On the Convergence of Laplace's Approximation and Its Implications for Bayesian Computation

Claudia Schillings, Björn Sprungk, Philipp Wacker

UQ for inverse problems in complex systems

Uncertainty quantification for complex systems: theory and methodologies

Outline

Bayesian Inverse Problem

Find the unknown data $u \in X$ from noisy observations

$$y = \mathcal{G}(u) + \eta$$

- $u \in X$ parameter function
- $\mathcal{G} : X \to Y$ forward response operator
- y observations, here $y \in Y = \mathbb{R}^K$
- evaluation of $\mathcal{G}$ expensive
- prior $u \sim \mu_0$
- noise model $\eta \sim \mathcal{N}(0, \Gamma)$

Bayesian Inverse Problem

Find the unknown data $u \in X$ from noisy observations

$$y = \mathcal{G}(u) + \eta$$

Assuming $\mathcal{G} \in C(X, Y)$ and $\mu_0(X) = 1$, then the posterior measure μ^y on $u|y$ is absolutely continuous w.r. to the prior on u and

$$\mu^y(du) = \frac{1}{Z} \exp(-\Phi(u))\mu_0(du)$$

with $\Phi : X \mapsto \mathbb{R}$, $\Phi(u) = \frac{1}{2}|y - \mathcal{G}(u)|_\Gamma^2$ and $Z = \int \exp(-\Phi(u))\mu_0(du)$.

Bayesian Inverse Problem

Find the unknown data $u \in X$ from noisy observations

$$y = \mathcal{G}(u) + \eta$$

Algorithms

- **MCMC**
  - dimension robust versions, multilevel strategies
- **Importance Sampling / Ratio Estimator**
  - MC, QMC, Sparse Grids
  - dimension robust versions, multilevel strategies

Bayesian Inverse Problem

Find the unknown data $u \in X$ from noisy observations

$$y = \mathcal{G}(u) + \eta$$

Small Noise Limit / Large Data Limit

- Concentration effect of the posterior
- Nonrobust behavior of the sampling methods w.r. to the size of the noise
- Highly desirable situation in practice (identification of the unknown parameters, design of experiments)

Design of robust methods via preconditioning using Laplace approximation

Bayesian Inverse Problem

Find the unknown data $u \in X$ from noisy observations

$$y = \mathcal{G}(u) + \eta$$

Let $X = \mathbb{R}^d$, $\mu_0(\mathrm{d}u) = \pi_0(u)\mathrm{d}u$ and $I(u) := \Phi(u) - \log \pi_0(u)$. Assume that $I \in C^2(\mathbb{R}^d; \mathbb{R})$, then the **Laplace approximation** of μ^y is given by the Gaussian measure

$$\tilde{\mu}^y := \mathcal{N}\left(u_{MAP}, C\right) ,$$

where u_{MAP} denotes the maximum a-posteriori estimate

$$u_{MAP} := \underset{u \in \{u \in \mathbb{R}^d : \pi_0(u) \neq 0\}}{\operatorname{argmin}} I(u), \qquad C^{-1} := \nabla^2 I(u_{MAP}).$$

Motivational Example

Model parametric elliptic problem

$$-\mathrm{div}(\hat{u}\nabla q) = f \quad \text{in } D := [0,1]\,,\ q = 0 \quad \text{in } \partial D\,,$$

with $f(x) = 100 \cdot x$ and

$$\hat{u}(x,y) = \exp(\psi_1(x)u_1 + \psi_2(x)u_2)\,,$$

with $\psi_1(x) = 0.1\sin(\pi x)$, $\psi_2(x) = 0.05\cos(\pi x)$ and $u_j \sim \mathcal{N}(0,1),\ j \in \mathbb{J}$.

Motivational Example

For given (noisy) data y,

$$y = \mathcal{G}(u) + \eta\,,$$

we are interested in the behavior of the posterior

$$\Theta(u) = \exp\big(-\Phi_{\Gamma_{obs}}(u; y)\big)$$

with

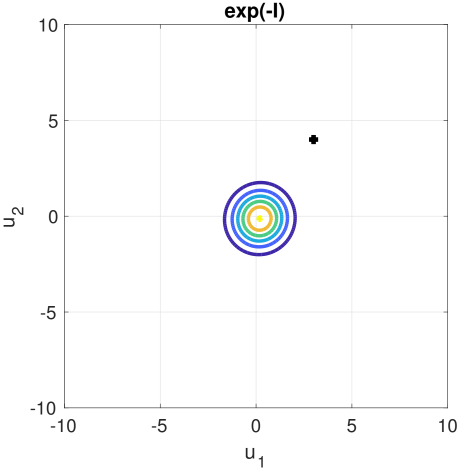
$$\Phi_{\Gamma_{obs}}(u; y) = \frac{1}{2}\Big((y - \mathcal{G}(u))^\top \Gamma_{obs}^{-1}(y - \mathcal{G}(u))\Big)$$

and variance of the noise $\Gamma_{obs} = \lambda \cdot id = 1/n \cdot id$. We set

$$I_{\Gamma_{obs}}(u; y) = \frac{1}{2}\Big((y - \mathcal{G}(u))^\top \Gamma_{obs}^{-1}(y - \mathcal{G}(u))\Big) + \frac{1}{2}\|u\|^2\;.$$

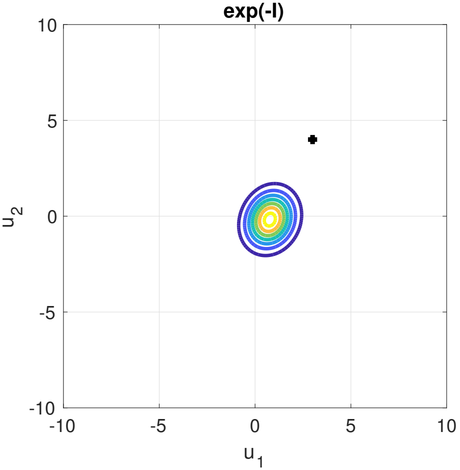
- QoI ϕ is the solution of the forward problem at $x = 0.5$.
- Observation operator $\mathcal{O}$ consists of 2 system responses at $x = 0.25$ and $x = 0.75$.

Motivational Example



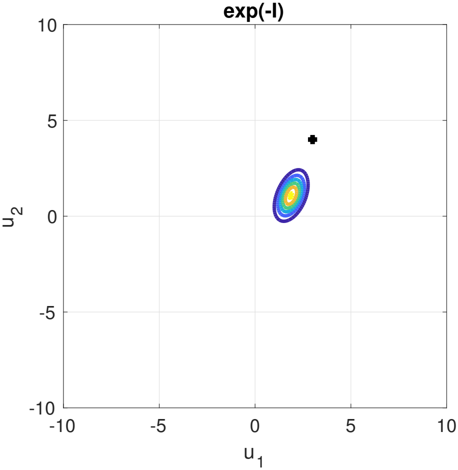
Contour plot of the posterior with observational noise $\lambda = 10^{-0}$.

Motivational Example



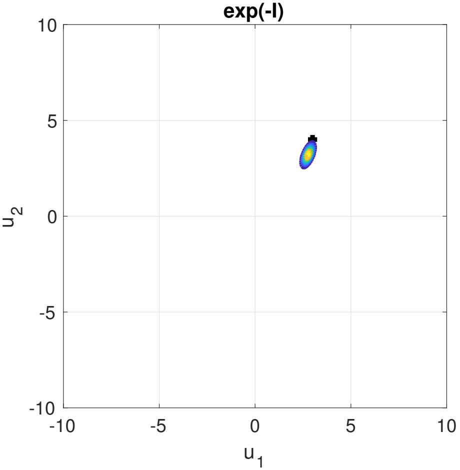
Contour plot of the posterior with observational noise $\lambda = 10^{-1}$.

Motivational Example



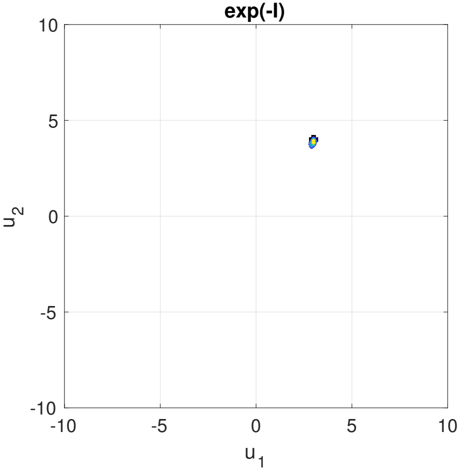
Contour plot of the posterior with observational noise $\lambda = 10^{-2}$.

Motivational Example



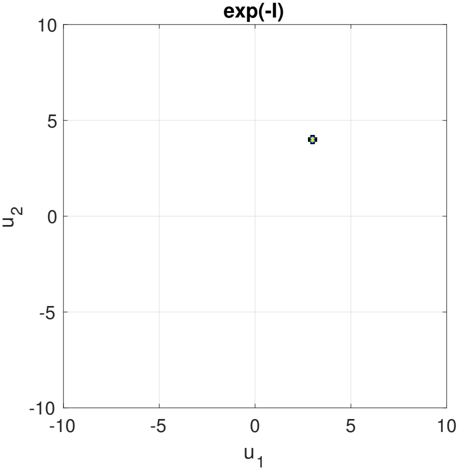
Contour plot of the posterior with observational noise $\lambda = 10^{-3}$.

Motivational Example



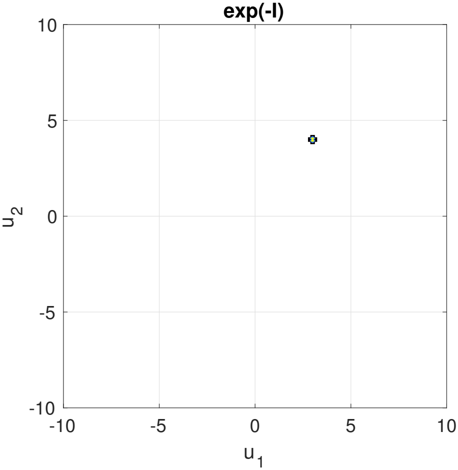
Contour plot of the posterior with observational noise $\lambda = 10^{-4}$.

Motivational Example



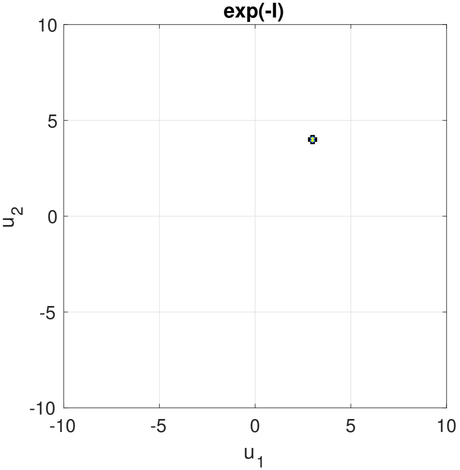
Contour plot of the posterior with observational noise $\lambda = 10^{-5}$.

Motivational Example



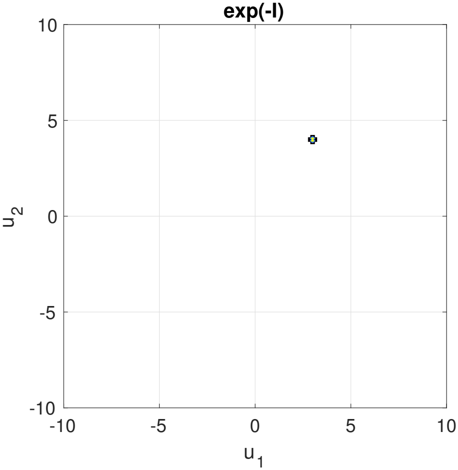
Contour plot of the posterior with observational noise $\lambda = 10^{-6}$.

Motivational Example



Contour plot of the posterior with observational noise $\lambda = 10^{-7}$.

Motivational Example



Contour plot of the posterior with observational noise $\lambda = 10^{-8}$.

Motivational Example

Laplace Approximation

$$\tilde{\mu}_n^y := \mathcal{N}\left(u_n, C_n\right) \,,$$

where u_n denotes the maximum a-posteriori estimate

$$u_n = \underset{u\in\mathbb{R}^2}{\operatorname{argmin}}\, I_n(u) = \underset{u\in\mathbb{R}^2}{\operatorname{argmin}}\, \frac{1}{2}\left((y-\mathcal{G}(u))^\top \Gamma_n^{-1}(y-\mathcal{G}(u))\right) + \frac{1}{2}\|u\|^2$$

and covariance

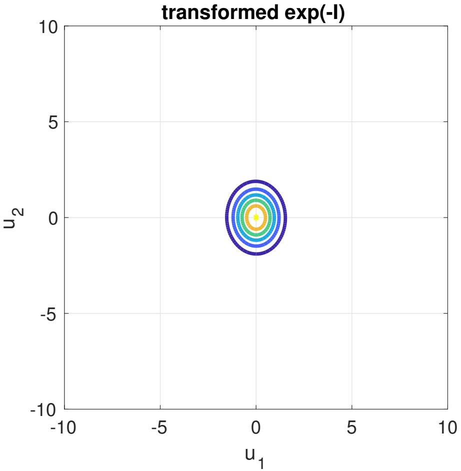
$$C_n^{-1} := \nabla^2 I_n(u_{MAP}).$$

Importance Sampling / Preconditioned Ratio Estimator

$$\mathbb{E}^{\mu^y}[\phi] = \int \phi \,\mathrm{d}\mu^y(u) = \int \phi \frac{1}{Z}\exp(-\Phi(u))\,\mathrm{d}\mu_0(u) = \int \phi\, \frac{\mathrm{d}\mu^y(u)}{\mathrm{d}\tilde{\mu}_n^y(u)}\mathrm{d}\tilde{\mu}_n^y(u)$$

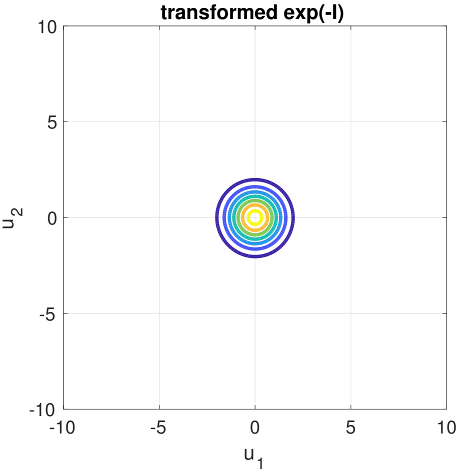
with Importance Sampling (MC, 10^5 samples), QMC (65000 samples), Gauss-Hermite quadrature (25 quadrature points).

Motivational Example



Contour plot of the Laplace approximation with $\lambda = 10^{-0}$.

Motivational Example



Contour plot of the Laplace approximation with $\lambda = 10^{-1,\dots,-8}$.

Motivational Example

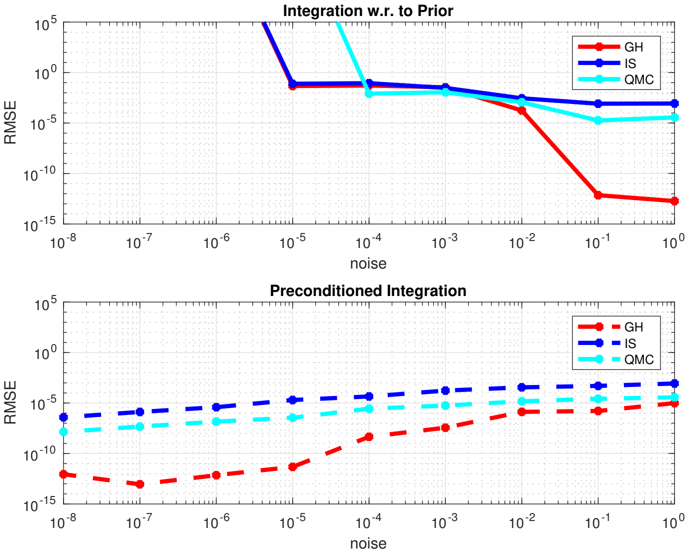


Figure: Convergence plots of $\mathbb{E}^{\mu^y}[\phi]$ w.r. to noise .

Motivational Example

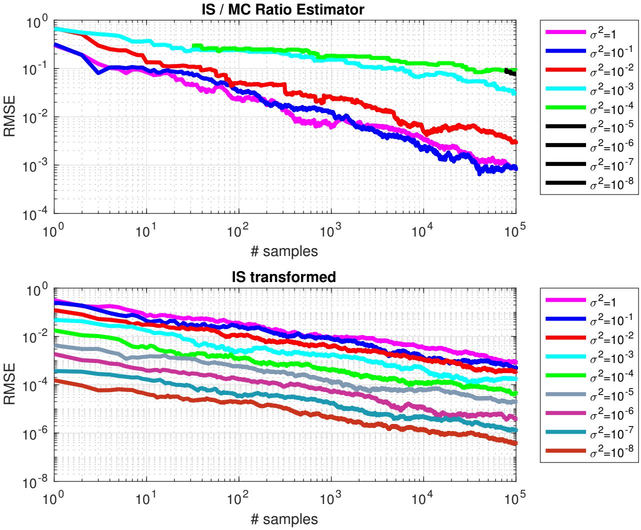


Figure: Convergence plots of $\mathbb{E}^{\mu^y}[\phi]$ for Importance Sampling .

Motivational Example

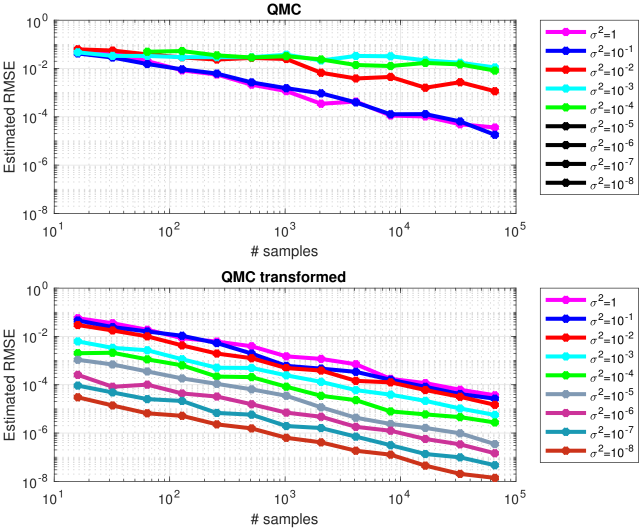


Figure: Convergence plots of $\mathbb{E}^{\mu^y}[\phi]$ for QMC.

Computation of the Laplace Approximation

Removing the degeneracy in the integrand function

- The maximizer u_{MAP} of the posterior measure is computed by minimizing the potential

$$\big((y-\mathcal{G}(u))^\top \Gamma_n^{-1}(y-\mathcal{G}(u))\big) - \log(\pi_0(u))$$

  using a trust-region Quasi-Newton approach with SR1 updates.

- Diagonalize the approximated Hessian $H_{SR1} = QMQ^\top$ and regularize the integrand by the curvature rescaling transformation

$$u_{MAP} + QM^{-1/2}u\,, \quad u \in \mathbb{R}^d\,.$$

Computation of the Laplace Approximation

- Quasi-Newton methods need in general only a few iterations (~ 20 iterations) to approximate the MAP, i.e. computational effort is negligible compared to sampling.
- Gradients can be efficiently computed via the adjoint approach,

$$\min_{u,p} I(u,p)$$
$$\text{s.t.} \quad g(u,p) = 0 \,.$$

The gradient is given by

$$\nabla_u I = g_u \tilde{p} + I_u$$

with

$$g_p^\top \tilde{p} = -I_u^\top \,.$$

- Automatic differentiation can be used in this context.

Asymptotic Analysis / Uniform Prior

Theorem

Assume that the parameter space is finite (possibly after dimension - truncation) and $\mathcal{G}(\cdot)$, y are such that the assumptions of Laplace's method hold; in particular, the minimum u_0 of

$$S(u) = \frac{1}{2}\Big((y - \mathcal{G}(u))^\top (y - \mathcal{G}(u))\Big)$$

is nondegenerate.

Then, as $\Gamma \downarrow 0$, the Bayesian estimate admits an asymptotic expansion

$$\mathbb{E}^{\mu^\delta}[\phi] = \frac{Z'_\Gamma}{Z_\Gamma} \sim a_0 + a_1\lambda + a_2\lambda^2 + \ldots.$$

where $a_0 = \phi(u_0)$.

Fedoryuk, Asimptotika: integraly i ryady Spravochnaya Matematicheskaya Biblioteka, 1987
CIS, Schwab, Scaling Limits in Computational Bayesian Inversion 2016

Preconditiong / Uniform Prior

- Results are the basis for preconditioned Smolyak quadrature methods for classes of operator equations with convergence rates independent of the number of parameters as well as of the observation noise variance.
- Generalization to preconditioned Quasi Monte Carlo methods is straightforward.

Motivation for new approximation results

- Generalization to other priors, in particular lognormal priors.
- Preconditioning strategies for MCMC, Importance Sampling.

Preconditiong / Uniform Prior

- Results are the basis for preconditioned Smolyak quadrature methods for classes of operator equations with convergence rates independent of the number of parameters as well as of the observation noise variance.
- Generalization to preconditioned Quasi Monte Carlo methods is straightforward.

Motivation for new approximation results

- Generalization to other priors, in particular lognormal priors.
- Preconditioning strategies for MCMC, Importance Sampling.

Approximation Results of Laplace Approximation

Setting

We consider a sequence of concentrating probability measures

$$\mu_n(\mathrm{d}u) = \frac{1}{Z_n} \exp\left(-n\Phi_n(u)\right) \mu_0(\mathrm{d}u), \quad Z_n := \int_{\mathbb{R}^d} \exp\left(-n\Phi_n(u)\right) \mu_0(\mathrm{d}u)$$

on $\mathbb{R}^d$, $n \in \mathbb{N}$ with $\mu_0(\mathrm{d}u) = \pi_0(u)\mathrm{d}u$, $\mathfrak{N} := \{u \in \mathbb{R}^d \colon \pi_0(u) > 0\}$.

The Laplace approximation of μ_n is given by the Gaussian measure

$$\tilde{\mu}_n := \mathcal{N}\left(u_n, \frac{1}{n} C_n\right)$$

with u_n given by

$$u_n := \operatorname*{argmin}_{u \in \mathfrak{N}} I_n(u) = \operatorname*{argmin}_{u \in \mathfrak{N}} \Phi_n(u) - \frac{1}{n} \log \pi_0(u). \qquad C_n^{-1} := \nabla^2 I_n(u_n).$$

Approximation Results of Laplace Approximation

Assumptions A

- $\Phi_n, \pi_0 \in C^3(\mathfrak{N}, \mathbb{R})$ and u_n is the unique global minimizer of I_n with

$$I_n(u_n) = 0, \qquad \nabla I_n(u_n) = 0, \qquad \nabla^2 I_n(u_n) > 0\,.$$

- There exist the limits

$$u_\star := \lim_{n\to\infty} u_n \qquad H_\star := \lim_{n\to\infty} \nabla^2 \Phi_n(u_n)$$

  in $\mathbb{R}^d$ and $\mathcal{L}(\mathbb{R}^d;\mathbb{R}^d)$ with $H_\star > 0$ and $u_\star$ belonging to the interior of $\mathfrak{N}$.

- For each ball $B_r(0)$, $r > 0$, there exists a constant $K_r < \infty$ and an $n_0 \in \mathbb{N}$ such that

$$\max_{u\in B_r(0)} \|\nabla^3 \log \pi_0(u)\| \leq K_r, \;\; \max_{u\in B_r(0)} \|\nabla^3 \Phi_n(u)\| \leq K_r \quad \forall n \geq n_0.$$

- For each $r > 0$ there exists a δ_r and an n_r such that

$$\delta_r \leq \inf_{u\notin B_r(u_n)} I_n(u) \qquad \forall n \geq n_r.$$

Approximation Results of Laplace Approximation

Theorem

Let the Assumptions A be satisfied.

Then, there holds

$$d_{\mathsf{H}}(\mu_n, \tilde{\mu}_n) \in \mathcal{O}(n^{-1/2}).$$

- Hellinger distance

$$d_{\mathsf{H}}^2(\mu, \tilde{\mu}) := \int_{\mathbb{R}^d} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(u)} - \sqrt{\frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\nu}(u)} \right|^2 \nu(\mathrm{d}u)\,.$$

- Key result for the proof:
  For $\pi_n(u) := \exp\left(-n\Phi_n(u)\right)\pi_0(u)$ and $\tilde{\pi}_n(u) := \exp\left(-\frac{n}{2}\|u - u_n\|^2_{C_n^{-1}}\right)$, for all $u \in \mathbb{R}^d$, there holds for any $p \in \mathbb{N}$

$$\int_{\mathbb{R}^d} \left| \left(\frac{\pi_n(u)}{\tilde{\pi}_n(u)}\right)^{1/p} - 1 \right|^p \tilde{\mu}_n(\mathrm{d}u) \in \mathcal{O}(n^{-p/2}).$$

Approximation Results of Laplace Approximation

Theorem

Let the Assumptions A be satisfied.

Then, there holds

$$d_{\mathsf{H}}(\mu_n, \tilde{\mu}_n) \in \mathcal{O}(n^{-1/2}).$$

- Hellinger distance

$$d_{\mathsf{H}}^2(\mu, \tilde{\mu}) := \int_{\mathbb{R}^d} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(u)} - \sqrt{\frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\nu}(u)} \right|^2 \nu(\mathrm{d}u)\,.$$

- Key result for the proof:
  For $\pi_n(u) := \exp\left(-n\Phi_n(u)\right)\pi_0(u)$ and $\tilde{\pi}_n(u) := \exp\left(-\frac{n}{2}\|u-u_n\|^2_{C_n^{-1}}\right)$, for all $u \in \mathbb{R}^d$, there holds for any $p \in \mathbb{N}$

$$\int_{\mathbb{R}^d} \left| \left(\frac{\pi_n(u)}{\tilde{\pi}_n(u)}\right)^{1/p} - 1 \right|^p \tilde{\mu}_n(\mathrm{d}u) \in \mathcal{O}(n^{-p/2}).$$

Approximation Results of Laplace Approximation

Theorem

Let the Assumptions A be satisfied.

Then, there holds

$$d_{\mathsf{H}}(\mu_n, \tilde{\mu}_n) \in \mathcal{O}(n^{-1/2}).$$

- **Generalization to the case of singular Hessian.**
- **No Gaussian prior assumption needed.**
- **Numerical discretization error can be incorporated.**

Laplace Approximation for Bayesian Inverse Problems

Consistency

Let μ_n^y denote the posterior of the sequence of problems

$$y = \mathcal{G}(u) + \eta$$

with $\eta \sim \mathcal{N}(0, 1/n\Gamma)$ satisfying Assumption A. Furthermore, we denote by $\tilde{\mu}_n^y$ the corresponding Laplace approximation.

For an estimator $\hat{Q}_n(\phi) \approx \mathbb{E}^{\mu_n^y}[\phi]$ with QoI $\phi : X \to \mathbb{R}$, the approximation error can be bounded by

$$|\hat{Q}_n(\phi) - \mathbb{E}^{\mu_n^y}[\phi]| \leq |\hat{Q}_n(\phi) - \mathbb{E}^{\tilde{\mu}_n^y}[\phi]| + \underbrace{|\mathbb{E}^{\tilde{\mu}_n^y}[\phi] - \mathbb{E}^{\mu_n^y}[\phi]|}_{\to 0 \quad (n \to \infty)}$$

- Importance Sampling: $|\hat{Q}(\phi) - \mathbb{E}^{\tilde{\mu}^y}[\phi]|$ converges assuming uniform boundedness of the variances.
- QMC / Sparse Quadrature: Faà di Bruno's formula gives bounds on the mixed derivatives.

Laplace Approximation for Bayesian Inverse Problems

Consistency

Let μ_n^y denote the posterior of the sequence of problems

$$y = \mathcal{G}(u) + \eta$$

with $\eta \sim \mathcal{N}(0, 1/n\Gamma)$ satisfying Assumption A. Furthermore, we denote by $\tilde{\mu}_n^y$ the corresponding Laplace approximation.

For an estimator $\hat{Q}_n(\phi) \approx \mathbb{E}^{\mu_n^y}[\phi]$ with QoI $\phi : X \to \mathbb{R}$, the approximation error can be bounded by

$$|\hat{Q}_n(\phi) - \mathbb{E}^{\mu_n^y}[\phi]| \leq |\hat{Q}_n(\phi) - \mathbb{E}^{\tilde{\mu}_n^y}[\phi]| + \underbrace{|\mathbb{E}^{\tilde{\mu}_n^y}[\phi] - \mathbb{E}^{\mu_n^y}[\phi]|}_{\to 0 \quad (n \to \infty)}$$

- Importance Sampling: $|\hat{Q}(\phi) - \mathbb{E}^{\tilde{\mu}^y}[\phi]|$ converges assuming uniform boundedness of the variances.
- QMC / Sparse Quadrature: Faà di Bruno's formula gives bounds on the mixed derivatives.

Summary and Outlook

- Convergence results for Laplace approximation in the Hellinger distance for vanishing noise.
- Efficient treatment of small observation noise variance.
- Development of preconditioning techniques to overcome the convergence problems.
- Combination of optimization and sampling techniques.

- Noise robustness of Laplace Importance Sampling and MCMC variants.

Björn Sprungk's poster on Variance-Independence of Random Walk Metropolis-Hastings Algorithms

References

- Fedoryuk M V 1987 Asimptotika: integraly i ryady Spravochnaya Matematicheskaya Biblioteka.
- Lu Y, Stuart A M and Weber H 2017 Gaussian Approximations for Probability Measures on R^d *SIAM/ASA J. Uncertainty Quantification* **5** 1136-1165.
- Scheichl R, Stuart A M, Teckentrup A L 2017 Quasi-Monte Carlo and Multilevel Monte Carlo Methods for Computing Posterior Expectations in Elliptic Inverse Problems *SIAM/ASA J. on Uncertainty Quantification* **5(1)** 493?518.
- Schillings C, Sprungk B and Wacker P 2018 On the Convergence of Laplace's Approximation and Its Implications for Bayesian Computation (in preparation).
- Schillings C and Schwab C 2014 Sparsity in Bayesian inversion of parametric operator equations *Inverse Problems* **30** 065007.
- Schillings C and Schwab C 2016 Scaling Limits in Computational Bayesian Inversion *ESAIM:M2AN* **50** 6.
- Sprungk B 2018 On Variance-Independence of Random Walk Metropolis-Hastings Algorithms (in preparation).