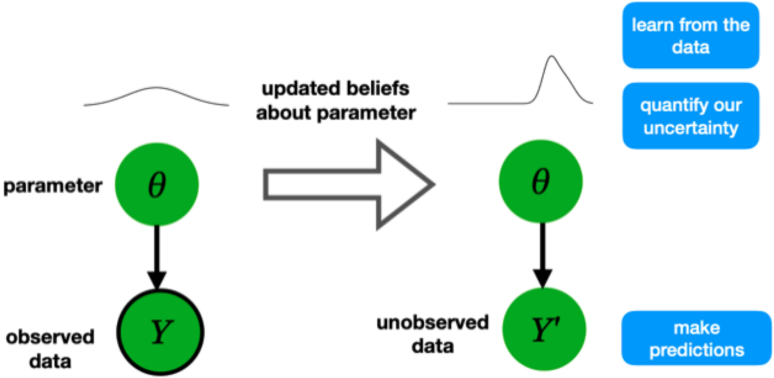
ENHANCING THE RELIABILITY OF GENERAL-PURPOSE ALGORITHMS FOR APPROXIMATE BAYESIAN INFERENCE

Yu Wang

December 4, 2024

BOSTON UNIVERSITY



$$\pi(\theta \mid Y) = \frac{P(Y \mid \theta)\pi_0(\theta)}{Z}$$

Bayesian Approximation

$$\pi(\theta \mid Y) = \frac{P(Y \mid \theta)\pi_0(\theta)}{Z}$$

- We want to learn about π, typically by calculating expectations
    - Find the mean and covariance of θ

Bayesian Approximation

$$\pi\left(\theta \mid Y\right) = \frac{P(Y \mid \theta)\pi_0(\theta)}{Z}$$

- We want to learn about π, typically by calculating expectations
  - Find the mean and covariance of θ
- However, in general, the expectations can't be done exactly

Bayesian Approximation

$$\pi\left(\theta \mid Y\right) = \frac{P(Y \mid \theta)\pi_0(\theta)}{Z}$$

- We want to learn about π, typically by calculating expectations
  - Find the mean and covariance of θ
- However, in general, the expectations can't be done exactly
- Approximate inference:
  - Markov chain Monte Carlo (MCMC)

Challenges in Modern Approximate Bayesian Inference

- **Challenges:**
  - high-dimensional $\theta \in \mathbb{R}^D$, D is large
  - complex relationship $p(Y|\theta)$
  - large-scale dataset

Challenges in Modern Approximate Bayesian Inference

- **Challenges:**
  - high-dimensional $\theta \in \mathbb{R}^D$, D is large
  - complex relationship $p(Y|\theta)$
  - large-scale dataset
- **Problems of MCMC:**
  - Slow convergence and poor mixing in complex cases e.g. those involving heterogeneity or high correlation between coordinates
  - Computational cost of likelihood evaluation is proportional to the dataset size

Challenges in Modern Approximate Bayesian Inference

- **Challenges:**
  - high-dimensional $\theta \in \mathbb{R}^D$, D is large
  - complex relationship $p(Y|\theta)$
  - large-scale dataset
- **Problems of MCMC:**
  - Slow convergence and poor mixing in complex cases e.g. those involving heterogeneity or high correlation between coordinates
  - Computational cost of likelihood evaluation is proportional to the dataset size
- **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)

- **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 $\mathcal{D}_{\pi}(\cdot)$: the canonical choice is *Kullback–Leibler (KL) divergence* out of convenience.

Related Works

- Existing evaluation tools:
  - Evidence Lower Bound (ELBO)
  - Kernel Stein Discrepancy (KSD)[1]
  - Pareto smoothed importance sampling (PSIS) $\hat{k}$[2]
  - W_2 upper bound[3]
- Problems:
  - Lack interpretability
  - Not applicable in high-dimensional parameter spaces.
  - Don't support marginal checks

---

1 Gorham et al. (2017). Measuring sample quality with kernels.International Conference on Machine Learning.

2 Yao et al. (2018) Yes, but did it work?: Evaluating variational inference. 35th International Conference on Machine Learning (ICML).

3 Huggins et al. (2020) Validated variational inference via practical posterior error bounds. AISTATS

TADDAA: Intuition

We want to quantify approximation error $\varepsilon^{(0)} := \mu(\hat{\pi}^{(0)}) - \mu(\pi)$

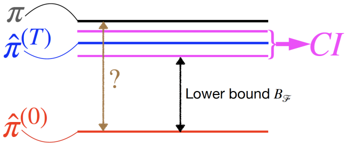


Figure 1: $\hat{\pi}^{(T)}$ significantly different from $\hat{\pi}^{(0)} \Rightarrow \hat{\pi}^{(0)}$ far from π.

- For another approximation $\hat{\pi}^{(T)}$ closer to taget posteriror π, $\varepsilon^{(T)} := \mu(\hat{\pi}^{(T)}) - \mu(\pi)$
- $$\varepsilon^{(0)} \geq |\mu(\hat{\pi}^{(0)}) - \mu(\hat{\pi}^{(T)})|$$

Example: Low-quality VI

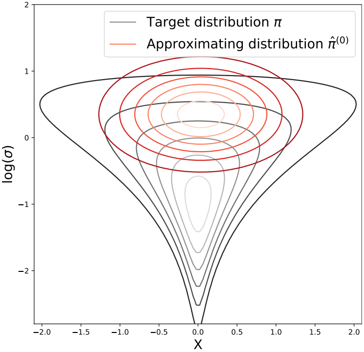
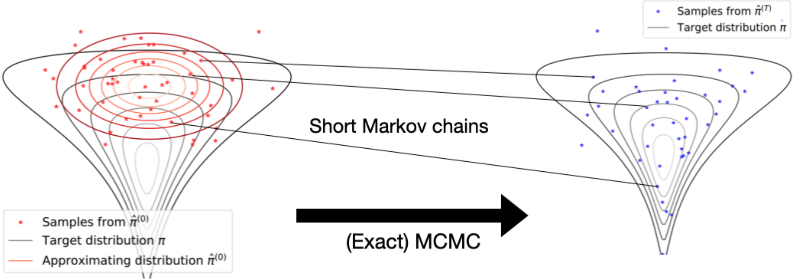


Figure 2: Variational approximation on 2-D Neal-Funnel shape model.

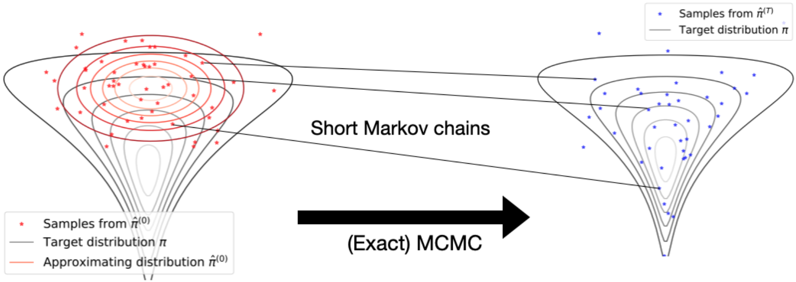
The parameterization of Neal-Funnel shape model is given as follows:

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

TADDAA Framework



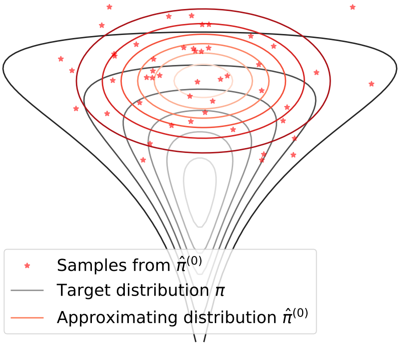
TADDAA Framework



Why not use MCMC directly?

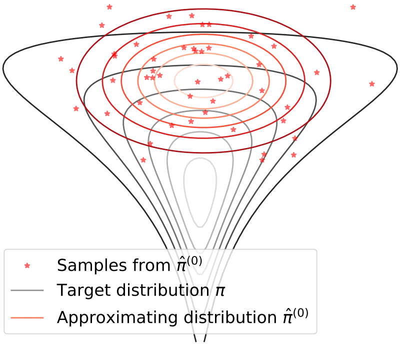
- Slow convergence in complicated cases.
- TADDAA **DOES NOT** rely on convergence of Markov chains.

TADDAA:Input

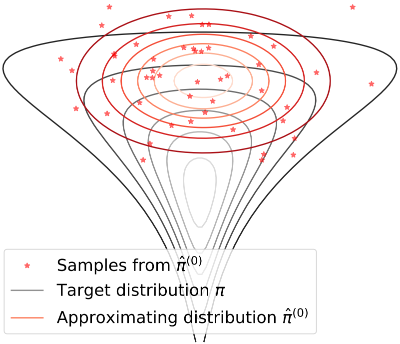


- log density of the target π
- approximating distribution $\hat{\pi}^{(0)}$

TADDAA:Input

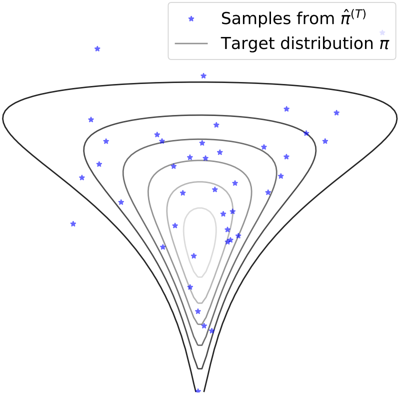


- log density of the target π
- approximating distribution $\hat{\pi}^{(0)}$
- functional of interest $\mathcal{F}$ (e.g. marginal mean)



- 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**:

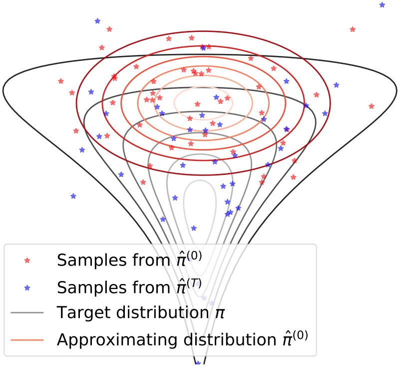
    $X_j^{(t+1)} \sim K_{h^{(t)}}(X_j^{(t)}, \cdot)$

  end for

  update step-size $h^{(t+1)}$ using INCA.

end for

TADDAA:Compute error lower bounds



- Compute a confidence interval $(\ell_{\mathcal{F}}, u_{\mathcal{F}})$ for $\mathcal{F}(\hat{\pi}^{(0)}) - \mathcal{F}(\hat{\pi}^{(T)})$ based on $X_{1:N}^{(0)}$ and $X_{1:N}^{(T)}$
- Compute lower bound $B_{\mathcal{F}}$

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

- Random Walk Metropolis-Hasting (RWMH).
- Metropolis-adjusted Langevin algorithm (MALA).
- Hamiltonian Monte Carlo (HMC).
- **Barker Proposal[4] (recommended choice)**
    - robust to precise step size and acceptance rate.
    - high sampling efficiency.

[4] Livingstone et al. The Barker proposal: combining robustness and efficiency in gradient-based MCMC. Journal of the Royal Statistical Society Series B: Statistical Methodology (2022)

Step size h

Step size adaption:

- **Inter-chain adaptation(INCA)**[5]
  - $Y_j^{(t+1)} \sim Q_{h^{(t)}}(X_j^{(t)}, \cdot)$, accept with probability $\alpha_j^{(t)}$.
  - $\bar{\alpha}^{(t)} = \frac{1}{N} \sum_{j=1}^{N} \alpha_j^{(t)}$.
- **Optimal Scaling**[6]
  - $\log h^{(t+1)} = \log h^{(t)} + \frac{1}{\sqrt{t+1}}(\bar{\alpha}^{(t)} - \bar{\alpha}_*)$, $\bar{\alpha}_*$ is optimal asymptotic acceptance.

[5] Craiu et al. Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. Journal of the American Statistical Association (2009).

[6] Roberts et al. Optimal scaling for various Metropolis-Hastings algorithms. Statistical science (2001).

Step size h

Step size adaption:

- **Inter-chain adaptation(INCA)**[5]
  - $Y_j^{(t+1)} \sim Q_{h^{(t)}}(X_j^{(t)}, \cdot)$, accept with probability $\alpha_j^{(t)}$.
  - $\bar{\alpha}^{(t)} = \frac{1}{N} \sum_{j=1}^{N} \alpha_j^{(t)}$.
- **Optimal Scaling**[6]
  - $\log h^{(t+1)} = \log h^{(t)} + \frac{1}{\sqrt{t+1}}(\bar{\alpha}^{(t)} - \bar{\alpha}_*)$, $\bar{\alpha}_*$ is optimal asymptotic acceptance.

Optimal initial step size $h^{(0)}$ and $\bar{\alpha}_*$:

- e.g. Barker: $h^{(0)} = 2.4^2/d^{1/3}$, $\bar{\alpha}_* = 0.576$.

---

[5] Craiu et al. Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. Journal of the American Statistical Association (2009).

[6] Roberts et al. Optimal scaling for various Metropolis-Hastings algorithms. Statistical science (2001).

Step size h

Step size adaption:

- **Inter-chain adaptation(INCA)**[5]
  - $Y_j^{(t+1)} \sim Q_{h^{(t)}}(X_j^{(t)}, \cdot)$, accept with probability $\alpha_j^{(t)}$.
  - $\bar{\alpha}^{(t)} = \frac{1}{N} \sum_{j=1}^{N} \alpha_j^{(t)}$.
- **Optimal Scaling**[6]
  - $\log h^{(t+1)} = \log h^{(t)} + \frac{1}{\sqrt{t+1}}(\bar{\alpha}^{(t)} - \bar{\alpha}_*)$, $\bar{\alpha}_*$ is optimal asymptotic acceptance.

Optimal initial step size $h^{(0)}$ and $\bar{\alpha}_*$:

- e.g. Barker: $h^{(0)} = 2.4^2/d^{1/3}$, $\bar{\alpha}_* = 0.576$.

Problem: INCA introduces dependence for $X_{1:N}^{(T)}$.

[5] Craiu et al. Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. Journal of the American Statistical Association (2009).

[6] Roberts et al. Optimal scaling for various Metropolis-Hastings algorithms. Statistical science (2001).

Asymptotic Independence of Adapted Markov Chains

$X_{1:N}^{(t)}$ are dependent $\rightarrowtail$ independence assumption of tests violated.

Asymptotic Independence of Adapted Markov Chains

$X_{1:N}^{(t)}$ are dependent $\rightarrowtail$ independence assumption of tests violated.

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

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.*

Number of Markov chains N

N is determined by user's tolerance for statistical test error δ, e.g.

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

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 [7].

- 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.

[7] Roberts et al. Optimal scaling for various Metropolis-Hastings algorithms. Statistical science (2001).

[8] 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 [7].

- 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

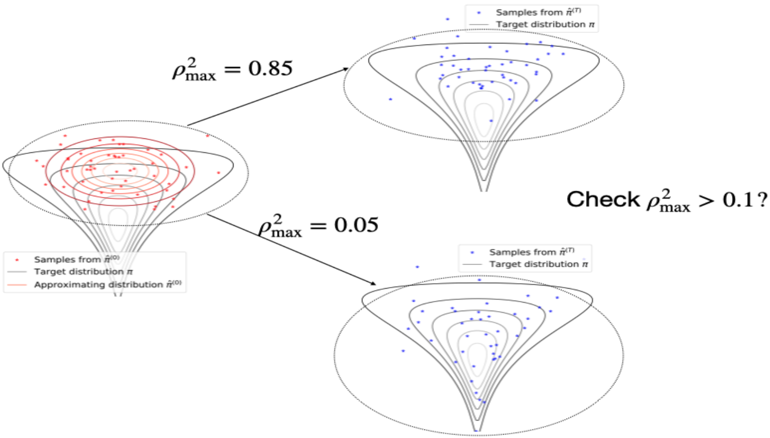
- Computational cost of TADDAA is comparable to VI:
  - Computational cost for VI: $\Theta(d^{1/3})$[8].
  - Computational cost for MALA and Barker: $\Theta(d^{1/3})$.
  - Computational cost for HMC: $\Theta(d^{1/4})$.

[7] Roberts et al. Optimal scaling for various Metropolis-Hastings algorithms. Statistical science (2001).

[8] 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:



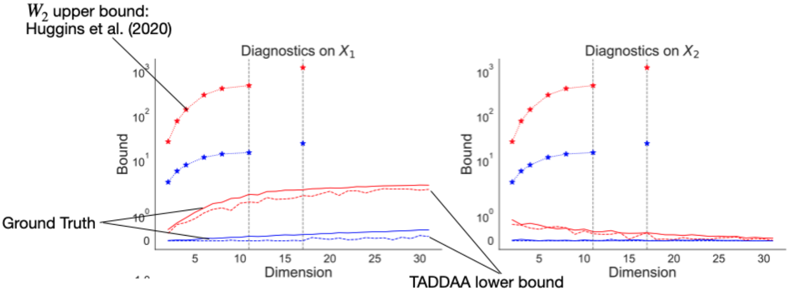


Figure 3: Diagnostics for Neal-funnel shape model, where TADDAA uses the Barker proposal. Here μ_i and σ_i^2 denote, respectively, the mean and variance of X_i.

Revisit (High-dimensional) Neal-Funnel Shape Model

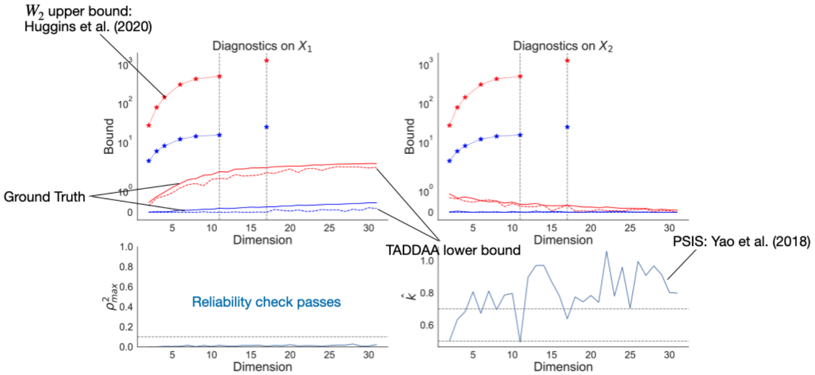
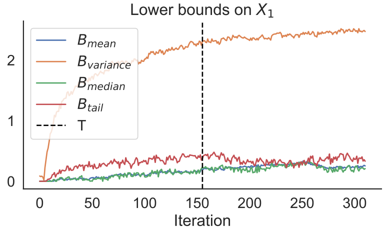
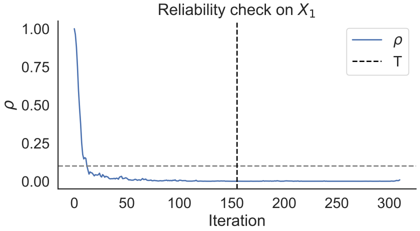


Figure 4: Diagnostics for Neal-funnel shape model, where TADDAA uses the Barker proposal. Here μ_i and σ_i^2 denote, respectively, the mean and variance of X_i.

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

Use a logistic regression model with a sparsity-inducing horseshoe prior on

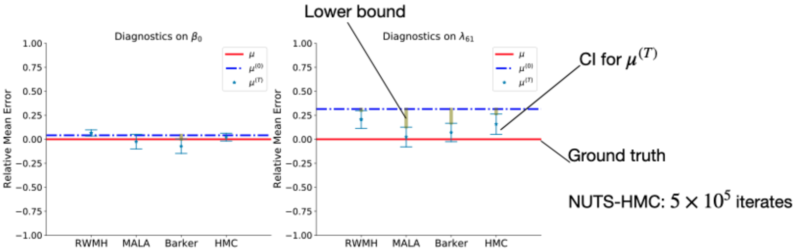
$$\begin{aligned}
y \mid \beta &\sim \mathrm{Bern}(\mathrm{logit}^{-1}(X\beta)), \\
\beta_j \mid \tau, \lambda, c &\sim \mathcal{N}(0, \tau^2 \tilde{\lambda}_j^2), \\
\lambda_j &\sim \mathrm{C}^+(0,1), \qquad \tau \sim \mathrm{C}^+(0, \tau_0), \\
c^2 &\sim \mathrm{InvGam}(2, 8),
\end{aligned}$$

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

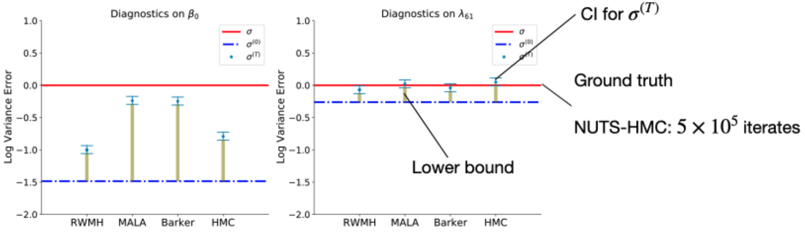
- $X \in \mathbb{R}^{71 \times 100}$.
- Parameter dimension $d = 203$.

Logistic Regression Using Horseshoe Prior: Mean Diagnostic

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

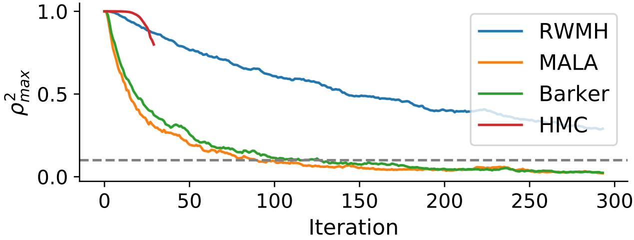


Logistic Regression Using Horseshoe Prior: Variance Diagnostic



Logistic Regression Using Horseshoe Prior: Reliability Check

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



Conclusion

We propose a robust 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

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

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

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 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

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

- Accurately estimate the stationary covariance Σ_θ under fixed learning rates:
    - Test loss
    - Escaping efficiency from a sharp minimal[9]
- Learning rate tuning guidance on optimal uncertainty quantification[10]

[9] Zhu et al. (2019). The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In International Conference on Machine Learning, pages 7654–7663. PMLR.

[10] Negrea et al. (2022). Tuning Stochastic Gradient Algorithms for Statistical Inference via Large-Sample Asymptotics.

Related Work: Quadratic Loss

Current works assume that the loss is well-approximated by a quadratic function[11]:

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

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

[11] Mandt et al. (2017). Stochastic Gradient Descent as Approximate Bayesian Inference. Journal of Machine Learning Research.

Related Work: Continuous-time Proxies

Approximate SGD by a continuous-time the Ornstein–Uhlenbeck (OU) process[12]

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

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

[12] Negrea et al. (2022). Tuning Stochastic Gradient Algorithms for Statistical Inference via Large-Sample Asymptotics.

Related Work: Continuous-time Proxies

Approximate SGD by a continuous-time the Ornstein–Uhlenbeck (OU) process[12]

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

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

- The covariance matrix of the stationary distribution $\Sigma_\vartheta := \mathrm{Cov}(\pi_\vartheta)$ satisfies

$$\Sigma_\vartheta \widehat{H} + \widehat{H} \Sigma_\vartheta = \Lambda \widehat{C}.$$

[12] Negrea et al. (2022). Tuning Stochastic Gradient Algorithms for Statistical Inference via Large-Sample Asymptotics.

Related Work: Continuous-time Proxies

Approximate SGD by a continuous-time the Ornstein–Uhlenbeck (OU) process[12]

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

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

- 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**.

[12] Negrea et al. (2022). Tuning Stochastic Gradient Algorithms for Statistical Inference via Large-Sample Asymptotics.

Related Work: Discrete-time proxies

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

[13] Liu et at. Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent. ICML (2021).

[14] Liu et at. Strength of Minibatch Noise in SGD. ICLR (2022).

Related Work: Discrete-time proxies

Under 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[13]:

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

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

- For well-specified linear model and assume $X \sim \mathcal{N}(0, A)$[14]:

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

[13] Liu et at. Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent. ICML (2021).

[14] Liu et at. Strength of Minibatch Noise in SGD. ICLR (2022).

Limitations of discrete-time proxies

Limitation:

- Assumptions often do not hold in practice:
  - Sample size $N \gg D$ and $N \to \infty$
  - 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 $(\theta_t)_{t \geq 0}$.

Limitations of discrete-time proxies

Limitation:

- Assumptions often do not hold in practice:
  - Sample size $N \gg D$ and $N \to \infty$
  - 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 $(\theta_t)_{t\geq 0}$.

Contribution:

- Propose a new discrete-time proxy algorithm that delivers more accurate stationary covariance estimates for:
  - Finite sample size N
  - More general convex loss
  - Misspecified model
- Provide quantitative, non-asymptotic error analysis of our approximation.

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

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

- 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:

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

Proposition

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

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

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*

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

where $\mathcal{I} := \frac{1}{N}\sum_{n=1}^{N} \nabla \ell_n\left(\widehat{\theta}\right)\nabla \ell_n\left(\widehat{\theta}\right)^\top$, $\mathcal{J}_n = \nabla^2 \ell_n\left(\widehat{\theta}\right)$.

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 [15]

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

[15] Huggins et al. Validated variational inference via practical posterior error bounds. AISTATS (2020)

Corollary

Under the same assumptions stated above and with $\beta = \infty$ (i.e., for the case of SGD), there exists $L > 0$, if $\lambda < L/4$, then there exists an explicit constant A such that

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

Experiments: Linear Regression

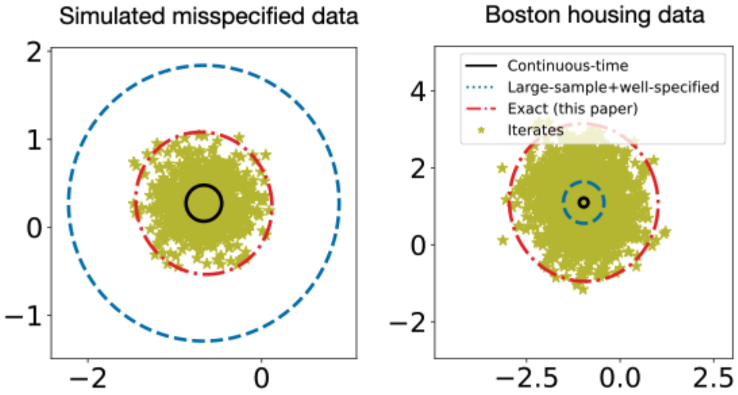
To validate our theory, we compare the predicted stationary covariance structure under the fixed learning rate obtained from other theory with others.

- *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



Experiments: Poisson Regression

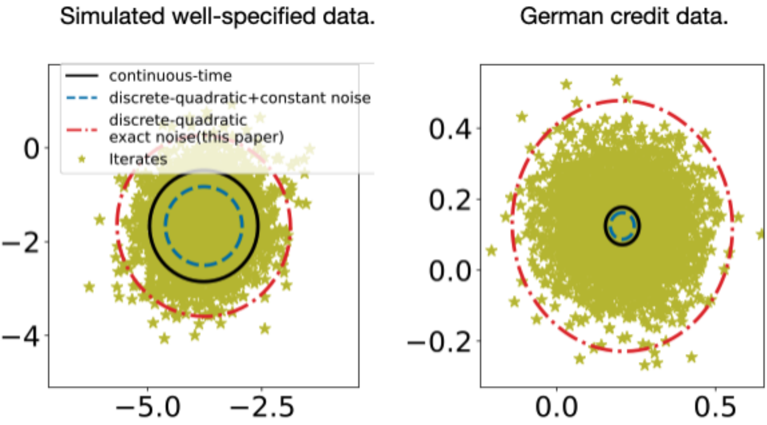
- *Simulated well-specified data.*:

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

  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



Conclusion

We have established a rigorous framework for understanding SGD and SGLD under:

- large learning rate.
- N is not large compared to D.
- model is incorrect.

Conclusion

Conclusion

1. Post hoc quality check for variational approximation:
   - support marginal checks
   - robust in high-dimensional parameter sapce
   - computationally efficient
2. Uncertainty quantification for subsampling methods:
   - Accurate stationary covariance structure estimation for stochastic gradient algorithms under fixed/nonvanishing learning rate
   - Optimal learning rate tuning guidance

Acknowledgments

Advisor

Jonathan H. Huggins

Committee

Emily Stephen

Luis Carvalho

Konstantinos Spiliopoulos

collaborator

Tamara Broderick

Mikolaj Kasprzak

Trevor Campbell