# Robust Approximate Sampling via Stochastic Gradient Barker Dynamics

Lorenzo Mauri[1], Giacomo Zanella[2]

[1]Duke University
[2]Bocconi University

May 2, 2024

# Main Contribution

## Stochastic Gradient Barker Dynamics Algorithm

- We develop the **Stochastic Gradient Barker Dynamics Algorithm (SGBD)** by extending the stochastic gradient MCMC framework to the Barker Proposal (Livingstone and Zanella [5]).
- We study the **bias** introduced by the stochastic gradient noise and devise **strategies to eliminate or reduce it**.
- We compare SGBD to the stochastic gradient Langevin dynamics algorithm (SGLD, Welling and Teh [9]) in numerical examples where SGBD displays **greater robustness to hyperparameter tuning**.

# Outline

# Analytical setting

**Task:** sampling from a target distribution of the form

$$\pi(\theta) \propto p(\theta) \prod_{i=1}^{N} p(y_i|x_i, \theta) = \exp\left(g(\theta)\right), \quad \theta \in \mathbb{R}^d \tag{1}$$

where $p(\theta)$ is the prior distribution and $p(y_i|x_i, \theta)$ is the likelihood of the $i$-th observation.

The **gradient** of $g(\theta)$ is a **sum of $N$ terms**,

$$\partial_j g(\theta) = \sum_{i=1}^{N} \partial_j g_i(\theta), \quad \partial_j g_i(\theta) = \frac{1}{N} \partial_j \log\left(p(\theta)\right) + \partial_j \log\left(p(y_i \mid x_i, \theta)\right)$$

where $\partial_j$ stands for the partial derivative with respect to the $j^{th}$ component of $\theta$.

$\rightarrow$ computing the gradient results in a **$\Theta(N)$ cost** per iteration.

# State of the Art

Gradient based MCMC:

- explore the space efficiently;
- can be **computationally expensive** (**$\Theta(N)$ cost** per iteration).

## Stochastic gradient MCMC

Stochastic gradient MCMC have gained popularity as they combine **scalability** to big size datasets with an **efficient exploration** of the space.

# Stochatic Gradient MCMC

SG-MCMC (e.g. SGLD, Welling and Teh [9]) replace the gradient with a mini-batch estimate

$$\hat{\partial}_j g(\theta) = \frac{N}{n} \sum_{i \in \mathcal{S}_n} \partial_j g_i(\theta) \qquad j = 1, \dots, p, \tag{2}$$

where $\mathcal{S}_n$ is a subset of $\{1, \dots, N\}$ of size $n \ll N$ sampled uniformly at random.

- Most SG-MCMC methods converge to the true posterior distribution if the **step-size is appropriately decreased to zero** (Welling and Teh [9], Chen et al. [3], Ding et al. [4], and Ma et al. [6]).
- This strategy **deteriorates mixing**, and practitioners usually keep the **step-size fixed**, which leads to **non-negligible bias** in the invariant distribution (Brosse et al. [2]).

# The Barker Proposal (I)

The Barker Proposal (Livingstone and Zanella [5]) is a novel MCMC that outperforms other gradient based algorithms in terms of **robustness to hyperparameter tuning**.

## Barker Proposal PDF

The Barker Proposal is a **first order approximation of a $\pi$-invariant process** and its PDF is given by

$$
\begin{aligned}
Q_B(\theta, \theta + w) &= \prod_{i=1}^{d} 2p(\partial_j g(\theta), w_j)\mu_\sigma(w_j) \quad \theta, w \in \mathbb{R}^d \\
p(\delta, z) &= (1 + \exp(-z\delta))^{-1} \qquad\qquad \delta, z \in \mathbb{R}
\end{aligned}
\tag{3}
$$

In our experiments, we take $\mu_\sigma = 0.5\mathcal{N}(-\sigma, (0.1\sigma)^2) + 0.5\mathcal{N}(\sigma, (0.1\sigma)^2)$, as recommended in Vogrinc et al. [8].

# The Barker Proposal (II)

**Algorithm 1:** Unadjusted Barker Proposal

**Input:** $\theta^{(0)} \in \mathbb{R}^d, \sigma > 0$

**for** $t = 1, \ldots, T$ **do**

  **for** $j = 1, \ldots, d$ **do**

    Draw $w_j^{(t)} \sim \mu_\sigma(\cdot)$;

    Set $b_j^{(t)} = 1$ with probability $p(\partial_j g(\theta^{(t-1)}), w_j^{(t)})$, otherwise $b_j^{(t)} = -1$;

    Update $\theta_j^{(t)} \leftarrow \theta_j^{(t-1)} + b_j^{(t)} w_j^{(t)}$ ;

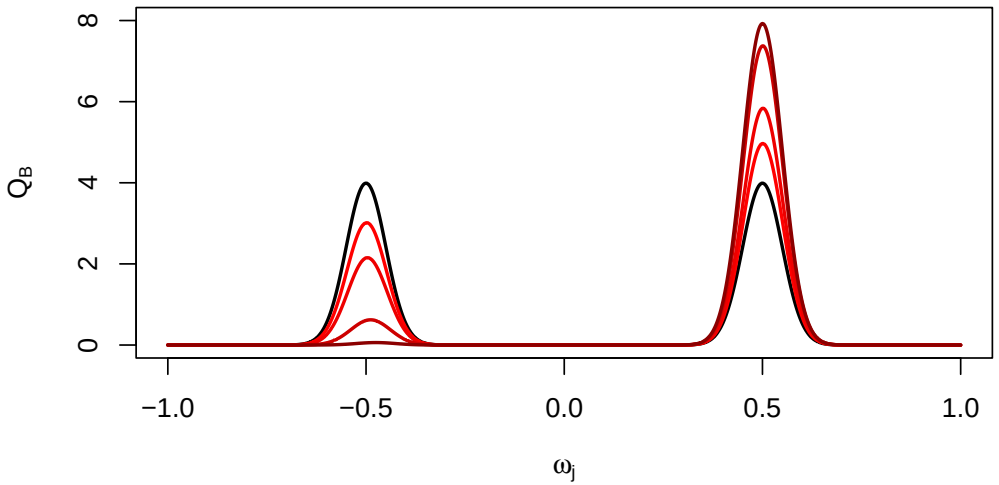



**Fig. 1.** $\mu_\sigma$ with $\sigma = 0.5$ (black) and Barker Proposal $Q_B$ for increasing values of $\partial_j g(\theta)$ (red).

# The Barker Proposal (III)

## Combining Robustness and Efficiency

- The size of the increment $|w_j|$ is **independent** of the gradient $\partial_j g(\theta)$ $\rightarrow$ **increased robustness** to hyperparameter tuning and target heterogeneity.
- $Q_B$ has the **favourable high-dimensional scaling** properties such as a scaling of order $d^{-1/3}$ as $d$ diverges (Vogrinc et al. [8]).

# The Stochastic Gradient Barker Dynamics Algorithm

We apply the stochastic gradient MCMC framework to the Barker proposal and develop three variants of the Stochastic Gradient Barker Dynamics Algorithm (SGBD):

1. Vanilla SGBD (v-SGBD),
2. Corrected SGBD (c-SGBD),
3. Extreme SGBD (e-SGBD).

# Vanilla SGBD

v-SGBD replaces $\partial_j g(\theta)$ with $\hat{\partial}_j g(\theta)$ in the Barker Proposal and is **equivalent to the Barker proposal with gradient shrunk towards zero**.

## Proposition 1

If $\hat{\partial}_j g(\theta) \sim \mathcal{N}\left(\partial_j g(\theta), \tau_\theta^2\right)$, we have

$$\left| \mathbb{E}\left[ p\left( \hat{\partial}_j g(\theta), w_j \right) \right] - p\left( c_{w_j,\tau_\theta} \partial_j g(\theta), w_j \right) \right| < 0.019, \tag{4}$$

where $c_{w_j,\tau_\theta} := \frac{1.702}{\sqrt{1.702^2 + w_j^2 \tau_\theta^2}}$.

The practical implication of (4) is an **inflation of the variance** of the stationary distribution.

# Corrected SGBD

c-SGBD solves the issue replacing $p\left(\partial_j g(\theta), w_j\right)$ with $\tilde{p}\left(\hat{\partial}_j g(\theta), w_j\right)$, where

$$\tilde{p}\left(\delta, z\right) := \begin{cases} p\left(\frac{1.702}{\sqrt{1.702^2 - \tau_\theta^2 z^2}}\delta, z\right) & \text{if } |z| < \frac{1.702}{\tau_\theta}, \\ \mathbf{1}\left(\delta z > 0\right) & \text{otherwise} \end{cases}, \tag{5}$$

with $\mathbf{1}(A)$ denoting the indicator function of the event $A$ and $\tau_\theta$ is the standard deviation of $\hat{\partial}_j g(\theta)$. In practice, we adopt an online estimate for $\tau_\theta$.

## Proposition 2 (Approximate unbiasedness of $\tilde{p}$) - Informal

If $\hat{\partial}_j g(\theta) \sim \mathcal{N}\left(\partial_j g(\theta), \tau_\theta^2\right)$, and $\tau_\theta$ is sufficiently small, then

$$\left|\mathbb{E}\left[\tilde{p}\left(\hat{\partial}_j g(\theta), w_j\right)\right] - p(\partial_j g(\theta), w_j)\right| < 0.019. \tag{6}$$

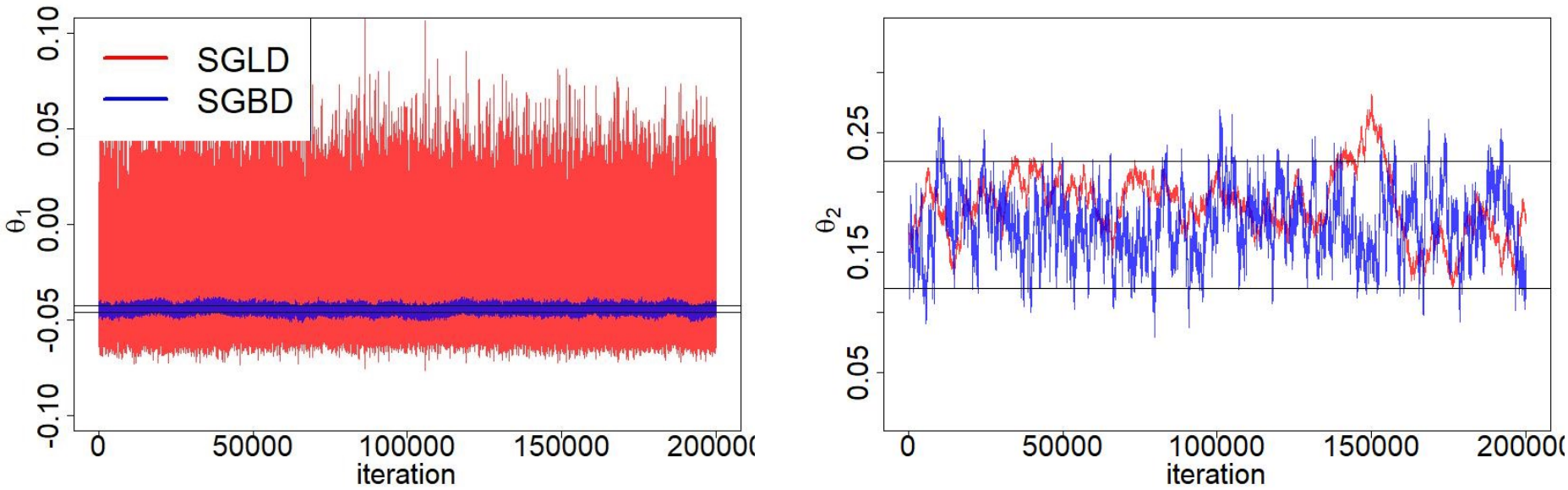
# Extreme SGBD

e-SGBD replaces $p\left(\partial_j g(\theta), w_j\right)$ with $\bar{p}\left(\hat{\partial}_j g(\theta), w_j\right) = \mathbf{1}_{\{\hat{\partial}_j g(\theta) w_j > 0\}}$.

- $\bar{p}$ minimizes the bias for $p$ if $\tau_\theta$ is large among all symmetric estimators for $p$ (Proposition 4 in the article).
- Hybrid between sampling and optimization algorithm (always moves in the direction of the stochastic gradient).

# Binary Regression with Scale Heterogeneity (N=80000, d=4, n=800)

We apply a Bayesian logistic regression to the Sepsis dataset from the UCI repository. We do not scale the covariates to induce scale eterogeneity in the posterior.



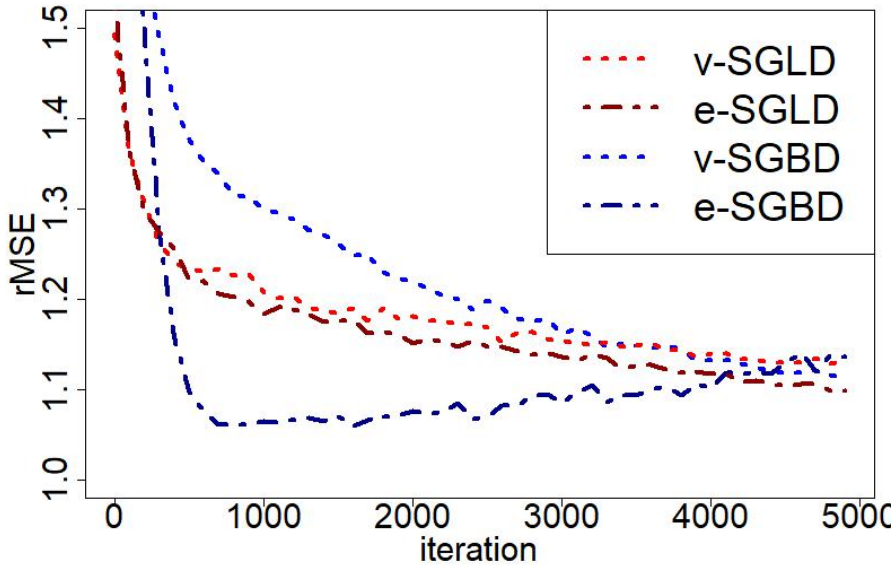

**(a)** Small posterior s.d.

**(b)** Large posterior s.d.

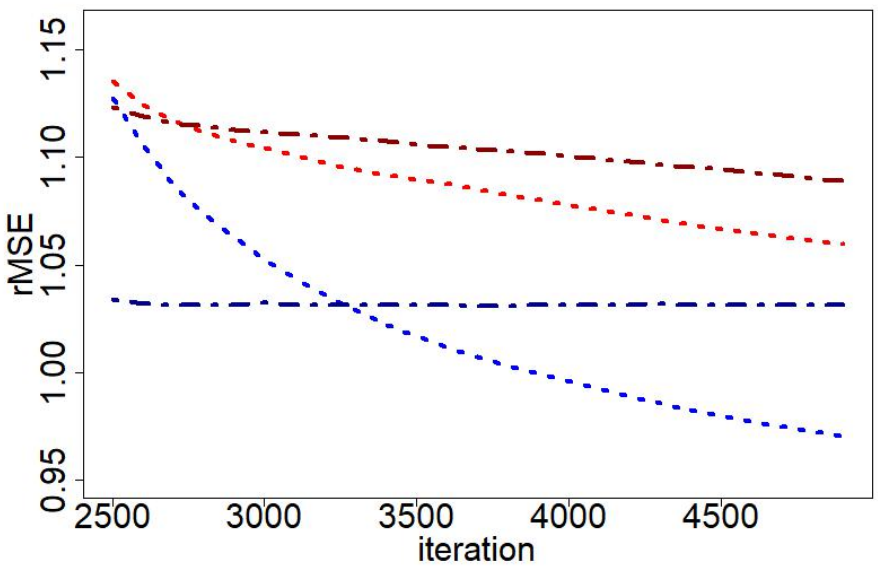
**Fig. 2.** Red refers to v-SGLD and blue to v-SGBD. Black horizontal lines indicate the interval centered at the posterior mean with two standard deviations width.

# Bayesian Matrix Factorization (N=80000, d=54080, n=800)

We apply a Bayesian matrix factorization model (Salakhutdinov and Mnih [7]) to the MovieLens dataset.




**(a)** Samples rMSE

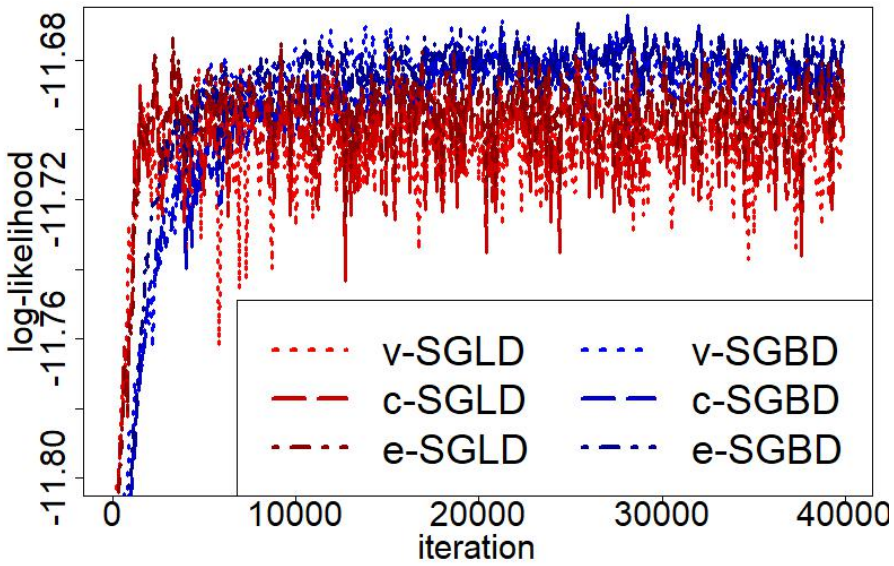



**(b)** MCMC rMSE

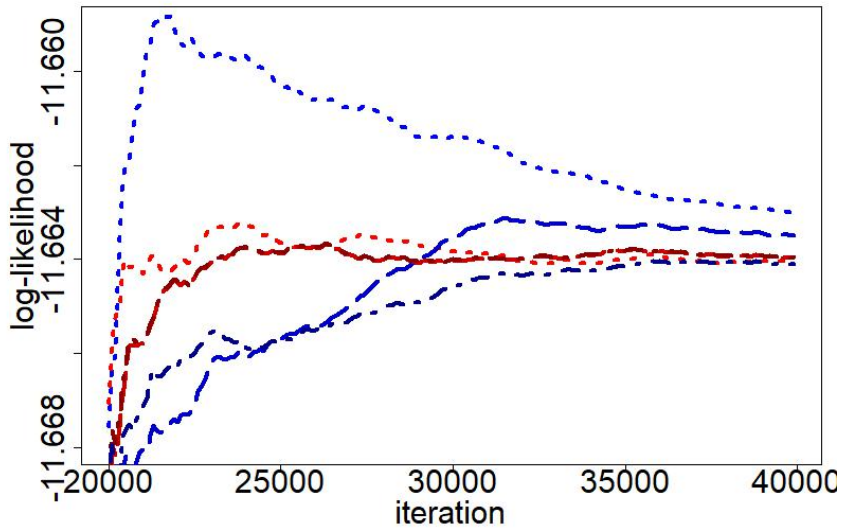
**Fig. 3.** Samples (left) and MCMC (right) estimates rMSE. Red refers to SGLD and blue to the SGBD. Lighter and dotted lines refer to vanilla implementations of the algorithm, darker and dashed-dotted lines to their extreme variants.

# Independent Component Analysis (N=14184, d=100, n=100)

We apply an independent component analysis (Amari et al. [1]) model to the MEG data.




**(a)** Samples log-likelihood




**(b)** MCMC log-likelihood

**Fig. 4.** Log likelihood of each sample (left) and MCMC estimates (right) on held-out data. Red refers to SGLD and blue to SGBD. For both algorithms, the vanilla (lighter dotted lines), corrected (medium scale dashed lines) and extreme (darker dotted-dashed lines) versions are displayed.

# Conclusion

1. We extended the **Barker proposal to the stochastic gradient setting**, leading to the SGBD algorithm.
2. We studied the bias induced by stochastic gradient noise in the Barker proposal and develop strategies to address it.
3. In numerical experiments, SGBD is **more robust to hyperparameters choice** and **to heterogeneity** in the target gradients (arising from e.g. skewness or ill-conditioning)
4. Interesting extensions include adding momentum (similarly to, e.g., SGHMC (Chen et al. [3])); developing adaptive variants that optimally tunes the step-size across iterations; studying connections with optimization schemes.

# References I


[1] Shun-ichi Amari, Andrzej Cichocki, and Howard Yang. "A New Learning Algorithm for Blind Signal Separation". In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky, M.C. Mozer, and M. Hasselmo. Vol. 8. MIT Press, 1995.

[2] Nicolas Brosse, Alain Durmus, and Eric Moulines. "The promises and pitfalls of Stochastic Gradient Langevin Dynamics". In: *Advances in Neural Information Processing Systems* 31 (2018).

[3] Tianqi Chen, Emily Fox, and Carlos Guestrin. "Stochastic Gradient Hamiltonian Monte Carlo". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, June 2014, pp. 1683–1691.

# References II


[4] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. "Bayesian Sampling Using Stochastic Gradient Thermostats". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014.

[5] Samuel Livingstone and Giacomo Zanella. "The Barker proposal: Combining robustness and efficiency in gradient-based MCMC". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2022).

[6] Yi-An Ma, Tianqi Chen, and Emily B. Fox. "A Complete Recipe for Stochastic Gradient MCMC". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 2917–2925.

# References III


[7] Ruslan Salakhutdinov and Andriy Mnih. "Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, 2008, pp. 880–887.

[8] Jure Vogrinc, Samuel Livingstone, and Giacomo Zanella. "Optimal design of the Barker proposal and other locally balanced Metropolis–Hastings algorithms". In: *Biometrika* (Oct. 2022).

[9] Max Welling and Yee Whye Teh. "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 681–688.

# Corrected SGBD - Pseudocode

**Algorithm 2:** Corrected Stochastic Gradient Barker Dynamics (c-SGBD)

**input:** $\theta^{(0)} \in \mathbb{R}^d, \sigma > 0, \beta \in (0,1), \{\hat{\tau}_j^{(0)}\}_{j=1,\dots,d}$

**for** $t = 1, \dots, T$ **do**

  Draw $\mathcal{S}_n \subset \{1, \dots, N\}$ uniformly at random;

  **for** $j = 1, \dots, d$ **do**

    Compute $\hat{\partial}_j g(\theta^{(t-1)})$ using (2);

    Update $\hat{\tau}_j^{(t)} \leftarrow$

    $(1-\beta)\hat{\tau}_j^{(t-1)} + \beta\sqrt{\sum_{i\in\mathcal{S}_n} \frac{(\partial_j g_i(\theta^{(t-1)}) - \frac{1}{n}\sum_{i\in\mathcal{S}_n}\partial_j g_i(\theta^{(t-1)}))^2}{n-1}}$;

    Draw $w_j^{(t)} \sim N(\sigma, (0.1\sigma)^2)$;

    Set $b_j^{(t)} = 1$ with probability $\tilde{p}\left(\hat{\partial}_j g(\theta^{(t-1)}), w_j^{(t)}\right)$, where $\tau_\theta$ is replaced by $\hat{\tau}_j^{(t)}$, otherwise $b_j^{(t)} = -1$;

    Update $\theta_j^{(t)} \leftarrow \theta_j^{(t-1)} + b_j^{(t)} w_j^{(t)}$ ;