Accelerated Convergence of Stochastic Heavy Ball Method under Anisotropic Gradient Noise

Rui Pan[1*] Yuxing Liu[2*] Xiaoyu Wang[1] Tong Zhang[3]

[1]Hong Kong University of Science and Technology [2]Fudan University
[3]University of Illinois Urbana-Champaign

ICLR 2024: The 12th International Conference on Learning Representations

Outline

Introduction: Motivation

Facts about Stochastic Heavy Ball (SHB) Method:

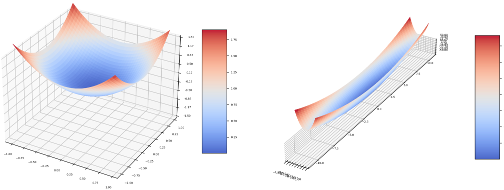
- In practice, SHB is widely adopted to provide acceleration.
- In theory, few results show SHB can provide acceleration.
- SHB cannot provide acceleration unless batch size is large or noise is special [Jain,2018].

Problem Setup

We focus on optimizing quadratic target function

$$\min_{\mathbf{w}} f(\mathbf{w}) \triangleq \mathbb{E}_{\xi}\left[f(\mathbf{w},\xi)\right], \text{ where } f(\mathbf{w},\xi) = \frac{1}{2}\mathbf{w}^{\top}\mathbf{H}(\xi)\mathbf{w} - \mathbf{b}(\xi)^{\top}\mathbf{w},$$

We denote $\mathbf{H} = \mathbb{E}_{\xi}\left[\mathbf{H}(\xi)\right]$ and $\kappa = \lambda_{\max}(\mathbf{H})/\lambda_{\min}(\mathbf{H})$.



Left: Loss surface of a typical quadratic objective; **Right**: Loss surface of a skewed quadratic objective when κ is large.

Problem Setup

We denote the gradient noise to be

$$\mathbf{n}_t \triangleq \nabla f(\mathbf{w}_t) - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \xi)$$

and make the following assumptions:

1. Independent gradient noise: $\{\mathbf{n}_t\}$ are pairwise independent.
2. Unbiased gradient noise: $\mathbb{E}[\mathbf{n}_t] = \mathbf{0}$.
3. Anisotropic gradient noise: $\mathbb{E}\left[\mathbf{n}_t \mathbf{n}_t^\top\right] \preceq \sigma^2 \mathbf{H}$.

Main Result

Acceleration of SHB is attainable while still achieving near-optimal convergence rates.

Corollary (main result)

Given a quadratic objective $f(\mathbf{w})$ and a step decay learning rate scheduler and momentum defined in the Theorem, with $T \geq \tilde{\Omega}\left(\sqrt{\kappa}\right)$, the output of the algorithm satisfies

$$\mathbb{E}\left[f(\mathbf{w}_T) - f(\mathbf{w}_*)\right] \leq \mathbb{E}\left[f(\mathbf{w}_0) + f(\mathbf{w}_1) - 2f(\mathbf{w}_*)\right] \cdot \exp\left(-\tilde{\Omega}\left(\frac{T}{\sqrt{\kappa}}\right)\right) + \tilde{\mathcal{O}}\left(\frac{d\sigma^2}{MT}\right),$$

where we use $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to hide the log factors.

Outline

Proof Sketch

We can divide our proof into 3 main parts as follow:

1. Bias-Variance Decomposition
2. **Dealing with Matrix Products**
3. Applying to Convergence Analysis

Bias-Variance Decomposition

$$\mathbb{E}\left[f(\mathbf{w}_T)\right] - f(\mathbf{w}_*) = B_T + V_T, \quad \text{where}$$

$$B_T \triangleq \sum_{j=1}^{d} \lambda_j \left\| \mathbf{T}_{T-1,j}\mathbf{T}_{T-2,j} \ldots \mathbf{T}_{1,j} \right\|^2 \mathbb{E} \left\| \left(\mathbf{\Pi}^\top \mathbf{V}^\top \begin{bmatrix} w_1 - w_* \\ w_0 - w_* \end{bmatrix} \right)_{2j-1:2j} \right\|^2,$$

$$V_T \triangleq \sigma^2 \sum_{j=1}^{d} \lambda_j^2 \sum_{\tau=1}^{T-1} \eta_\tau^2 \left\| \mathbf{T}_{T-1,j}\mathbf{T}_{T-2,j} \ldots \mathbf{T}_{\tau+1,j} \right\|^2.$$

Here the momentum matrix is defined as

$$\mathbf{T}_{t,j} \triangleq \begin{bmatrix} 1 + \beta - \eta_t \lambda_j & -\beta \\ 1 & 0 \end{bmatrix},$$

where λ_j is the j-th eigenvalue of $\mathbf{H}$.

One Possible Approach

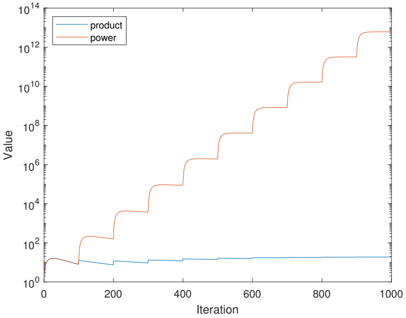
Lemma (bounding matrix power)

Given momentum matrices $\mathbf{T}_{t,j}$, for all positive integers k, it holds that

$$\left\|\mathbf{T}_{t,j}^{k}\right\|_F \leq \min\left(8k, \frac{8}{\sqrt{|(1+\beta-\eta_t\lambda_j)^2-4\beta|}}\right)\rho(\mathbf{T}_{t,j})^k.$$

However, in this way there will be additional κ every stage, which makes the bound not tight and even causes loss explode.

Loss Explosion



Blue: $\left\|\mathbf{T}_{T-1,j}\mathbf{T}_{T-2,j}\ldots\mathbf{T}_{1,j}\right\|$

Orange: $\left\|\left(\mathbf{T}'_{n_l,j}\right)^{k_l}\right\|\left\|\left(\mathbf{T}'_{n_l-1,j}\right)^{k_l}\right\|\ldots\left\|\left(\mathbf{T}'_{1,j}\right)^{k_l}\right\|$

Novel Technique

The key is utilizing the fact that every $T_{t,j}$ does not differ too much.

Lemma (from matrix power to matrix product)

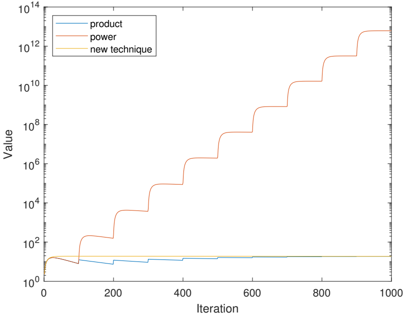
Given matrices $\mathbf{T}_{t,j}$ *and* $\mathbf{\Delta}_i$, $\mathbf{\Delta}$ *defined as*

$$\mathbf{T}_{t,j} = \begin{bmatrix} 1+\beta-\eta_t\lambda_j & -\beta \\ 1 & 0 \end{bmatrix}, \quad \mathbf{\Delta}_i = \begin{bmatrix} \delta_i & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{\Delta} = \begin{bmatrix} \delta & 0 \\ 0 & 0 \end{bmatrix},$$

where $\delta_i \geq 0$, $\delta = \max_{1\leq i\leq k} \delta_i$, *if* $(1+\beta-\eta_t\lambda_j)^2 - 4\beta \geq 0$, *it holds that*

$$\left\|(\mathbf{T}_{t,j}+\mathbf{\Delta}_1)(\mathbf{T}_{t,j}+\mathbf{\Delta}_2)\ldots(\mathbf{T}_{t,j}+\mathbf{\Delta}_k)\right\|_F \leq \left\|(\mathbf{T}_{t,j}+\mathbf{\Delta})^k\right\|_F.$$

Novel Technique



Blue: $\|\mathbf{T}_{T-1,j}\mathbf{T}_{T-2,j}\ldots\mathbf{T}_{1,j}\|$ Yellow: $\left\|\left(\mathbf{T}'_{n_l,j}\right)^{T}\right\|$

Orange: $\left\|\left(\mathbf{T}'_{n_l,j}\right)^{k_l}\right\|\left\|\left(\mathbf{T}'_{n_l-1,j}\right)^{k_l}\right\|\ldots\left\|\left(\mathbf{T}'_{1,j}\right)^{k_l}\right\|$

From Matrix Power to Matrix Product

Proof: discuss in two kinds of product of $\mathbf{T}$ and $\mathbf{\Delta}$

$$
\begin{aligned}
& (\mathbf{T}_{t,j} + \mathbf{\Delta}_1)(\mathbf{T}_{t,j} + \mathbf{\Delta}_2)...(\mathbf{T}_{t,j} + \mathbf{\Delta}_n) \\
\Rightarrow \quad & \mathbf{T}_{t,j}^{k_1}\mathbf{\Delta}_1 \ldots \mathbf{T}_{t,j}^{k_2} \ldots \mathbf{\Delta} \ldots \\
& \text{or } \mathbf{\Delta} \ldots \mathbf{T}_{t,j}^{k_1}\mathbf{\Delta} \ldots \mathbf{T}_{t,j}^{k_2} \ldots \\
\Rightarrow \quad & \mathbf{T}_{t,j}^{k_1}\mathbf{\Delta} = \begin{bmatrix} \frac{\gamma_1^{k+1}-\gamma_2^{k+1}}{\gamma_1-\gamma_2}\delta & 0 \\ \frac{\gamma_1^{k}-\gamma_2^{k}}{\gamma_1-\gamma_2}\delta & 0 \end{bmatrix}, \quad \mathbf{\Delta}\mathbf{T}_{t,j}^{k_1} = \begin{bmatrix} \frac{\gamma_1^{k+1}-\gamma_2^{k+1}}{\gamma_1-\gamma_2}\delta & -\beta\frac{\gamma_1^{k}-\gamma_2^{k}}{\gamma_1-\gamma_2}\delta \\ 0 & 0 \end{bmatrix}
\end{aligned}
$$

Two key properties:

- The left column is nonnegative, the right column is nonpositive.
- Absolute value of each element is a monotonically increasing function of δ.

Key Result

Lemma (bounding matrix product)

Given $\beta \in [0, 1)$, $\mathbf{T}_{t,j}$, *if* $\mathbf{T}_{t,j}$ *only has real eigenvalues, which is equivalent to that the discriminant of* $\mathbf{T}_{t,j}$ *satisfies that* $(1 + \beta - \eta_t \lambda_j)^2 - 4\beta \geq 0$, *it holds that*

$$\|\mathbf{T}_{t+1,j}\mathbf{T}_{t+2,j}...\mathbf{T}_{t+k,j}\| \leq \min\left(8k, \frac{8}{\sqrt{(1 + \beta - \eta_{t+k}\lambda_j)^2 - 4\beta}}\right)\rho(\mathbf{T}_{t+k,j})^k.$$

Main application: loss will not worsen too much when step size is small.

Convergence Analysis: Bias

$$\|\mathbf{T}_{T-1,j}\mathbf{T}_{T-2,j}\ldots\mathbf{T}_{1,j}\|^2 \leq \|\mathbf{T}_{T-1,j}\mathbf{T}_{T-2,j}\ldots\mathbf{T}_{\tau+1,j}\|^2 \cdot \left\|\left(\mathbf{T}'_{1,j}\right)^{k_1}\right\|^2$$

1. In the first stage: bias exponentially decays after $\tilde{\mathcal{O}}(\sqrt{\kappa})$ iterations.
2. In the remaining stages, the bias won't get much worse (at most κ times of that after the first stage)

$$\|\mathbf{T}_{t+1,j}\mathbf{T}_{t+2,j}\ldots\mathbf{T}_{t+k,j}\| \leq \min\left(8k, \frac{8}{\sqrt{(1+\beta-\eta_{t+k}\lambda_j)^2-4\beta}}\right)\rho(\mathbf{T}_{t+k,j})^k$$
$$\leq \frac{8}{\sqrt{(1+\beta-\eta_{t+k}\lambda_j)^2-4\beta}} \approx \sqrt{\kappa}.$$

Convergence Analysis: Variance

$$V = \sigma^2 \sum_{j=1}^{d} \lambda_j^2 \sum_{\tau=1}^{T-1} \eta_\tau^2 \left\| \mathbf{T}_{T-1,j} \mathbf{T}_{T-2,j} \ldots \mathbf{T}_{\tau+1,j} \right\|^2 = \sigma^2 \sum_{j=1}^{d} \sum_{\tau=1}^{T-1} V_{\tau,j}.$$

- $\eta_t \lambda_j > \left(1 - \sqrt{\beta}\right)^2 = 1/\kappa$, allows geometric decay of variance, $\mathbf{T}_{t,j}$ has complex eigenvalues.
- $\eta_t \lambda_j \in [h/(T\sqrt{\kappa}), 1/\kappa]$, allows geometric decay of variance, $\mathbf{T}_{t,j}$ has real eigenvalues.
- $\eta_t \lambda_j < h/(T\sqrt{\kappa})$, variance no longer decay, but will not worsen too much due to small step sizes.

The balance point h is around poly$(\log(T\sqrt{\kappa}))$.

Main Result

Theorem (main result)

Given a quadratic objective $f(\mathbf{w})$ and a step decay learning rate scheduler with $\beta = (1 - 1/\sqrt{\kappa})^2$, and $n_\ell \equiv T/k_\ell$ with settings that

1. *stepsize η'_ℓ: $\eta'_\ell = \frac{1}{L} \cdot \frac{1}{C^{\ell-1}}$*
2. *the stage length k_ℓ: $k_\ell = \frac{T}{\log_c(T\sqrt{\kappa})}$*
3. *The total iteration number T: $\frac{T}{\ln(2^{14}T^8)\cdot\ln(2^6T^4)\cdot\log_c(T^2)} \geq 2C\sqrt{\kappa}$,*

then such scheduler exists, and the output of the algorithm satisfies

$$\begin{aligned}\mathbb{E}[f(\mathbf{w}_T) - f(\mathbf{w}_*)] \leq & \mathbb{E}\left[f(\mathbf{w}_0) + f(\mathbf{w}_1) - 2f(\mathbf{w}_*)\right] \\ & \cdot \exp\left(14\ln 2 + 2\ln T + 2\ln\kappa - \frac{2T}{\sqrt{\kappa}\log_c\left(T\sqrt{\kappa}\right)}\right) \\ & + \frac{4096 d\sigma^2}{MT}\ln^2\left(2^6T^4\right) \cdot \log_c^2\left(T\sqrt{\kappa}\right).\end{aligned}$$

Outline

Experiments: Ridge Regressions

Table 1: Training loss statistics of ridge regression in a4a dataset over 5 runs.

Methods/Schedules	$(f(\mathbf{w}) - f(\mathbf{w}_*)) \times 10^{-2}$			
	Batch size $M = 512$	$M = 128$	$M = 32$	$M = 8$
SGD + constant η_t	2.10±0.46	1.17±0.81	1.27±0.27	0.94±0.83
SGD + step decay	2.44±0.45	0.64±0.04	0.11±0.01	**0.04±0.04**
SHB + constant η_t	0.86±0.55	0.55±0.26	1.03±0.35	0.97±0.58
SHB + step decay	**0.13±0.03**	**0.01±0.00**	**0.03±0.02**	0.06±0.05

References

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. 2018

Accelerating stochastic gradient descent for least squares regression

COLT 2018.

Rui Pan, Haishan Ye, Tong Zhang. 2018

Eigencurve: Optimal Learning Rate Schedule for SGD on Quadratic Objectives with Skewed Hessian Spectrums

ICLR 2022.