Provable Acceleration of Heavy Ball beyond Quadratics for a Class of Polyak-Łojasiewicz Functions when the Non-Convexity is Averaged-Out

ICML 2022

Speaker: Jun-Kun Wang (Yale)

Joint work with Chi-Heng Lin (Georgia Tech), Andre Wibisono (Yale), Bin Hu (UIUC)

$$\min_{w} f(w)$$

for $t = 0$ to T **do**

Given current iterate w_t, compute gradient $\nabla f(w_t)$.

Update iterate $w_{t+1} = w_t - \underbrace{\eta}_{\text{step size}} \nabla f(w_t) + \underbrace{\underbrace{\beta}_{\text{momentum parameter}} (w_t - w_{t-1})}_{\text{momentum}}$

end for

Heavy Ball (Polyak 1964)

for $t = 0$ to T **do**

Given current iterate w_t, obtain gradient $\nabla \ell(w_t)$.

Update momentum $M_t = \beta M_{t-1} + \nabla \ell(w_t)$.

Update iterate $w_{t+1} = w_t - \eta M_t$.

end for

Heavy Ball (Polyak 1964)

for $t = 0$ to T **do**

Given current iterate w_t, compute gradient $\nabla f(w_t)$.

Update iterate $w_{t+1} = w_t - \eta \nabla f(w_t) + \beta_t (w_t - w_{t-1})$.

end for

allow a non-constant $\beta_t \in [0,1]$

Heavy Ball (Polyak 1964)

(Known results) Acceleration over Gradient Descent

- **1. (strongly convex quadratics)**

$\min_{w \in R^d} \frac{1}{2} w^\top M w + b^\top w$ Condition number: $\kappa := \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$

Faster ! **HB:** $\left(1 - \Theta\left(\frac{1}{\sqrt{\kappa}}\right)\right)^T$ GD: $\left(1 - \Theta\left(\frac{1}{\kappa}\right)\right)^T$

- **2. (Over-parametrized Neural Network)**

κ_0 : condition number of the neural tangent kernel matrix at initialization

Faster ! **HB:** $\left(1 - \Theta\left(\frac{1}{\sqrt{\kappa_0}}\right)\right)^T$ GD: $\left(1 - \Theta\left(\frac{1}{\kappa_0}\right)\right)^T$

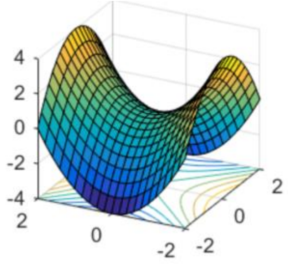
(ICML 2021) Jun-Kun Wang, Chi-Heng Lin, Jacob Abernethy

A Modular Analysis of Provable Acceleration via Polyak's Momentum: Training a Wide ReLU Network and a Deep Linear Network

- **3. (Escape Saddle Points)**

(ICLR 2020) Jun-Kun Wang, Chi-Heng Lin, and Jacob Abernethy.

Escape Saddle Points Faster with Stochastic Momentum.



HB escapes saddle points faster than GD

(Known results) Acceleration over Gradient Descent

- **1. (strongly convex quadratics)**

$$\min_{w \in R^d} \frac{1}{2} w^\top M w + b^\top w \qquad \text{Condition number: } \kappa := \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$$

Faster ! **HB:** $\left(1 - \Theta\left(\frac{1}{\sqrt{\kappa}}\right)\right)^T$ $\qquad$ ***GD***: $\left(1 - \Theta\left(\frac{1}{\kappa}\right)\right)^T$

Quadratic or close-to-quadratic

- **2. (Over-parametrized Neural Network)**

$\boldsymbol{\kappa_0}$ **: condition number of the neural tangent kernel matrix at initialization**

Faster ! **HB:** $\left(1 - \Theta\left(\frac{1}{\sqrt{\kappa_0}}\right)\right)^T$ $\qquad$ ***GD***: $\left(1 - \Theta\left(\frac{1}{\kappa_0}\right)\right)^T$

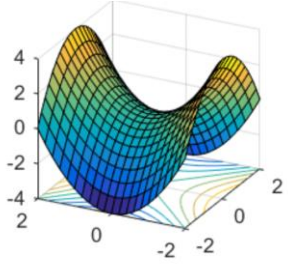
(ICML 2021) Jun-Kun Wang, Chi-Heng Lin, Jacob Abernethy

A Modular Analysis of Provable Acceleration via Polyak's Momentum: Training a Wide ReLU Network and a Deep Linear Network

- **3. (Escape Saddle Points)**

(ICLR 2020) Jun-Kun Wang, Chi-Heng Lin, and Jacob Abernethy.

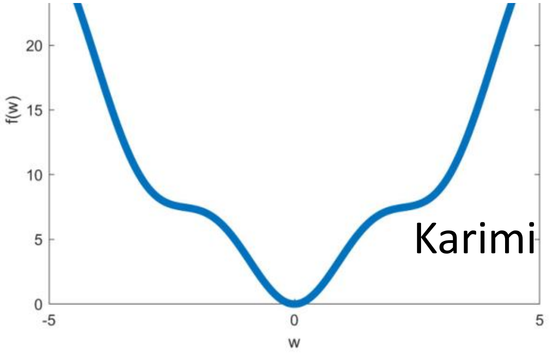
Escape Saddle Points Faster with Stochastic Momentum.



HB escapes saddle points faster than GD

Provable Acceleration of HB beyond Quadratics

$f(w) = w^2 + 4sin^2(w)$



Karimi et al.

- **Average Hessian:** w_* a global minimizer of $f(\cdot)$

$$H_f(w) := \int_0^1 \nabla^2 f(\theta w + (1-\theta)w_*)d\theta.$$

The non-convexity of $f(\cdot)$ between w and w_* is averaged-out with parameter λ_* when the smallest eigenvalue of the average Hessian satisfies $\lambda_{min}(H_f) \geq \lambda_* > 0$.

λ_*Averaged-out implies λ_*PL!

- **Polyak-Łojasiewicz (PL) condition:**

$$\|\nabla f(w)\|^2 \geq 2\mu(f(w) - \min_w f(w))$$

PL holds at point w

Our main result (informal version)

Suppose $f(\cdot)$ is twice differentiable, satisfies μ-PL, has L-Lipschitz gradient and Hessian. Apply HB to solve $\min_w f(w)$. Assume that the averaged-out condition holds for all t, i. e., $\lambda_{\min}(H_f(w_t)) > 0$.

Then, there exists a time $t_0 = \tilde{\Theta}(\frac{L}{\mu})$ such that for all $T > t_0$, the iterate w_T of HB satisfies:

$$\|w_T - w_*\| = O\left(\prod_{t=t_0}^{T-1}\left(1 - \Theta\left(\frac{1}{\sqrt{\kappa_t}}\right)\right)\right)\|w_{t_0} - w_*\|$$

where $\kappa_t := \frac{L}{\lambda_{\min}(H_f(w_t))}$ is the condition number of the average Hessian at w_t.

$$\begin{bmatrix} w_{t+1} - w_* \\ w_t - w_* \end{bmatrix} = \underbrace{\begin{bmatrix} I_d - \eta H_t + \beta_t I_d & -\beta_t I_d \\ I_d & 0_d \end{bmatrix}}_{:= A_t} \begin{bmatrix} w_t - w_* \\ w_{t-1} - w_* \end{bmatrix}$$

Goal: Showing $\left\| \begin{bmatrix} w_{t+1} - w_* \\ w_t - w_* \end{bmatrix} \right\|_2$ decays at an accelerated rate

HB dynamics

Fact: Spectral norm $\|A_t\|_2 \geq 1$ for any $\eta \leq \frac{1}{L}$ and $\beta_t \in [0,1]$

$$\begin{bmatrix} w_{t+1} - w_* \\ w_t - w_* \end{bmatrix} = \underbrace{\begin{bmatrix} I_d - \eta H_t + \beta_t I_d & -\beta_t I_d \\ I_d & 0_d \end{bmatrix}}_{:=A_t} \begin{bmatrix} w_t - w_* \\ w_{t-1} - w_* \end{bmatrix}$$

HB dynamics

Goal: Showing $\left\| \begin{bmatrix} w_{t+1} - w_* \\ w_t - w_* \end{bmatrix} \right\|_2$ decays at an accelerated rate

Fact: Spectral norm $\|A_t\|_2 \geq 1$ for any $\eta \leq \frac{1}{L}$ and $\beta_t \in [0,1]$

If the momentum parameter satisfies $1 \geq \beta_t > \left(1 - \sqrt{\eta \lambda_{t,i}}\right)^2$
then A_t has a decomposition in the complex field: $A_t = P_t D_t P_t^{-1}$, where $\|D_t\|_2 = \sqrt{\beta_t}$

$$\begin{aligned}
\begin{bmatrix} w_{T+1} - w_* \\ w_T - w_* \end{bmatrix} &= A_T A_{T-1} A_{T-2} \cdots A_{t_0} \begin{bmatrix} w_{t_0} - w_* \\ w_{t_0-1} - w_* \end{bmatrix} \\
&= \left(P_T D_T P_T^{-1}\right)\left(P_{T-1} D_{T-1} P_{T-1}^{-1}\right)\left(P_{T-2} D_{T-2} P_{T-2}^{-1}\right) \\
&\quad \cdots \left(P_{t_0} D_{t_0} P_{t_0}^{-1}\right) \begin{bmatrix} w_{t_0} - w_* \\ w_{t_0-1} - w_* \end{bmatrix}, \\
&= P_T \underbrace{\left(D_T P_T^{-1} P_{T-1}\right)}_{:=\Psi_T} \underbrace{\left(D_{T-1} P_{T-1}^{-1} P_{T-2}\right)}_{:=\Psi_{T-1}} \cdots \\
&\quad \cdot \underbrace{\left(D_{T-2} P_{T-2}^{-1} P_{T-3}\right)}_{:=\Psi_{T-2}} \cdots \underbrace{\left(D_{t_0+1} P_{t_0+1}^{-1} P_{t_0}\right)}_{:=\Psi_{t_0}} D_{t_0} P_{t_0}^{-1} \begin{bmatrix} w_{t_0} - w_* \\ w_{t_0-1} - w_* \end{bmatrix}.
\end{aligned}$$

$$\left\| \begin{bmatrix} w_{t+1} - w_* \\ w_t - w_* \end{bmatrix} \right\| = O\left(\textstyle\prod_{t=t_0+1}^{T} \|\Psi_t\|_2\right) \left\| \begin{bmatrix} w_{t_0} - w_* \\ w_{t_0-1} - w_* \end{bmatrix} \right\|$$

$$\begin{bmatrix} w_{t+1} - w_* \\ w_t - w_* \end{bmatrix} = \underbrace{\begin{bmatrix} I_d - \eta H_t + \beta_t I_d & -\beta_t I_d \\ I_d & 0_d \end{bmatrix}}_{:=A_t} \begin{bmatrix} w_t - w_* \\ w_{t-1} - w_* \end{bmatrix}$$

HB dynamics

Goal:
Showing $\left\| \begin{bmatrix} w_{t+1} - w_* \\ w_t - w_* \end{bmatrix} \right\|_2$
decays at an accelerated rate

Fact: Spectral norm $\|A_t\|_2 \geq 1$ for any $\eta \leq \frac{1}{L}$ and $\beta_t \in [0,1]$

If the momentum parameter satisfies $1 \geq \beta_t > \left(1 - \sqrt{\eta \lambda_{t,i}}\right)^2$
then A_t has a decomposition in the complex field: $A_t = P_t D_t P_t^{-1}$, where $\|D_t\|_2 = \sqrt{\beta_t}$

$$\begin{aligned}
\begin{bmatrix} w_{T+1} - w_* \\ w_T - w_* \end{bmatrix} &= A_T A_{T-1} A_{T-2} \cdots A_{t_0} \begin{bmatrix} w_{t_0} - w_* \\ w_{t_0-1} - w_* \end{bmatrix} \\
&= \left(P_T D_T P_T^{-1}\right)\left(P_{T-1} D_{T-1} P_{T-1}^{-1}\right)\left(P_{T-2} D_{T-2} P_{T-2}^{-1}\right) \\
&\quad \cdots \left(P_{t_0} D_{t_0} P_{t_0}^{-1}\right) \begin{bmatrix} w_{t_0} - w_* \\ w_{t_0-1} - w_* \end{bmatrix}, \\
&= P_T \underbrace{\left(D_T P_T^{-1} P_{T-1}\right)}_{:=\Psi_T} \underbrace{\left(D_{T-1} P_{T-1}^{-1} P_{T-2}\right)}_{:=\Psi_{T-1}} \cdots \\
&\quad \cdot \underbrace{\left(D_{T-2} P_{T-2}^{-1} P_{T-3}\right)}_{:=\Psi_{T-2}} \cdots \underbrace{\left(D_{t_0+1} P_{t_0+1}^{-1} P_{t_0}\right)}_{:=\Psi_{t_0}} D_{t_0} P_{t_0}^{-1} \begin{bmatrix} w_{t_0} - w_* \\ w_{t_0-1} - w_* \end{bmatrix}.
\end{aligned}$$

$$\left\| \begin{bmatrix} w_{t+1} - w_* \\ w_t - w_* \end{bmatrix} \right\| = O\left(\prod_{t=t_0+1}^{T} \|\Psi_t\|_2\right) \left\| \begin{bmatrix} w_{t_0} - w_* \\ w_{t_0-1} - w_* \end{bmatrix} \right\|$$

come by our poster!