\sqrt{1-s^2}x, s^2 C_0),$$
$$\text{H-RW: } P_n(x) = N(x, s^2 C_n), \qquad \text{gpCN: } P_n(x) = N(A_s x, s^2 C_n)$$

- **Results:**



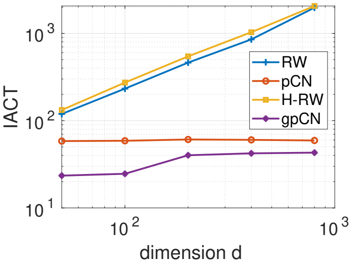
Numerical Experiment

- **Problem:** Infer coefficient in 1D BVP by observing solution at 4 points
- **Proposals:**

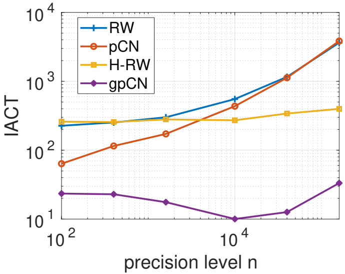
$$\text{RW: } P_0(x) = N(x, s^2 C_0), \qquad \text{pCN: } P_0(x) = N(\sqrt{1-s^2}x, s^2 C_0),$$
$$\text{H-RW: } P_n(x) = N(x, s^2 C_n), \qquad \text{gpCN: } P_n(x) = N(A_s x, s^2 C_n)$$

- **Results:**

IACT vs. dimension



IACT vs. precision level



Noise-Level Robustness of MH Algorithms

- Given μ_n-invariant Markov chains $(X_m)_{m\in\mathbb{N}}$ we can study if

$$\lim_{n\to\infty} \sum_{m=0}^{\infty} \operatorname{Corr}\left(f(X_1), f(X_{1+m})\right) < \infty, \qquad f \in L^2_{\mu_0}(\mathbb{R})$$

Noise-Level Robustness of MH Algorithms

- Given μ_n-invariant Markov chains $(X_m)_{m\in\mathbb{N}}$ we can study if

$$\lim_{n\to\infty} \sum_{m=0}^{\infty} \mathrm{Corr}\left(f(X_1), f(X_{1+m})\right) < \infty, \qquad f \in L^2_{\mu_0}(\mathbb{R})$$

- To start, we consider simpler efficiency indicators:

Mean acceptance rate: $\mathbb{E}\left[\alpha_n(X_m, Y_m)\right],$

Lag-1-Autocorrelation: $\mathrm{Corr}(a^\top X_m, a^\top X_{m+1}), \; a \in \mathbb{R}^d$

- Noise-level robust efficiency defined as

$$\lim_{n\to\infty} \mathbb{E}\left[\alpha_n(X_m, Y_m)\right] > 0, \qquad \lim_{n\to\infty} \mathrm{Corr}(a^\top X_m, a^\top X_{m+1}) < 1$$

Noise-Level Robustness of MH Algorithms cont'd

- [S., 2017]: For Gaussian posteriors $\mu_n = \mathscr{L}_{\mu_n} = N(x_n, C_n)$ the proposals

$$P_n(x) = N(x, s^2 C_n), \qquad P_n(x) = N(A_{s,n} x, s^2 C_n)$$

yield

$$\lim_{n\to\infty} \mathbb{E}\left[\alpha_n(X_m, Y_m)\right] > 0, \qquad \lim_{n\to\infty} \mathrm{Corr}(a^\top X_m, a^\top X_{m+1}) < 1 \tag{1}$$

Noise-Level Robustness of MH Algorithms cont'd

- [S., 2017]: For Gaussian posteriors $\mu_n = \mathscr{L}_{\mu_n} = N(x_n, C_n)$ the proposals

$$P_n(x) = N(x, s^2 C_n), \qquad P_n(x) = N(A_{s,n} x, s^2 C_n)$$

yield

$$\lim_{n\to\infty} \mathbb{E}\left[\alpha_n(X_m, Y_m)\right] > 0, \quad \lim_{n\to\infty} \mathrm{Corr}(a^\top X_m, a^\top X_{m+1}) < 1 \tag{1}$$

- Convergence of the Laplace approximation lifts this to the non-Gaussian case:

Theorem ([Rudolf, S., 2019])

Given $d_H(\mu_n, \mathscr{L}_{\mu_n}) \to 0$ we have for the H-RW and gpCN proposal

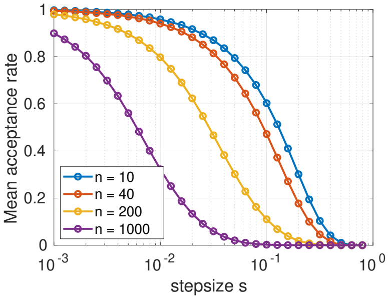
$$P_n(u) = N(u, s^2 C_n), \qquad P_n(u) = N(A_{s,n} u, s^2 C_n),$$

that (1) *holds.*

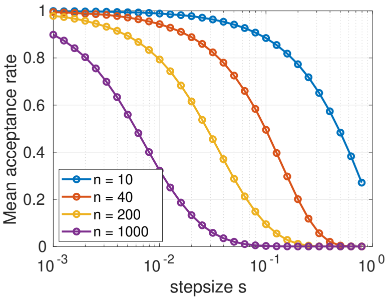
Numerical Experiment for Increasing Concentration

- Linear forward map G (convolution operator) applied to unknown function
- Gaussian prior and noise $\varepsilon \sim N(0, n^{-1} I_4)$ yield Gaussian posterior
- Examine mean acceptance rate vs. proposal stepsize s:

$$P(u) = N(u, s^2 C_0)$$



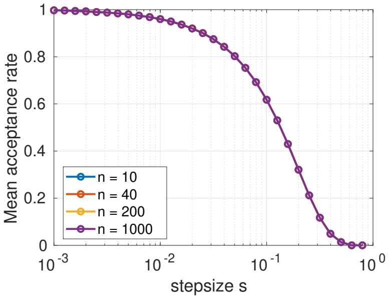
$$P(u) = N(\sqrt{1 - s^2}u, s^2 C_0)$$



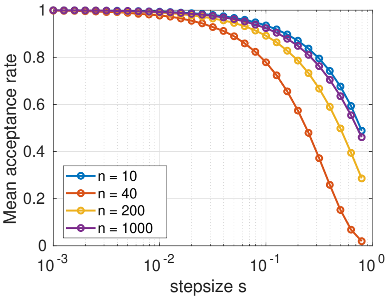
Numerical Experiment for Increasing Concentration

- Linear forward map G (convolution operator) applied to unknown function
- Gaussian prior and noise $\varepsilon \sim N(0, n^{-1} I_4)$ yield Gaussian posterior
- Examine mean acceptance rate vs. proposal stepsize s:

$P(u) = N(u, s^2 C_n)$



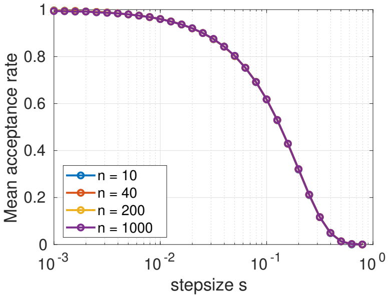
$P(u) = N(A_{s,n} u, s^2 C_n)$



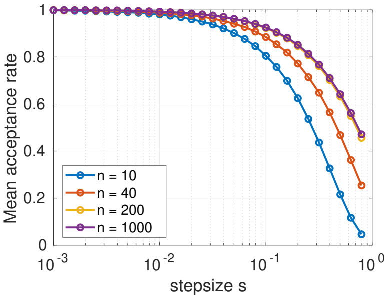
Numerical Experiment for Increasing Concentration

- **Nonlinear** forward map G ($\exp \circ$ convolution operator)
- Gaussian prior and noise $\varepsilon \sim N(0, n^{-1} I_4)$ yield **non**-Gaussian posterior
- Examine mean acceptance rate vs. proposal stepsize s:

$$P(u) = N(u, s^2 C_n)$$



$$P(u) = N(A_{s,n}u, s^2 C_n)$$



Next

Self-Normalizing Importance Sampling

- Given importance distribution ν and i.i.d. samples $X_m \sim \nu$, $m = 1, \ldots, M$, use

$$\mathbb{E}_{\mu_n}[f] \;=\; \frac{\int_{\mathbb{R}^d} f \; \mathrm{e}^{-n\Phi} \mathrm{d}\mu_0}{\int_{\mathbb{R}^d} \mathrm{e}^{-n\Phi} \mathrm{d}\mu_0} \approx \frac{\sum_{m=1}^{M} w_n(X_m) \; f(X_m)}{\sum_{i=1}^{M} w_n(X_m)} \qquad w_n \propto \frac{\mathrm{d}\mu_n}{\mathrm{d}\nu}$$

Self-Normalizing Importance Sampling

- Given importance distribution ν and i.i.d. samples $X_m \sim \nu$, $m = 1, \ldots, M$, use

$$\mathbb{E}_{\mu_n}[f] \;=\; \frac{\int_{\mathbb{R}^d} f \; \mathrm{e}^{-n\Phi} \mathrm{d}\mu_0}{\int_{\mathbb{R}^d} \mathrm{e}^{-n\Phi} \mathrm{d}\mu_0} \approx \frac{\sum_{m=1}^{M} w_n(X_m)\; f(X_m)}{\sum_{i=1}^{M} w_n(X_m)} \qquad w_n \propto \frac{\mathrm{d}\mu_n}{\mathrm{d}\nu}$$

- SLLN yields $\frac{\sum_{m=1}^{M} w_n(X_m)\; f(X_m)}{\sum_{i=1}^{M} w_n(X_m)} \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_{\mu_n}[f]$ and given that

$$V_{\mu_n,\nu}(f) := \mathbb{E}_\nu \left[\left(\frac{\mathrm{d}\mu_n}{\mathrm{d}\nu} \right)^2 (f - \mathbb{E}_{\mu_n}[f])^2 \right] < \infty$$

there holds a CLT with asymptotic variance $V_{\mu_n,\nu}(f)$ as $M \to \infty$

Self-Normalizing Importance Sampling

- Given importance distribution ν and i.i.d. samples $X_m \sim \nu$, $m = 1, \ldots, M$, use

$$\mathbb{E}_{\mu_n}[f] \;=\; \frac{\int_{\mathbb{R}^d} f \; \mathrm{e}^{-n\Phi} \mathrm{d}\mu_0}{\int_{\mathbb{R}^d} \mathrm{e}^{-n\Phi} \mathrm{d}\mu_0} \approx \frac{\sum_{m=1}^{M} w_n(X_m) \; f(X_m)}{\sum_{i=1}^{M} w_n(X_m)} \qquad w_n \propto \frac{\mathrm{d}\mu_n}{\mathrm{d}\nu}$$

- SLLN yields $\frac{\sum_{m=1}^{M} w_n(X_m) \; f(X_m)}{\sum_{i=1}^{M} w_n(X_m)} \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}_{\mu_n}[f]$ and given that

$$V_{\mu_n,\nu}(f) := \mathbb{E}_\nu \left[\left(\frac{\mathrm{d}\mu_n}{\mathrm{d}\nu} \right)^2 (f - \mathbb{E}_{\mu_n}[f])^2 \right] < \infty$$

  there holds a CLT with asymptotic variance $V_{\mu_n,\nu}(f)$ as $M \to \infty$

- How does $V_{\mu_n,\nu}(f)$ behave as $n \to \infty$ for suitable ν?

Prior Importance Sampling

- Choose prior measure as importance distribution $\nu = \mu_0$, i.e.,

$$X_m \sim \mu_0 \text{ i.i.d. }, \quad w_n(x) = \exp(-n\Phi(x))$$

- Asymptotic variance of prior importance sampling given by

$$V_{\mu_n,\mu_0}(f) = \frac{1}{Z_n^2} \int_{\mathbb{R}^d} (f - \mathbb{E}_{\mu_n}[f])^2 \, \mathrm{e}^{-2n\Phi} \, \mathrm{d}\mu_0, \qquad Z_n = \int_{\mathbb{R}^d} \mathrm{e}^{-n\Phi} \, \mathrm{d}\mu_0$$

Prior Importance Sampling

- Choose prior measure as importance distribution $\nu = \mu_0$, i.e.,

$$X_m \sim \mu_0 \text{ i.i.d. }, \quad w_n(x) = \exp(-n\Phi(x))$$

- Asymptotic variance of prior importance sampling given by

$$V_{\mu_n,\mu_0}(f) = \frac{1}{Z_n^2} \int_{\mathbb{R}^d} (f - \mathbb{E}_{\mu_n}[f])^2 \, \mathrm{e}^{-2n\Phi} \, \mathrm{d}\mu_0, \qquad Z_n = \int_{\mathbb{R}^d} \mathrm{e}^{-n\Phi} \, \mathrm{d}\mu_0$$

Theorem ([Schillings, S., Wacker, 2019])

Given $d_H(\mu_n, \mathscr{L}_{\mu_n}) \to 0$*, sufficiently smooth* Φ *and* $f \in L^1_{\mu_0}(\mathbb{R})$ *with* $\nabla f(x_\star) \neq 0$*, then*

$$V_{\mu_n,\mu_0}(f) \sim n^{d/2-1}, \qquad n \to \infty.$$

$\Rightarrow$ Prior importance sampling becomes **less efficient** as posterior concentrates

Laplace-Based Importance Sampling

- Choose $\nu = \mathscr{L}_{\mu_n}$, i.e.,

$$X_m \sim N(x_n, C_n), \qquad w_n(x) = \exp(-n[\Phi(x) - T_2\Phi(x; x_n)])$$

  where $T_2\Phi(\cdot; x_n)$ denotes Taylor polynomial of order 2 of Φ at MAP point x_n

- Applied, e.g., for fast Bayesian optimal experimental design [Beck et al., 2018]

Laplace-Based Importance Sampling

- Choose $\nu = \mathscr{L}_{\mu_n}$, i.e.,

$$X_m \sim N(x_n, C_n), \qquad w_n(x) = \exp(-n[\Phi(x) - T_2\Phi(x; x_n)])$$

  where $T_2\Phi(\cdot; x_n)$ denotes Taylor polynomial of order 2 of Φ at MAP point x_n

- Applied, e.g., for fast Bayesian optimal experimental design [Beck et al., 2018]

- Existence of $V_{\mu_n, \mathscr{L}_{\mu_n}}(f)$ not ensured & requires at least quadratic growth of Φ

Laplace-Based Importance Sampling

- Choose $\nu = \mathscr{L}_{\mu_n}$, i.e.,

$$X_m \sim N(x_n, C_n), \qquad w_n(x) = \exp(-n[\Phi(x) - T_2\Phi(x; x_n)])$$

where $T_2\Phi(\cdot; x_n)$ denotes Taylor polynomial of order 2 of Φ at MAP point x_n

- Applied, e.g., for fast Bayesian optimal experimental design [Beck et al., 2018]
- Existence of $V_{\mu_n, \mathscr{L}_{\mu_n}}(f)$ not ensured & requires at least quadratic growth of Φ

Theorem ([Schillings, S., Wacker, 2019])

Given $d_H(\mu_n, \mathscr{L}_{\mu_n}) \to 0$, *sufficiently smooth* Φ, *and* $f \in L^2_{\mu_0}(\mathbb{R})$, *we have*

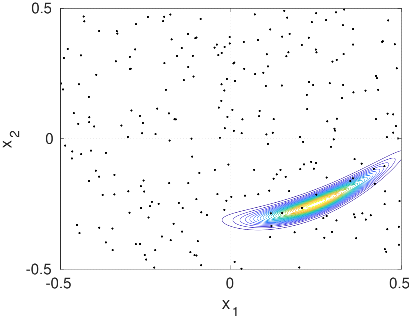
$$\left| \frac{\sum_{m=1}^M w_n(X_m)\ f(X_m)}{\sum_{m=1}^M w_n(X_m)} - \mathbb{E}_{\mu_n}[f] \right| \in o_{\mathbb{P}}(n^{-\delta}), \qquad \delta < 1/2.$$

$\Rightarrow$ Laplace-based importance sampling becomes **more efficient** as $n \to \infty$

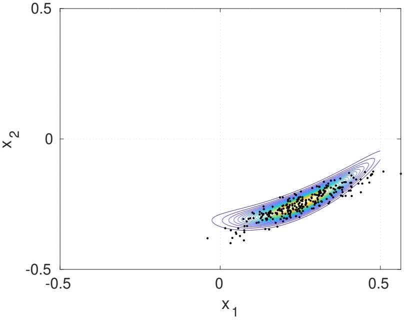
Simple Example

Prior: $\mu_0 = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$, noise: $\varepsilon \sim N(0, n^{-1} I_d)$, forward: $G = (G_1, \ldots, G_d)$,

$$G_1(x) = \exp(x_1/5), \quad G_2(x) = x_2 - x_1^2, \quad G_3(x) = x_3, \quad G_4(x) = 2x_4 + x_1^2$$



256 prior samples

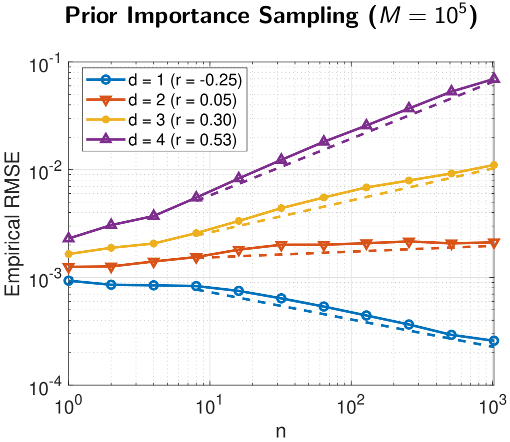


256 Laplace-based samples

Simple Example

Prior: $\mu_0 = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$, noise: $\varepsilon \sim N(0, n^{-1} I_d)$, forward: $G = (G_1, \ldots, G_d)$,

$$G_1(x) = \exp(x_1/5), \quad G_2(x) = x_2 - x_1^2, \quad G_3(x) = x_3, \quad G_4(x) = 2x_4 + x_1^2$$

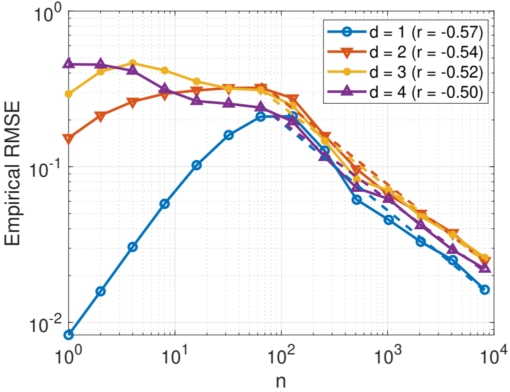


Simple Example

Prior: $\mu_0 = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$, noise: $\varepsilon \sim N(0, n^{-1} I_d)$, forward: $G = (G_1, \ldots, G_d)$,

$$G_1(x) = \exp(x_1/5), \quad G_2(x) = x_2 - x_1^2, \quad G_3(x) = x_3, \quad G_4(x) = 2x_4 + x_1^2$$

Laplace-based Importance Sampling ($M = 10^5$)



Next

1. Laplace Approximation

2. Markov Chain Monte Carlo

3. Importance Sampling

4. Quasi Monte Carlo

Quasi-Monte Carlo Integration

- For uniform prior $\mu_0 = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$ approximate integrals

$$\int_{[-\frac{1}{2},\frac{1}{2}]^d} f \, \mathrm{e}^{-n\Phi} \, \mathrm{d}\mu_0 \approx \frac{1}{M} \sum_{m=1}^{M} \mathrm{e}^{-n\Phi(X_m)} \, f(X_m)$$

using randomly shifted lattice rules [Sloan, Kuo, Joe, 2002] where

$$X_m = \mathrm{frac}\Big(\frac{mz}{M} + \Delta\Big) - \frac{1}{2}, \quad z \in \{1, \dots, N-1\}^d, \quad \Delta \sim \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$$

Quasi-Monte Carlo Integration

- For uniform prior $\mu_0 = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$ approximate integrals

$$\int_{[-\frac{1}{2},\frac{1}{2}]^d} f \, \mathrm{e}^{-n\Phi} \, \mathrm{d}\mu_0 \approx \frac{1}{M} \sum_{m=1}^{M} \mathrm{e}^{-n\Phi(X_m)} \, f(X_m)$$

  using randomly shifted lattice rules [Sloan, Kuo, Joe, 2002] where

$$X_m = \mathrm{frac}\left(\frac{mz}{M} + \Delta\right) - \frac{1}{2}, \quad z \in \{1, \dots, N-1\}^d, \quad \Delta \sim \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$$

- **Problem:** For increasing $n \to \infty$ the usual bound for the mean squared error

$$\mathbb{E}\left[\left|Z_n - \frac{1}{M} \sum_{m=1}^{M} \mathrm{e}^{-n\Phi(X_m)}\right|^2\right]$$

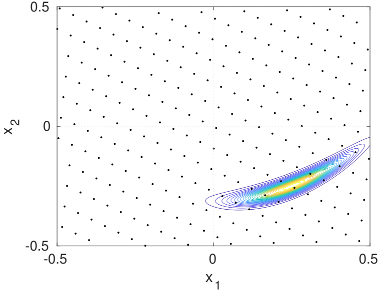
  behaves like $n^{d/2}$ [Schillings, S., Wacker, 2019]

Laplace-based Quasi-Monte Carlo

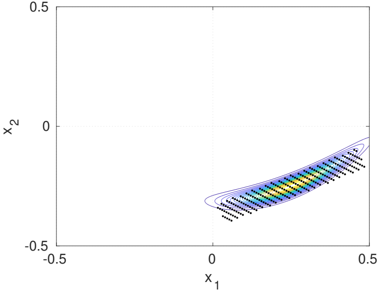
Apply Laplace-based transform

$$T_n(x) := x_n + \tau C_n^{1/2} x, \qquad x \in [-\frac{1}{2}, \frac{1}{2}]^d$$

to move lattice points X_m where μ_n concentrates (τ ensuring $T_n(x) \in [-\frac{1}{2}, \frac{1}{2}]^d$).



256 shifted lattice points



Transformed points

Laplace-based Quasi-Monte Carlo

Apply Laplace-based transform

$$T_n(x) := x_n + \tau C_n^{1/2} x, \qquad x \in [-\frac{1}{2}, \frac{1}{2}]^d$$

to move lattice points X_m where μ_n concentrates (τ ensuring $T_n(x) \in [-\frac{1}{2}, \frac{1}{2}]^d$).

Lemma ([Schillings, S., Wacker, 2019])

Given $d_H(\mathscr{L}_{\mu_n}, \mu_n) \to 0$ and sufficiently smooth Φ we obtain for the transformed shifted lattice rule

$$\frac{1}{Z_n^2} \mathbb{E}\left[\left| Z_n - \frac{\det(\tau C_n^{1/2})}{M} \sum_{m=1}^{M} \mathrm{e}^{-n\Phi(T_n(X_m))} \right|^2\right] \leq C(\tau, M) \in \mathscr{O}(n^0).$$

$\Rightarrow$ Bounded relative error for computing decaying $Z_n \to 0$

Simple Example cont'd

Prior: $\mu_0 = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$, noise: $\varepsilon \sim N(0, n^{-1} I_d)$, forward: $G = (G_1, \dots, G_d)$,

$$G_1(x) = \exp(x_1/5), \quad G_2(x) = x_2 - x_1^2, \quad G_3(x) = x_3, \quad G_4(x) = 2x_4 + x_1^2$$

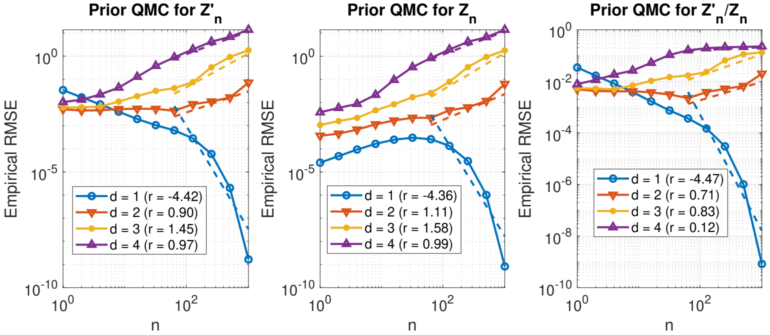
Relative errors for: Z_n, $Z_n' = \int_{[-\frac{1}{2}, \frac{1}{2}]^d} f \, \mathrm{e}^{-n\Phi} \, \mathrm{d}\mu_0$, $\mathbb{E}_{\mu_n}[f] = \frac{Z_n'}{Z_n}$

Simple Example cont'd

Prior: $\mu_0 = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$, noise: $\varepsilon \sim N(0, n^{-1} I_d)$, forward: $G = (G_1, \dots, G_d)$,

$$G_1(x) = \exp(x_1/5), \quad G_2(x) = x_2 - x_1^2, \quad G_3(x) = x_3, \quad G_4(x) = 2x_4 + x_1^2$$

Relative errors for: Z_n, $Z_n' = \int_{[-\frac{1}{2},\frac{1}{2}]^d} f \, \mathrm{e}^{-n\Phi} \, \mathrm{d}\mu_0$, $\mathbb{E}_{\mu_n}[f] = \frac{Z_n'}{Z_n}$

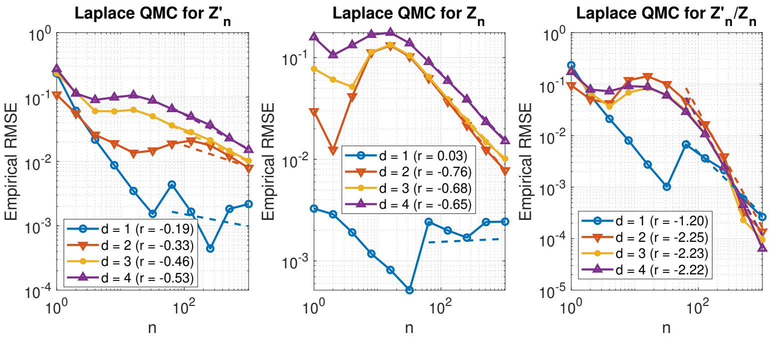


Simple Example cont'd

Prior: $\mu_0 = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}]^d)$, noise: $\varepsilon \sim N(0, n^{-1} I_d)$, forward: $G = (G_1, \ldots, G_d)$,

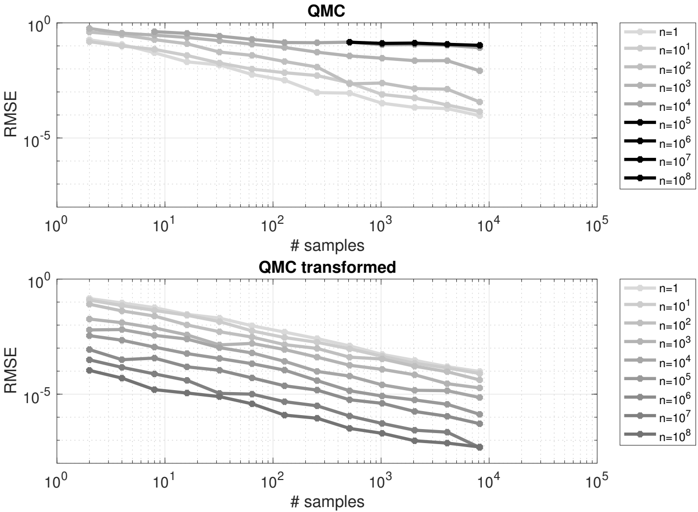
$$G_1(x) = \exp(x_1/5), \quad G_2(x) = x_2 - x_1^2, \quad G_3(x) = x_3, \quad G_4(x) = 2x_4 + x_1^2$$

Relative errors for: Z_n, $Z_n' = \int_{[-\frac{1}{2}, \frac{1}{2}]^d} f \, \mathrm{e}^{-n\Phi} \, \mathrm{d}\mu_0$, $\mathbb{E}_{\mu_n}[f] = \frac{Z_n'}{Z_n}$



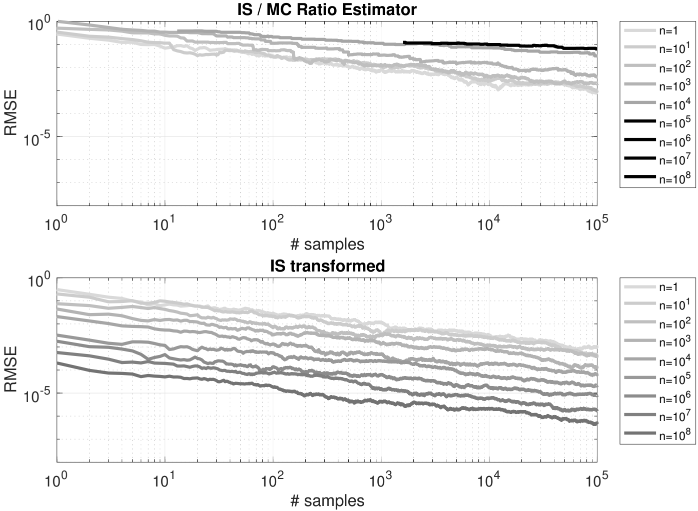
Example: Lognormal Elliptic PDE

Computing posterior mean of log coefficient given noisy data with $\varepsilon \sim N(0, \frac{1}{n} I_d)$:



Example: Lognormal Elliptic PDE

Computing posterior mean of log coefficient given noisy data with $\varepsilon \sim N(0, \frac{1}{n} I_d)$:



Summary

- Bayesian inference with informative data requires noise-level robust sampling
- Prior-based sampling methods suffer from a decreasing observational noise
- Robust sampling methods obtainable by using the Laplace approximation
- First theoretical results on noise-level robustness of importance sampling, MCMC, and QMC

Some open issues:

- Spectral gap-robustness for Laplace-based MCMC
- Convergence of Laplace approximation and sampling analysis in Hilbert spaces
- Beyond Laplace: What to do if posterior concentrates along nonlinear manifolds?

References

A. Alexanderian, N. Petra, G. Stadler, O. Ghattas. A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems. *SIAM J. Sci. Comput.* 38(1):A243–A272 (2016).

A. Beskos, G. O. Roberts, A. Thiery, and N. Pillai. Asymptotic Analysis of the Random-Walk Metropolis Algorithm on Ridged Densities *Ann. Appl. Probab.* 28(5):2966–3001 (2018).

Q. Long, M. Scavino, R. Tempone, S. Wang. Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering* 259:24–39 (2013).

P. Chen, U. Villa, and O. Ghattas. Hessian-based adaptive sparse quadrature for infinite-dimensional Bayesian inverse problems. *Computer Methods in Applied Mechanics and Engineering*, 327:147–172 (2017).

D. Rudolf, B. Sprungk. On a generalization of the preconditioned Crank-Nicolson Metropolis algorithm. *Found. Comput. Math.*, 18(2):309–343 (2018).

C. Schillings, Ch. Schwab. Scaling limits in computational Bayesian inversion. *ESAIM M2AN*, 50(6):1825–1856 (2016).

C. Schillings, B. Sprungk and P. Wacker. On the Convergence of the Laplace Approximation and Noise-Level-Robustness of Laplace-based Monte Carlo Methods for Bayesian Inverse Problems. arXiv:1901:03958, (2019).