ENHANCING THE RELIABILITY OF GENERAL-PURPOSE ALGORITHMS FOR APPROXIMATE BAYESIAN INFERENCE

Yu Wang

November 20, 2024

Bayesian Approximation

$$\pi\left(\theta \mid \{Y_i\}_{i=1}^N\right) = \frac{\prod_{i=1}^N p(Y_i \mid \theta)\pi_0(\theta)}{Z}$$

- We want to learn about π, typically by calculating expectations

$$\mathbb{E}\{f(X)\} := \pi(f) := \int f(x)\pi(\mathrm{d}x).$$

Bayesian Approximation

$$\pi\left(\theta \mid \{Y_i\}_{i=1}^N\right) = \frac{\prod_{i=1}^N p(Y_i \mid \theta)\pi_0(\theta)}{Z}$$

- We want to learn about π, typically by calculating expectations

$$\mathbb{E}\{f(X)\} := \pi(f) := \int f(x)\pi(\mathrm{d}x).$$

- However, in general, the expectations can't be done exactly.

Bayesian Approximation

$$\pi\left(\theta \mid \{Y_i\}_{i=1}^N\right) = \frac{\prod_{i=1}^N p(Y_i \mid \theta)\pi_0(\theta)}{Z}$$

- We want to learn about π, typically by calculating expectations

$$\mathbb{E}\{f(X)\} := \pi(f) := \int f(x)\pi(\mathrm{d}x).$$

- However, in general, the expectations can't be done exactly.
- Approximate inference: Markov chain Monte Carlo (MCMC) constructs a Markov chain $Y_0, Y_1, Y_2, \ldots$, such that $\rho_n \to \pi$

$$\sum_{n=1}^N f(Y_n)/N \overset{a.s.}{\to} \pi(f) \quad \text{for } N \to \infty.$$

Challenges in Modern Approximate Bayesian Inference

- **Challenges:**
  - high-dimensional $\theta \in \mathbb{R}^D$, D is large.
  - complex relationship $p(Y_i|\theta)$.
  - large-sclae dataset $\{Y_i\}_{i=1}^N$.

Challenges in Modern Approximate Bayesian Inference

- **Challenges:**
  - high-dimensional $\theta \in \mathbb{R}^D$, D is large.
  - complex relationship $p(Y_i|\theta)$.
  - large-sclae dataset $\{Y_i\}_{i=1}^N$.
- **Problems of MCMC:**
  - Slow convergence and mixing in high-dimensional spaces.
  - Computational cost of likelihood evaluation is proportional to the dataset size.
  - Diagnosing convergence is harder in high dimensions.

Challenges in Modern Approximate Bayesian Inference

- **Challenges:**
  - high-dimensional $\theta \in \mathbb{R}^D$, D is large.
  - complex relationship $p(Y_i|\theta)$.
  - large-sclae dataset $\{Y_i\}_{i=1}^N$.
- **Problems of MCMC:**
  - Slow convergence and mixing in high-dimensional spaces.
  - Computational cost of likelihood evaluation is proportional to the dataset size.
  - Diagnosing convergence is harder in high dimensions.
- **Mitigations:**
  - Variational Inference (VI).
  - Subsampling methods (e.g. Stochastic Gradient Langevin Dynamics (SGLD)).

Overview

- A priori finite-time, finite-data guarantees:
  A Unifying Framework for Understanding General-purpose Bayesian Posterior Approximation Methods,
  Huggins, Kasprzak, **Wang**, Campbell, Broderick. In prep

Overview

- A priori finite-time, finite-data guarantees:
  A Unifying Framework for Understanding General-purpose Bayesian Posterior Approximation Methods,
  Huggins, Kasprzak, **Wang**, Campbell, Broderick. In prep
- **A post hoc quality check for VI**:
  A Targeted Accuracy Diagnostic for Variational Approximations (TADDAA), **Wang**, Kasprzak, Huggins. AISTATS 2023.
- **Uncertainty quantification for Subsampling methods**:
  Stationary Analysis of Fixed Learning Rate Stochastic Gradient Algorithms, **Wang** & Huggins. (Under Review)

TADDAA

Markov chain Monte Carlo (MCMC)

MCMC sampling methods provide a general-purpose framework for obtaining samples that are asymptotically exact.

- **Proposal distribution:** $Q_\psi(x, \mathrm{d}y)$ parameterized by ψ with current state x and corresponding density $q_\psi(x, y)$.
- **Metropolis–Hastings (MH) correction:** to construct a Markov kernel with the desired stationary distribution π, a proposed state $Y \sim Q_\psi(x, \cdot)$ is accepted with probability

$$\alpha(x, Y) = \min\left\{1, \frac{\pi(Y) q_\psi(Y, x)}{\pi(x) q_\psi(x, Y)}\right\}.$$

Variational Inference (VI)

Variational inference (VI) provides a potentially faster alternative to MCMC when models are complex and/or the dataset size is large.

$$\hat{\pi} = \arg\min_{\xi \in \mathcal{Q}} \mathcal{D}_{\pi}(\xi).$$

- Variational family $\mathcal{Q}$: we are able to efficiently calculate expectations of interest (e.g. mean and variance).
- Measure of discrepancy $\mathcal{D}_{\nu}(\cdot)$: the canonical choice is *Kullback–Leibler (KL) divergence* out of convenience.

$$\mathcal{D}_{\pi}(\xi) = \mathrm{KL}(\xi \mid \pi) := \int \log\left(\frac{\mathrm{d}\xi}{\mathrm{d}\pi}\right) \mathrm{d}\xi.$$

Related Works

- Existing evaluation tools:
  - Evidence Lower Bound (ELBO).
  - Kernel Stein Discrepancy (KSD).
  - Pareto smoothed importance sampling (PSIS) $\hat{k}$.
- Problems:
  - Lack interpretability.
  - Not applicable in high-dimensional parameter spaces.
  - Don't support marginal checks.

TADDAA:Intuition

We want to quantify approximation error $\varepsilon^{(0)}(\mathcal{F}) := \mathcal{F}(\hat{\pi}^{(0)}) - \mathcal{F}(\pi)$ for a posterior functional of interest $\mathcal{F}$ such as a mean or variance.

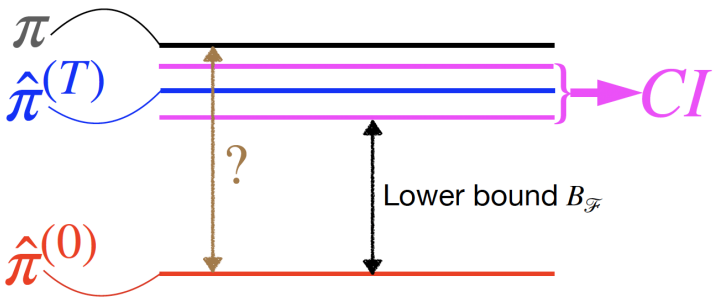
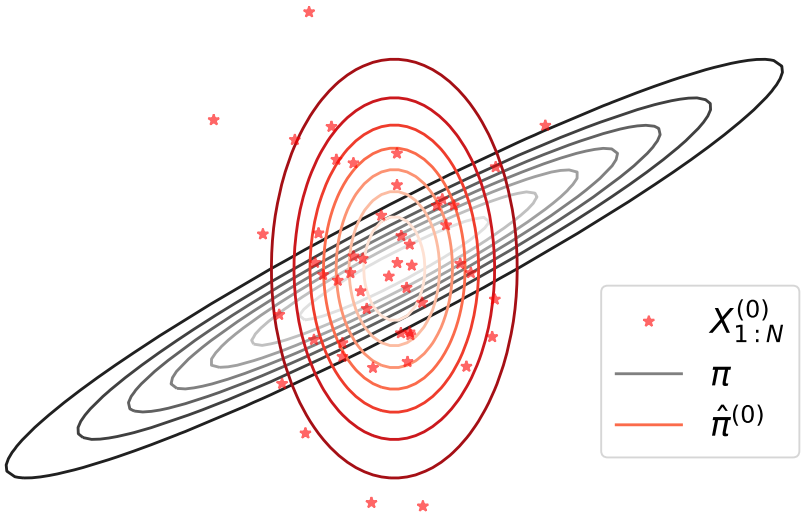


Figure 1: $\hat{\pi}^{(T)}$ significantly different from $\hat{\pi}^{(0)} \Rightarrow \hat{\pi}^{(0)}$ far from π.

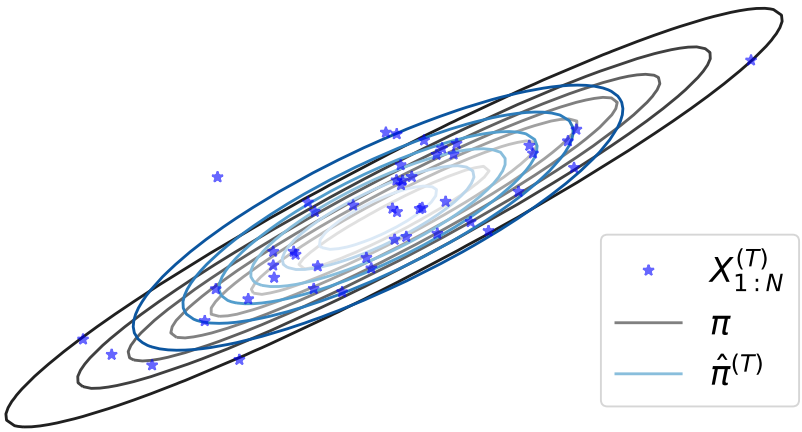
- For another approximation $\hat{\pi}^{(T)}$ closer to taget posteriror π, $\varepsilon^{(T)}(\mathcal{F}) := \mathcal{F}(\hat{\pi}^{(T)}) - \mathcal{F}(\pi)$
- $$\varepsilon^{(0)}(\mathcal{F}) \geq |\mathcal{F}(\hat{\pi}^{(0)}) - \mathcal{F}(\hat{\pi}^{(T)})|$$
$$\geq \begin{cases} 0 & \text{if } \ell_{\mathcal{F}} \leq 0 \leq u_{\mathcal{F}} \\ \min(|\ell_{\mathcal{F}}|, |u_{\mathcal{F}}|) & \text{otherwise} \end{cases}$$
$$= \mathbf{1}\{0 \notin (\ell_{\mathcal{F}}, u_{\mathcal{F}})\} \times \min(|\ell_{\mathcal{F}}|, |u_{\mathcal{F}}|)$$
$$=: B_{\mathcal{F}}.$$

TADDAA:Input



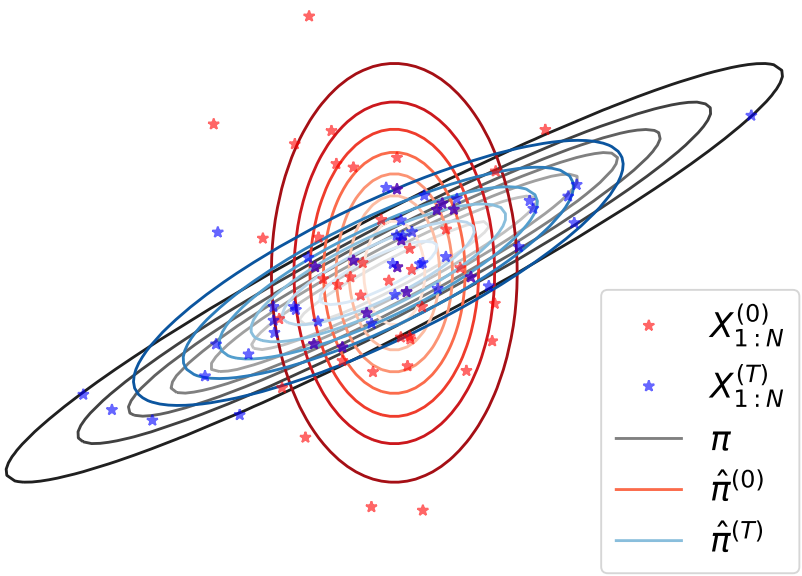
- log density of the target π
- approximating distribution $\hat{\pi}^{(0)}$
- functional of interest $\mathcal{F}$ (e.g. marginal mean)
- transition kernel $K_h(x, \mathrm{d}y)$ (e.g. Barker, HMC)
- number of Markov chains N and iterations T

TADDAA:Run MCMC with inter-chain adaptation (INCA)



for $t = 0$ to $T - 1$ **do**:
  for $j = 1$ to N **do**:
    $X_j^{(t+1)} \sim K_{h^{(t)}}(X_j^{(t)}, \cdot)$
  end for
  update step-size $h^{(t+1)}$ using INCA.
end for

TADDAA:Compute error lower bounds and reliability check



- Compute correlation check $\rho^2_{\max}(T)$
- Compute a confidence interval $(\ell_{\mathcal{F}}, u_{\mathcal{F}})$ for $\mathcal{F}(\hat{\pi}^{(0)}) - \mathcal{F}(\hat{\pi}^{(T)})$ based on $X^{(0)}_{1:N}$ and $X^{(T)}_{1:N}$
- Compute lower bound $B_{\mathcal{F}}$

Transition kernel $K_h(x, \mathrm{d}y)$

Let $x \in \mathbb{R}^d$ denote the current state, $h \in \mathbb{R}_+$ the step size, and $G \in \mathbb{R}^{d \times d}$ a positive semi-definition preconditioning matrix.

- Random Walk Metropolis-Hasting (RWMH).
- Metropolis-adjusted Langevin algorithm (MALA).
- Hamiltonian Monte Carlo (HMC).
- **Barker Proposal (recommended choice)**: robust to precise step size and acceptance rate.

Step size h

All four kernels rely on a step-size parameter $h^{(t)}$, to guarantee superior sampling efficiency when dimension d is large, $h^{(t)}$ should be adapted according to *inter-chain adaptation* (INCA) and *optimal scaling*.

- Step size adaption:
  - Generate proposals $Y_j^{(t+1)} \sim Q_{\psi^{(t)}}(X_j^{(t)}, \cdot)$, then accept with probability $\alpha_j^{(t)}$.
  - $$\psi^{(t+1)} = \psi^{(t)} + \frac{1}{\sqrt{t+1}}(\bar{\alpha}^{(t)} - \bar{\alpha}_*),$$

    where $\psi^{(t)} = \log h^{(t)}$ and $\bar{\alpha}_*$ is optimal asymptotic acceptance.

Step size h

All four kernels rely on a step-size parameter $h^{(t)}$, to guarantee superior sampling efficiency when dimension d is large, $h^{(t)}$ should be adapted according to *inter-chain adaptation* (INCA) and *optimal scaling*.

- Step size adaption:
  - Generate proposals $Y_j^{(t+1)} \sim Q_{\psi^{(t)}}(X_j^{(t)}, \cdot)$, then accept with probability $\alpha_j^{(t)}$.
  - $$\psi^{(t+1)} = \psi^{(t)} + \frac{1}{\sqrt{t+1}}(\bar{\alpha}^{(t)} - \bar{\alpha}_*),$$

    where $\psi^{(t)} = \log h^{(t)}$ and $\bar{\alpha}_*$ is optimal asymptotic acceptance.
- Optimal initial step size $h^{(0)}$ and $\bar{\alpha}_*$:
  - RWMH: $h^{(0)} = 2.4^2/d$, $\bar{\alpha}_* = 0.234$.
  - MALA and Barker: $h^{(0)} = 2.4^2/d^{1/3}$, $\bar{\alpha}_* = 0.576$.
  - HMC: $h^{(0)} = 2.4^2/d^{1/4}$, $\bar{\alpha}_* = 0.4$.

Step size h

All four kernels rely on a step-size parameter $h^{(t)}$, to guarantee superior sampling efficiency when dimension d is large, $h^{(t)}$ should be adapted according to *inter-chain adaptation* (INCA) and *optimal scaling*.

- Step size adaption:
  - Generate proposals $Y_j^{(t+1)} \sim Q_{\psi^{(t)}}(X_j^{(t)}, \cdot)$, then accept with probability $\alpha_j^{(t)}$.
  - $$\psi^{(t+1)} = \psi^{(t)} + \frac{1}{\sqrt{t+1}}(\bar{\alpha}^{(t)} - \bar{\alpha}_*),$$

    where $\psi^{(t)} = \log h^{(t)}$ and $\bar{\alpha}_*$ is optimal asymptotic acceptance.
- Optimal initial step size $h^{(0)}$ and $\bar{\alpha}_*$:
  - RWMH: $h^{(0)} = 2.4^2/d$, $\bar{\alpha}_* = 0.234$.
  - MALA and Barker: $h^{(0)} = 2.4^2/d^{1/3}$, $\bar{\alpha}_* = 0.576$.
  - HMC: $h^{(0)} = 2.4^2/d^{1/4}$, $\bar{\alpha}_* = 0.4$.
- Problem: INCA introduces dependence between the Markov chains, which could invalidate the statistical tests and confidence intervals.

Number of Markov chains N

The number of Markov chains must be sufficiently large that the confidence intervals are small enough to detect meaningful errors. Hence, if the user's tolerance is δ_{mean} for relative mean error and δ_{var} for log variance error, then

$$N = \max\left(N_{\text{mean}}, N_{\text{variance}}\right),$$

where

$$N_{\text{mean}} := \min\left\{n \in \mathbb{N} : \frac{t_{n-1}(\alpha/2)}{\sqrt{n}} \leq \delta_{\text{mean}}\right\},$$

$$N_{\text{variance}} := \min\left\{n \in \mathbb{N} : \log\left(\frac{\chi^2_{n-1}(1-\alpha/2)}{\chi^2_{n-1}(\alpha/2)}\right) \leq \delta_{\text{var}}\right\}.$$

Number of iterations T

Markov chain requires $\Theta(d^\gamma)$ iterations to mix according to theory of optimal scaling.

- For RWMH, MALA, Barker: $T = \lfloor cd^{1/3} \rfloor$.
- For HMC: $T = \lfloor cd^{1/4}/L \rfloor$, where L is the number of leapfrog steps in HMC.

[1] Bhatia, Kush, et al. Statistical and computational trade-offs in variational inference: A case study in inferential model selection.

Number of iterations T

Markov chain requires $\Theta(d^\gamma)$ iterations to mix according to theory of optimal scaling.

- For RWMH, MALA, Barker: $T = \lfloor cd^{1/3} \rfloor$.
- For HMC: $T = \lfloor cd^{1/4}/L \rfloor$, where L is the number of leapfrog steps in HMC.

Remark

- Based on our ablation studies, $c = 50$ is a reasonable choice.
- Theories [1] suggest computational cost of TADDAA is comparable to VI:
  - Computational cost for VI: $\Theta(d^{1/3})$.
  - Computational cost for MALA and Barker: $\Theta(d^{1/3})$.
  - Computational cost for HMC: $\Theta(d^{1/4})$.

[1] Bhatia, Kush, et al. Statistical and computational trade-offs in variational inference: A case study in inferential model selection.

A Reliability Check for the Diagnostic

The reliability of TADDAA depends on the mixing behavior of the Markov chains:

- If the Markov chains are mixing well, we expect the diagnostic results can be trusted.
- If the Markov chains are not mixing well, diagnosis of a poor approximation can still be trusted but diagnosis of a good approximation may not be reliable. [2]

[2] In the latter case, we should consider increasing the length of Markov chains or otherwise improving the Markov kernel.

A Reliability Check for the Diagnostic (Continued)

- We propose to use the worst-case correlation coefficient $\rho^2_{\max}(T) := \max_i \rho^2_i(T)$, where

$$\rho^2_i(T) := \frac{\sum_{j=1}^{N}(X^{(0)}_{j,i} - \hat{\mu}^{(0)}_i)(X^{(T)}_{j,i} - \hat{\mu}^{(T)}_i)}{\sqrt{\sum_{j=1}^{N}(X^{(0)}_{j,i} - \hat{\mu}^{(0)}_i)^2}\sqrt{(\sum_{j=1}^{N} X^{(T)}_{j,i} - \hat{\mu}^{(T)}_i)^2}}.$$

- Check passes: $\rho^2_{\max}(T) < 0.1$.

Asymptotic Independence of Adapted Markov Chains

The Markov chains $X_{1:N}^{(t)}$ are not independent once $t > 1$, so the final samples $X_{1:N}^{(T)}$ violate the independence requirement of statistical tests.

Asymptotic Independence of Adapted Markov Chains

The Markov chains $X_{1:N}^{(t)}$ are not independent once $t > 1$, so the final samples $X_{1:N}^{(T)}$ violate the independence requirement of statistical tests.

Definition
Let $X_{N,1:N} = (X_{N,1}, \ldots, X_{N,N})$ denote a random vector. The sequence of random vectors $\{X_{N,1:N}\}_{N=1}^{\infty}$ is *$\bar{\nu}$-chaotic* if, for any $r \in \mathbb{N}$ and any bounded continuous real-valued functions $g_1, g_2, \ldots, g_r$,

$$\lim_{N\to\infty} \mathbb{E}_{X_{N,1:N}} \left\{ \prod_{i=1}^{r} g_i \left(X_{N,i} \right) \right\} = \prod_{i=1}^{r} \int g_i(x) \bar{\nu}(\mathrm{d}x).$$

Adapted Markov Chains are Chaotic

Assumption

1. *The proposal probability density* $q_h(y, x)$ *is continuous with respect to* (x, y, h).
2. *The target distribution has a continuous probability density function* $\pi(\cdot)$.
3. *Samples generated from the Markov transition kernel* $T(x, y, h, \cdot, \cdot)$ *satisfy* $\mathbb{E}\|X_j^{(t)}\|^2 < \infty$ *and* $\mathbb{E}\|Y_j^{(t)}\|^2 < \infty$ *for any* $t \in \mathbb{N}$.

Theorem

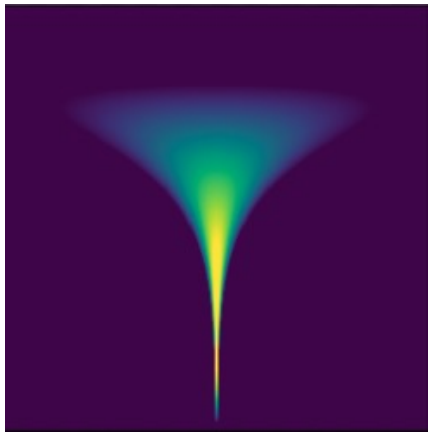
Under some mild assumptions, for any $t \in \mathbb{N}$, *there exists a probability distribution* $\bar{\nu}^{(t)}$ *such that the sequence* $\{X_{1:N}^{(t)}\}_{N=1}^{\infty}$ *is* $\bar{\nu}^{(t)}$*-chaotic.*

Experiment: Neal-Funnel Shape Model

Neal's example has support for $\log(\sigma) \in \mathbb{R}$ and $x \in \mathbb{R}^{d-1}$. The parameterization of this model is given as follows:

$$\log(\sigma) \sim \mathcal{N}(0, \sigma_0^2), \quad x_i \sim \mathcal{N}(0, \sigma).$$

For illustrative purpose, let $\beta_1 = \log(\sigma)$ and $\beta_2 = x_1$. In our experiment, $\sigma_0 = 1$.



Experiment: Neal-Funnel Shape Model

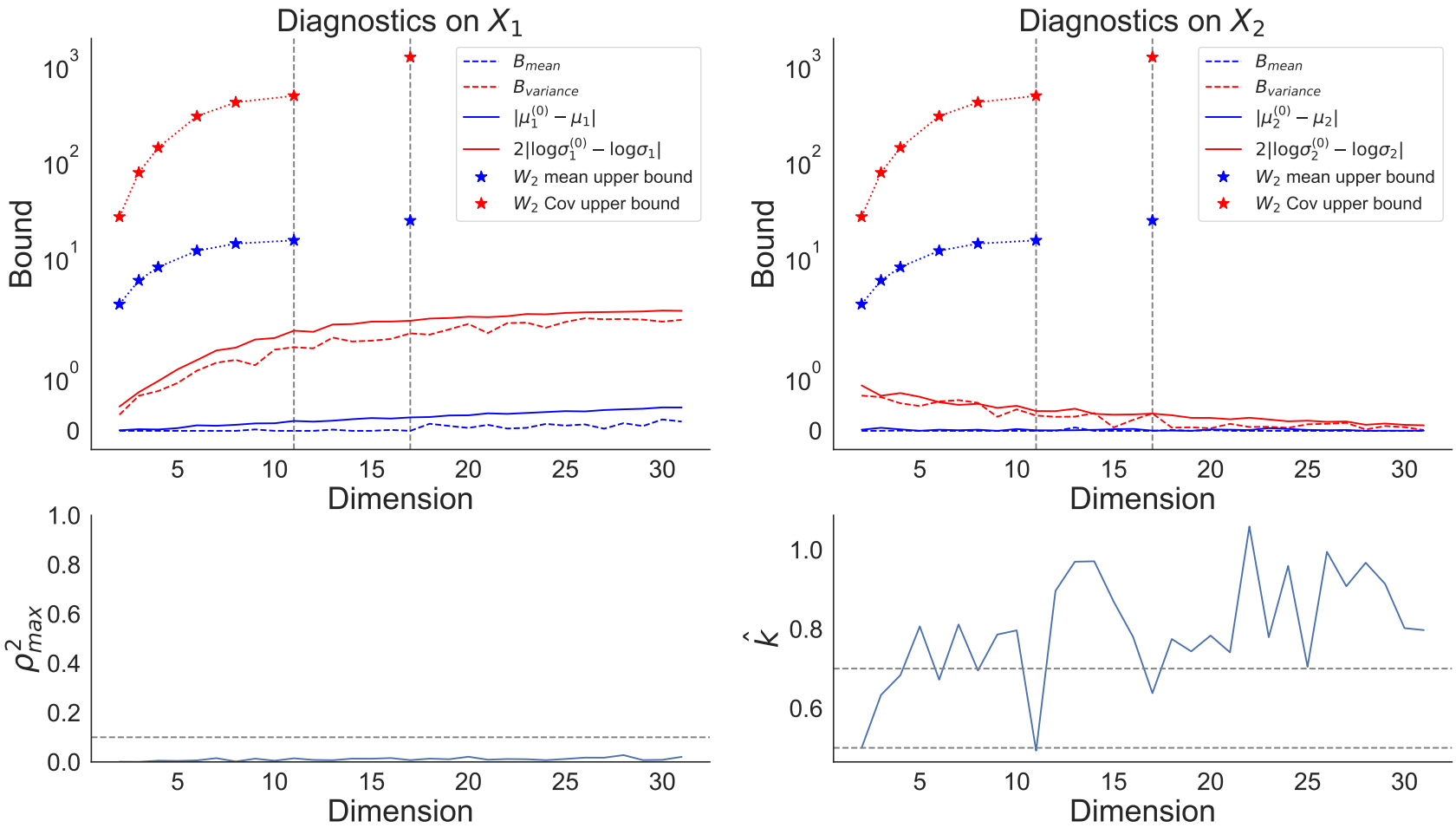
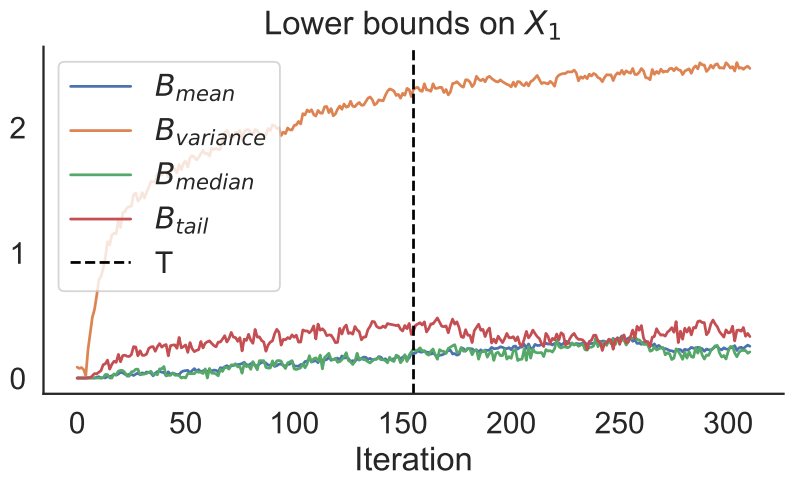
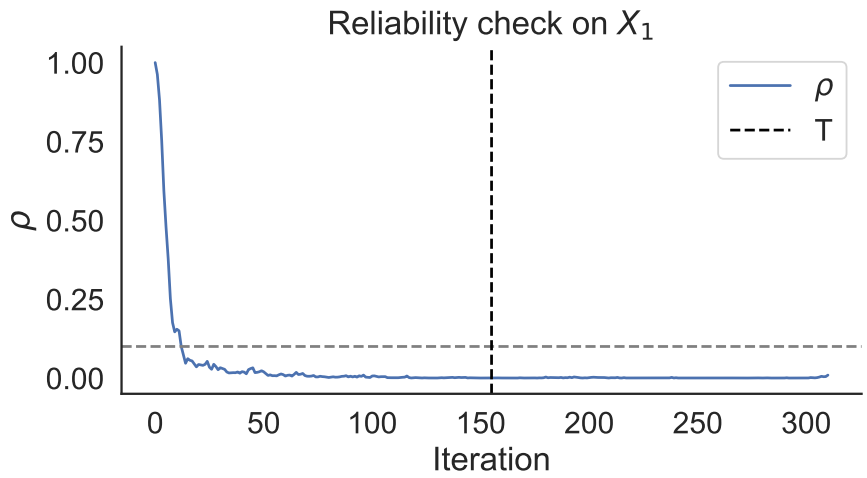


Figure 2: Diagnostics for Neal-funnel shape model, where TADDAA uses the Barker proposal. Here μ_i and σ_i^2 denote, respectively, the mean and variance of X_i.

Experiment: Neal-Funnel Shape Model

Ablation study on $d = 30$: the lower bounds become nearly constant at our proposed number of iterations T.





Experiment: Logistic Regression Using Horseshoe Prior

We use a logistic regression model with a sparsity-inducing horseshoe prior on

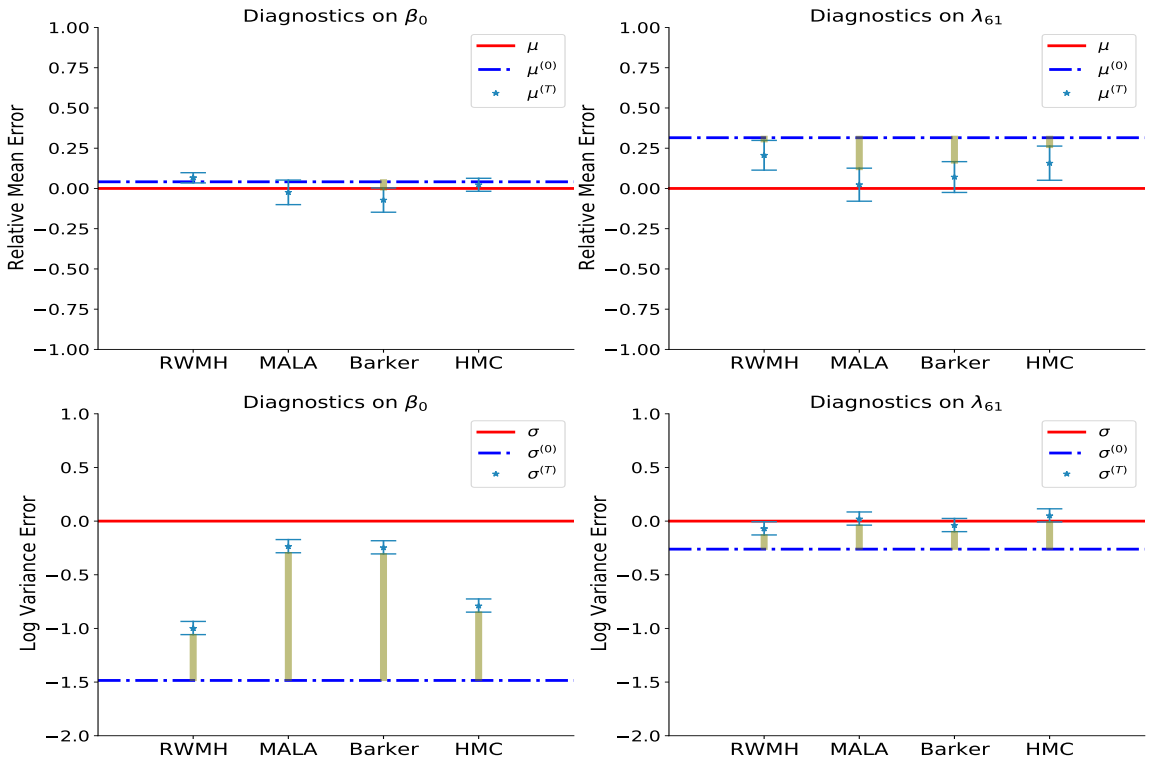
$$
\begin{aligned}
y \mid \beta &\sim \text{Bern}(\text{logit}^{-1}(X\beta)),\\
\beta_j \mid \tau, \lambda, c &\sim \mathcal{N}(0, \tau^2 \tilde{\lambda}_j^2),\\
\lambda_j &\sim \text{C}^+(0,1), \qquad \tau \sim \text{C}^+(0, \tau_0),\\
c^2 &\sim \text{InvGam}(2,8),
\end{aligned}
$$

where y denotes the binary outcomes, $\tau > 0$ and $\lambda > 0$ are global and local shrinkage parameters.

- $X \in \mathbb{R}^{71\times 100}$ is the features matrix.
- Parameter dimensionality is $d = 203$.

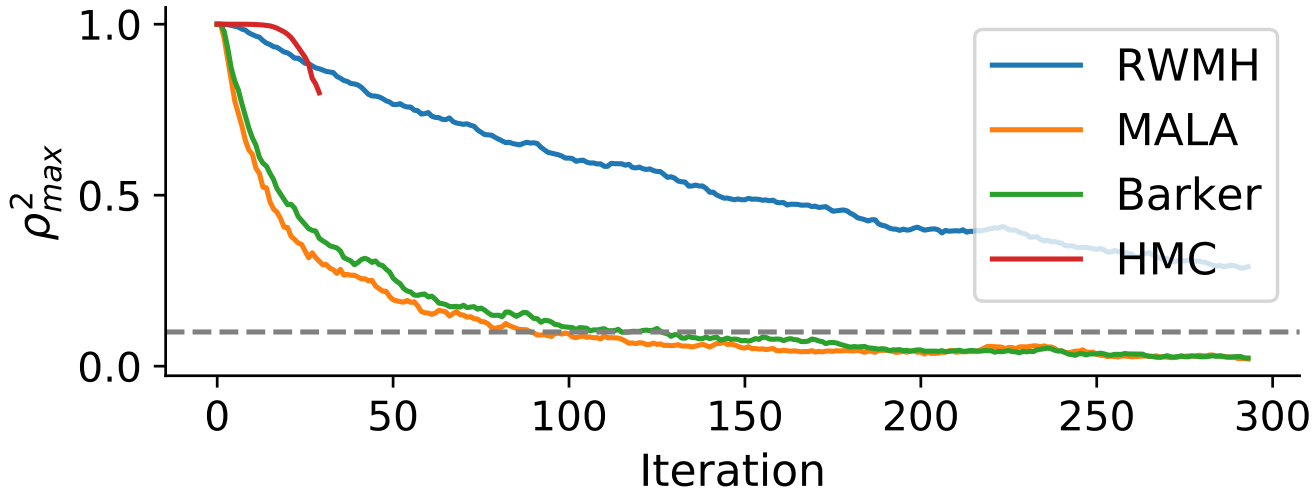
Experiment: Logistic Regression Using Horseshoe Prior

- Mean and variance confidence intervals: capture both accurate and inaccurate marginal estimates, provide quite precise lower bounds.
- Computational efficiency: use 28% as many gradient evaluations as VI.



Experiment: Logistic Regression Using Horseshoe Prior

Reliability check: Barker and MALA pass reliability check, RWMH and HMC chains fail to mix.



Stationary Analysis of Fixed Learning Rate Stochastic Gradient Algorithms

Stochastic optimization

Consider data $\{x_n\}_{n=1}^N$ with $x_n \in \mathbb{X}$. For a parameter $\theta \in \mathbb{R}^D$, observation-level differentiable loss $\ell : \mathbb{X} \times \mathbb{R}^D \to \mathbb{R}$, and regularizer $\mathcal{R} : \mathbb{R}^D \to \mathbb{R}$, we aim to minimize the loss function

$$\mathcal{L}(\theta) := N^{-1} \sum_{n=1}^N \ell(x_n, \theta) + N^{-1}\mathcal{R}(\theta).$$

- Gradient Descnet (GD):

$$\theta_t = \theta_{t-1} - \Lambda \nabla \mathcal{L}(\theta_{t-1})$$

- Stochastic Gradient Descnet (SGD):

$$\theta_t = \theta_{t-1} - \Lambda G_t(\theta_{t-1}),$$

where $G_t(\theta) := B^{-1} \sum_{n \in S_t} \nabla \ell(x_n, \theta) + N^{-1} \nabla \mathcal{R}(\theta)$ is the stochastic gradient.

Subsampling Markov chain Monte Carlo (SGLD)

SGLD is a Markov chain Monte Carlo (MCMC) algorithm equivalent to modifying SGD to include an additional Gaussian noise term

$$\theta_t = \theta_{t-1} - \Lambda\, G_t(\theta_{t-1}) + \sqrt{2\beta^{-1}\Lambda}\,\xi_{t-1},$$

- $\beta \in (0, \infty]$ is the inverse temperature (canonically set to $\beta = N$).
- $\xi_{t-1} \overset{\text{iid}}{\sim} \mathcal{N}(0, I)$.

Goal

We would like to accurately estimate the stationary covariance structure

$$\Sigma_\theta := \lim_{t \to \infty} \mathrm{Cov}(\theta_t).$$

Characterizing the stationary covariance can help quantify

- Test loss.
- Escaping efficiency from a sharp minimal.

Related Work: Quadratic Loss

Current works assume that the loss is well-approximated by a quadratic function:

$$\mathcal{L}(\theta_t) \approx \tilde{\mathcal{L}}(\theta_t) := \tfrac{1}{2}\big(\theta_t - \widehat{\theta}^{(N)}\big)^\top \widehat{H}\big(\theta_t - \widehat{\theta}^{(N)}\big) + \text{const},$$

where $\widehat{H} := \nabla^2 \mathcal{L}(\widehat{\theta})$ is the Hessian of the loss (evaluated at $\widehat{\theta}$).

Related Work: Continuous-time Proxies

The most popular proxy approach is to replace discrete dynamics of the iterative algorithm by a continuous-time the Ornstein–Uhlenbeck (OU) process

$$\mathrm{d}\vartheta_t = -\Lambda \widehat{H} \vartheta_t \mathrm{t} + \Lambda \widehat{C}^{1/2} \mathrm{d}W_t,$$

where W_t be a d-dimensional Brownian motion and $\widehat{C} = \mathrm{Cov}(G_1(\widehat{\theta}))$ denotes the gradient noise covariance at the minimizer.

- The covariance matrix of the stationary distribution $\Sigma_\vartheta := \mathrm{Cov}(\pi_\vartheta)$ satisfies
$$\Sigma_\vartheta \widehat{H} + \widehat{H} \Sigma_\vartheta = \Lambda \widehat{C}.$$
- Limitation: continuous-time proxies provide close approximation to SGD only for **small learning rates**.

Related Work: Discrete-time proxies

Assuming the loss is well-approximated by quadratic loss, the discrete-time proxy algorithm updates

$$\psi_t = \psi_{t-1} - \frac{\Lambda}{B} \sum_{n \in S_t} \widehat{H}_n (\psi_{t-1} - \widehat{\theta}),$$

where $\widehat{H}_n := \nabla^2 \ell(x_n, \widehat{\theta})$.

- *Implicit* characterization of Σ_ψ:

$$\Lambda \widehat{H} \Sigma_\psi + \Sigma_\psi \widehat{H} \Lambda = \Lambda \left(\overline{C}_\psi + \widehat{H} \Sigma_\psi \widehat{H} \right) \Lambda,$$

  where $\Sigma_\psi := \mathrm{Cov}(\pi_\psi)$, and $\overline{C}_\psi := \mathbb{E}[\mathrm{Cov}\{G_1(\psi_\infty)\}]$ denotes the expected covariance of the gradient noise.

- For well-specified linear model and assume $X \sim \mathcal{N}(0, A)$:

$$\overline{C}_\psi \approx B^{-1} \left(A \Sigma_\psi A + \mathrm{Tr}\left[A \Sigma_\psi \right] A + \sigma^2 A \right).$$

Limitations of discrete-time proxies

- Assumptions often do not hold in practice:
  - Sample size $N >> D$.
  - Mean Squared Error (MSE) loss.
  - The model is well-specified.
- There is no guarantee that the proxy process $(\psi_t)_{t\geq 0}$ is close to the original process $(\theta_t)_{t\geq 0}$.

A New Proxy Algorithm for Analyzing SG(L)D

Our approach is to apply a second-order Taylor approximation to each loss term $\ell_n(\theta) := \ell(x_n, \theta)$:

$$\tilde{\ell}_n(\theta) := \ell_n(\widehat{\theta}) + \nabla \ell_n^\top(\widehat{\theta})(\theta - \widehat{\theta}) + (\theta - \widehat{\theta})^\top \nabla^2 \ell_n(\widehat{\theta})(\theta - \widehat{\theta}),$$

A New Proxy Algorithm for Analyzing SG(L)D

Our approach is to apply a second-order Taylor approximation to each loss term $\ell_n(\theta) := \ell(x_n, \theta)$:

$$\tilde{\ell}_n(\theta) := \ell_n(\widehat{\theta}) + \nabla \ell_n^\top(\widehat{\theta})(\theta - \widehat{\theta}) + (\theta - \widehat{\theta})^\top \nabla^2 \ell_n(\widehat{\theta})(\theta - \widehat{\theta}),$$

- Minimizer $\widehat{\theta}$ satisfies

$$\nabla \mathcal{L}(\widehat{\theta}) = \tfrac{1}{N} N^{-1} \textstyle\sum_{n=1}^{N} \nabla \ell(x_n, \theta) + N^{-1} \nabla \mathcal{R}(\theta) = 0$$

- In general,

$$B^{-1} \sum_{n \in S_t} \nabla \ell(x_n, \widehat{\theta}) + N^{-1} \nabla \mathcal{R}(\widehat{\theta}) \neq 0.$$

Stationary Fluctuation

Our new proxy algorithm update as follows:

$$\begin{aligned}\psi_t = \psi_{t-1} - \frac{\Lambda}{B}\sum_{n\in S_t}\Big\{\nabla \ell_n\big(\widehat{\theta}\big) + \mathcal{J}_n\big(\psi_{t-1} - \widehat{\theta}\big)\Big\} \\ - \frac{\Lambda}{N}\nabla\mathcal{R}(\psi_{t-1}) + \sqrt{2\beta^{-1}\Lambda}\,\xi_{t-1}.\end{aligned} \tag{1}$$

Proposition

Assuming the iterates $(\psi_t)_{t\geq 0}$ have a well-defined stationary distribution, the stationary covariance Σ_ψ satisfies

$$\Lambda\widehat{H}\Sigma_\psi + \Sigma_\psi\widehat{H}\Lambda = \Lambda\big(\overline{C}_\psi + \widehat{H}\Sigma_\psi\widehat{H}\big)\Lambda + 2\beta^{-1}\Lambda.$$

Stationary Gradient Noise

Theorem

For the proxy algorithm, if $\mathcal{R}(\theta) = \frac{1}{2}\theta^\top \Gamma \theta^\top$ *and the mini-batches are sampled with replacement, then*

$$\overline{C}_\psi = \frac{1}{B}\left(\mathcal{I} - \frac{1}{N^2}\Gamma\widehat{\theta\theta}^\top\Gamma^\top + \frac{1}{N}\sum_{n=1}^{N}\mathcal{J}_n\Sigma_\psi\mathcal{J}_n - \mathcal{J}\Sigma_\psi\mathcal{J}\right),$$

where $\mathcal{I} := \frac{1}{N}\sum_{n=1}^{N}\nabla\ell_n\big(\widehat{\theta}\big)\nabla\ell_n\big(\widehat{\theta}\big)^\top$. *If the mini-batches are sampled without replacement, the same result holds but with the right-hand side multiplied by* $(N-B)/(N-1)$.

Wasserstein Distance

How to assess the accuracy of our proxy algorithm? **Wasserstein Distance**

- $$W_2(\pi, \tilde{\pi}) = \inf \mathbb{E}(\|\theta - \tilde{\theta}\|^2)^{1/2},$$

  where the infimum is over all joint distributions of $(\theta, \tilde{\theta})$ such that $\theta \sim \pi$ and $\tilde{\theta} \sim \tilde{\pi}$.
- $W_2(\pi_\theta, \pi_\psi) \leq \varepsilon$ implies that

$$|\sigma_{\theta,d} - \sigma_{\psi,d}| \leq \varepsilon \ (d = 1, \ldots, D)$$
$$\|\Sigma_\theta - \Sigma_\psi\| \leq 2\varepsilon(\|\Sigma_\theta\|^{1/2} \wedge \|\Sigma_\psi\|^{1/2} + \varepsilon).$$

Error Analysis

Theorem

Under standard assumptions and $\Lambda = \lambda I_D$ *for some* $\lambda \in (0, 1/(2L))$, *then, letting* $\beta := 1 - \lambda\mu(1 - 2\lambda L)$, $\overline{M} := \{N^{-1}\sum_{n=1}^{N} M_n^2\}^{1/2}$, *and* $C_s := \mathbb{E}(\|\psi_s - \widehat{\theta}\|^4)$, *for all* $t = 1, 2, \ldots,$

$$W_2^2(\pi_{\theta,t}, \pi_{\psi,t})$$
$$\leq \beta^t W_2^2(\theta_0, \psi_0) + \lambda\overline{M}\left\{\frac{\lambda\overline{M}}{2} + \frac{2}{\mu}\right\}\sum_{s=1}^{t}\beta^{t-s}C_{s-1}.$$

Corollary

Under the same assumptions stated above and with $\beta = \infty$ *(i.e., for the case of SGD), if* $\lambda < L/4$, *then there exists an explicit constant A such that*

$$W_2(\pi_\theta, \pi_\psi) \leq A\frac{\lambda}{B}.$$

Experiments: Linear Regression

To validate our theory, we compare the predicted stationary covariance structure under the fixed learning rate obtained from other theory with others.

- *Simulated misspecified data.*:

$$y_n \sim \mathcal{N}(x_n^\top \theta_\star, 1 + \|x_i\|_2^2),$$

  where $\theta_\star \sim \mathcal{N}(0, I_D)$ is fixed and $x_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I_D)$.
- Real-world dataset: *Boston housing data.*

Experiments: Linear Regression

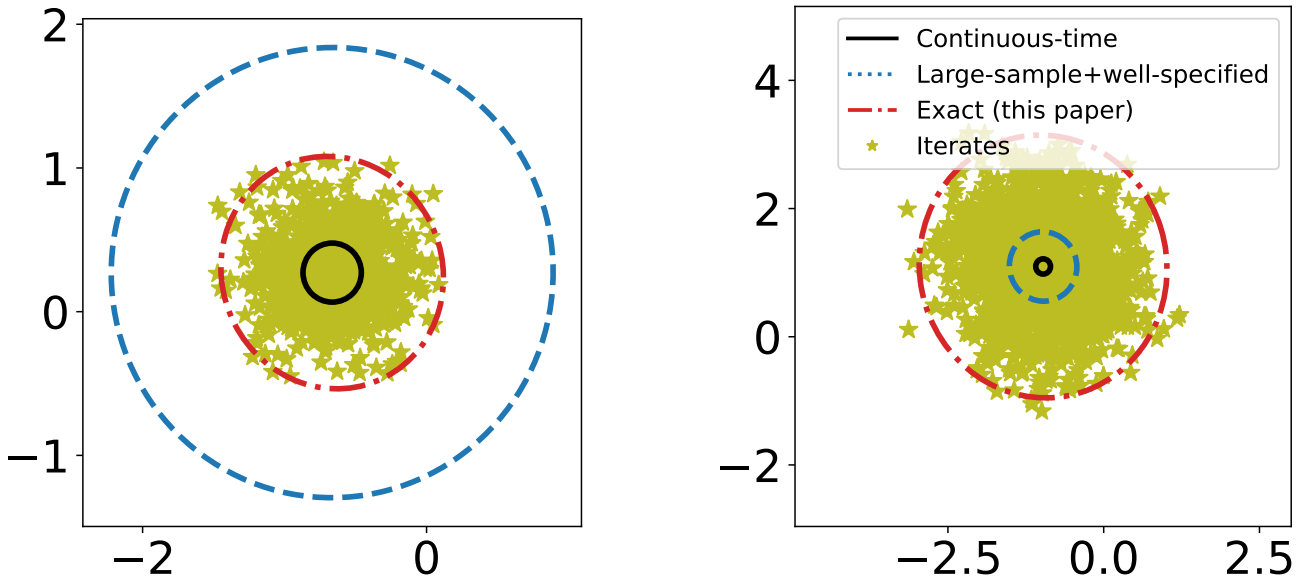


Figure 3: Comparison of estimated stationary covariance structure for linear regression at 3σ confidence region on **(left)** simulated misspecified data with heteroskedastic noise and **(right)** the classic Boston housing dataset with $\lambda = 0.1$ and $B = 32$. Our theory provides more accurate stationary covariance predictions in both cases.

Experiments: Poisson Regression

- *Simulated misspecified data.*:

$$y_n \sim \text{Poisson}(\exp\{x_n^\top \theta_\star\}),$$

  where $\theta_\star \sim \mathcal{N}(0, I_D)$ is fixed and $x_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I_D)$.
- Real-world dataset: *German credit data.*

Experiments: Poisson Regression

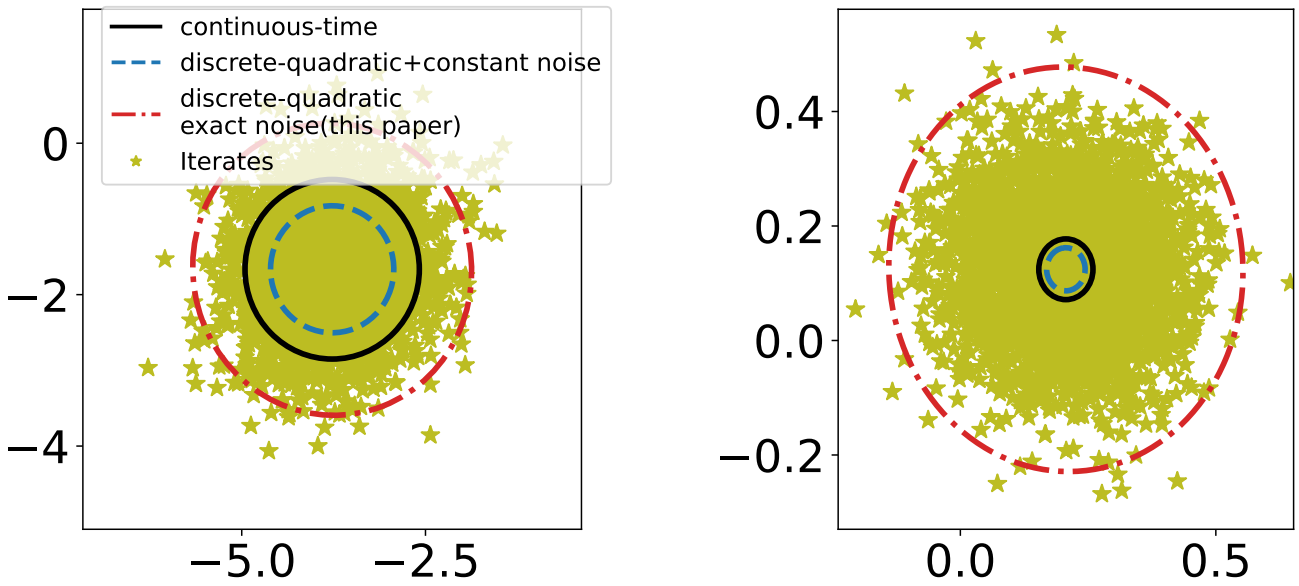


Figure 4: Comparison of estimated stationary covariance structure for Poisson regression at 3σ confidence region with **(left)** simulated well-specified data and **(right)** the German credit data by setting batch size $\lambda = 0.1$, and $B = 32$.

Conclusion

Conclusion

- We propose a diagnostic tool for VI:
  - supports marginal checks and is applicable to high-dimensional parameter spaces
  - provides lower bounds on the error of specific posterior summaries
  - is computationally efficient
  - can be validated using a simple correlation-based reliability check
- We propose tools for stationary covariance analysis of stochastic gradient algorithms:
  - has minimal checkable assumptions.
  - Nonasymptotic error analysis.

Acknowledgments

Supervisor

Committee

Collaborator