SGD with memory: fundamental properties and stochastic acceleration[1]

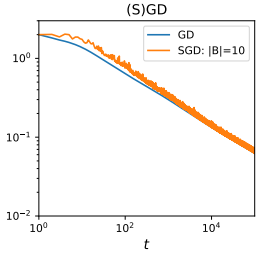
Dmitry Yarotsky
Maksim Velikanov

ICOMP 2024, Innopolis
October 11, 2024

[1] arXiv:2410.04228

A motivating example: (S)GD for MNIST

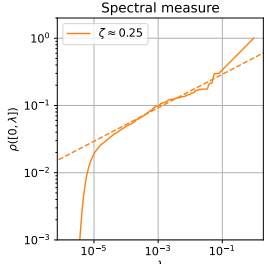
SGD with random batches B_t: $\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla L_{B_t}(\mathbf{w}_t)$
GD: limit of SGD as $|B| \to \infty$



MNIST NTK-classifier with MSE loss: loss falls off as **power law**:

$$L(\mathbf{w}_t) \sim C_L t^{-\zeta}, \quad \zeta \approx 0.25$$

This can be deduced from **spectral power law**:



$$L(\mathbf{w}) \approx \frac{1}{2}\sum_{k=1}^{\infty} \lambda_k \langle \Delta\mathbf{w}, \mathbf{e}_k \rangle^2 = \frac{1}{2}\langle \Delta\mathbf{w}, H\Delta\mathbf{w} \rangle, \quad \Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_*$$

$$\sum_{k:\lambda_k<\lambda} \lambda_k \langle \mathbf{w}_*, \mathbf{e}_k \rangle^2 \sim Q\lambda^{\zeta}, \quad \lambda \to 0 \quad (\text{“source condition”})$$

Accelerating power-law GD: Heavy Ball with Jacobi schedules

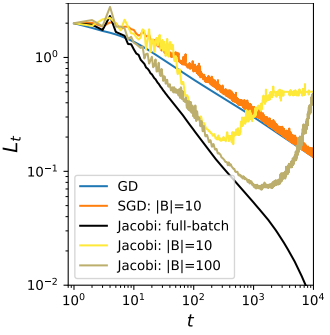
Jacobi-scheduled HB (JHB):

$$\begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_t \\ \mathbf{u}_{t+1} \end{pmatrix} = \begin{pmatrix} -\alpha & \beta_t \\ -\alpha & \beta_t \end{pmatrix} \begin{pmatrix} \nabla L(\mathbf{w}_t) \\ \mathbf{u}_t \end{pmatrix}, \quad \beta_t \sim 1 - \frac{const}{t}$$

Without sampling noise (i.e., when $|B| = \infty$), JHB **doubles** the convergence exponent:

$$L(\mathbf{w}_t) = O(t^{-2\zeta})$$

But with sampling noise ($|B| < \infty$) JHB works only for a limited number of steps, then **diverges**



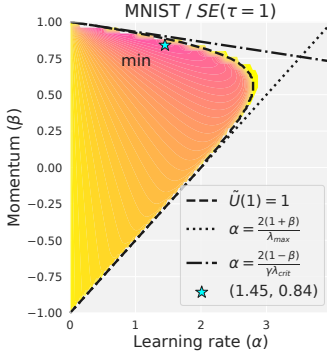
Why stochastic Jacobi Heavy Ball fails?

Jacobi Heavy Ball gradually increases **effective learning rate**:

$$\alpha_{\text{eff}} = \frac{\alpha}{1-\beta} = \alpha + \alpha\beta + \alpha\beta^2 + \dots$$

$$\alpha_{\text{eff}} \to \infty \text{ as } \beta \to 1$$

But under sampling noise, only a limited range of α_{eff} is stable

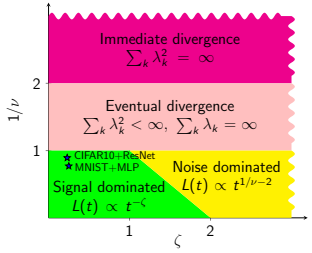


Our approach-1

So, can we still accelerate SGD?

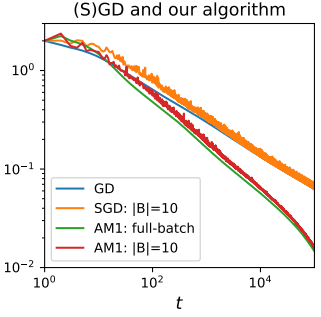
Yes! Our approach:

- Consider general **SGD with memory size** M
- They have **two equivalent forms** (matrix and sequential)
- Perform **propagator expansion** of the loss
- All **stationary** memory-M algorithms have the same **phase diagram** with **divergent**, **signal-** and **noise-dominated phases**



Our approach-2

- Find the **leading term** of loss $L_t \sim C_L t^{-\zeta}$ (with explicit constant C_L!)
- In the signal dominated phase, C_L depends on the algorithm only through $|B|$ and generalized **effective learning rate** α_{eff}
- Acceleration: increase $\alpha_{\text{eff}} \to \infty$ while preserving stability
- Accelerated regime exists for $M = 1$! But we need to go beyond Heavy Ball
- Accelerate $L_t = O(t^{-\zeta})$ to $L_t = O(t^{-\zeta(1+\overline{\alpha})}), \overline{\alpha} > 0$, by adiabatically adjusting parameters with time (heuristic)



SGD with memory M

$$\begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_t \\ \mathbf{u}_{t+1} \end{pmatrix} = \begin{pmatrix} -\alpha_t & \mathbf{b}_t^T \\ \mathbf{c}_t & D_t \end{pmatrix} \begin{pmatrix} \nabla L_{B_t}(\mathbf{w}_t + \mathbf{a}_t^T \mathbf{u}_t) \\ \mathbf{u}_t \end{pmatrix}, \quad t = 0, 1, 2, \ldots$$

Here

- $\mathbf{w}_t \in \mathcal{H}$: the main current state vector
- $\mathbf{u}_t \in \mathbb{R}^M \otimes \mathcal{H}$: a set of M "generalized momentum vectors"
- $\alpha_t \in \mathbb{R}, \mathbf{a}_t, \mathbf{b}_t, \mathbf{c}_t \in \mathbb{R}^M, D_t \in \mathbb{R}^{M \times M}$: parameters of the algorithm

$M = 0$: plain SGD
$M = 1$: includes Heavy Ball, averaging

Equivalence of stationary matrix and sequential representations

Stationary (t-independent) memory-M GD:

$$\begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_t \\ \mathbf{u}_{t+1} \end{pmatrix} = \begin{pmatrix} -\alpha & \mathbf{b}^T \\ \mathbf{c} & D \end{pmatrix} \begin{pmatrix} \nabla L(\mathbf{w}_t + \mathbf{a}^T \mathbf{u}_t) \\ \mathbf{u}_t \end{pmatrix} \quad \text{(matrix form)}$$

$$\mathbf{w}_{t+M+1} = \sum_{m=0}^{M} p_m \mathbf{w}_t + \sum_{m=0}^{M} q_m \nabla L(\mathbf{w}_t) \text{ with } \sum_{m=0}^{M} p_m = 1 \quad \text{(sequential form)}$$

Theorem (informal). Matrix and sequential forms of stationary memory-M GD are equivalent for quadratic losses L, with $(p_m)_{m=0}^M, (q_m)_{m=0}^M$ connected to $\alpha, \mathbf{a}, \mathbf{b}, \mathbf{c}, D$ by

$$\sum_{m=0}^{M} p_m x^m = x^{M+1} - (x-1)\det(x - D),$$

$$\sum_{m=0}^{M} q_m x^m = \det \begin{pmatrix} \mathbf{a}^T \mathbf{c} - \alpha & \mathbf{a}^T(1-D) - \mathbf{b}^T \\ \mathbf{c} & x - D \end{pmatrix}.$$

Propagator expansion of the loss

Under a "spectrally expressible" approximation, for stationary algorithms with $\mathbf{a} = 0$ and quadratic losses

$$L_t \equiv \mathbb{E}L(\mathbf{w}_t) = \frac{1}{2}\Big(V_{t+1} + \sum_{m=1}^{t} \sum_{0<t_1<\ldots<t_m<t+1} U_{t+1-t_m} U_{t_m-t_{m-1}} U_{t_{m-1}-t_{m-2}} \cdots U_{t_2-t_1} V_{t_1}\Big)$$

with

- **Signal propagators**

$$V_t = \sum_{\lambda_k \in \mathrm{spec}(H)} \lambda_k \langle \mathbf{w}_*, \mathbf{e}_k \rangle^2 \left| \begin{pmatrix} 1 & 0 \end{pmatrix} S_{\lambda_k}^{t-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right|^2$$

- **Noise propagators**

$$U_t = \frac{1}{|B|} \sum_{\lambda_k \in \mathrm{spec}(H)} \lambda_k^2 \left| \begin{pmatrix} 1 & 0 \end{pmatrix} S_{\lambda_k}^{t-1} \begin{pmatrix} -\alpha \\ \mathbf{c} \end{pmatrix} \right|^2$$

Here S_λ determine the noiseless propagation in spectral subspace λ:

$$S_\lambda = \begin{pmatrix} 1 & \mathbf{b}^T \\ 0 & D \end{pmatrix} + \lambda \begin{pmatrix} -\alpha \\ \mathbf{c} \end{pmatrix} (1, \mathbf{a}^T)$$

GD: $U_t \equiv 0$ and $L_t = \frac{1}{2} V_{t+1}$

Convergence of SGD = Convergence of GD & $U_\Sigma < 1$

Let $U_\Sigma = \sum_{t \geq 1} U_t$

Theorem.

1. **[Convergence]** Suppose that $U_\Sigma < 1$. If V_t is bounded (resp., $V_t \to 0$), then also L_t is bounded (resp., $L_t \to 0$).
2. **[Divergence]** If $U_\Sigma > 1$ and $V_t > 0$ for at least one t, then $\sup_{t=1,2,\ldots} L_t = \infty$.

Signal- and noise-dominated phases

Theorem. Assume $V_t = C_V t^{-\xi_V}(1 + o(1))$, $U_t = C_U t^{-\xi_U}(1 + o(1))$, and $U_\Sigma < 1$.

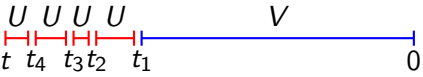
1. **[Signal-dominated regime]** If $\xi_U > \xi_V$, then

$$L_t = \frac{C_V}{2(1 - U_\Sigma)} t^{-\xi_V}(1 + o(1)).$$

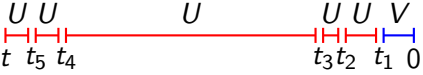
2. **[Noise-dominated regime]** If $\xi_V > \xi_U$, then

$$L_t = \frac{V_\Sigma C_U}{2(1 - U_\Sigma)^2} t^{-\xi_U}(1 + o(1)).$$

Signal-dominated



Noise-dominated



Stability and generalized effective learning rate

$S_{\lambda=0} = \left(\begin{smallmatrix} 1 & \mathbf{b}^T \\ 0 & D \end{smallmatrix}\right)$ has eigenvalue 1. Assume it's the largest eigenvalue.

For stability, we need the respective eigenvalue μ_λ of S_λ to decrease as λ increases from 0

Theorem.

$$\mu_\lambda = 1 - \alpha_{\text{eff}}\lambda + O(\lambda^2), \quad \lambda \searrow 0$$

with the **effective learning rate**

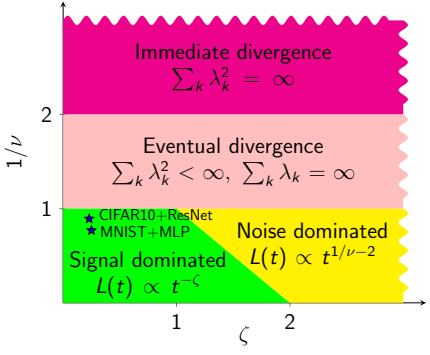
$$\alpha_{\text{eff}} = \alpha - \mathbf{b}^T(1-D)^{-1}\mathbf{c} = \frac{\sum_{m=0}^{M} q_m}{\sum_{m=0}^{M} m p_m - M - 1},$$

Power law phase diagram

Assume
$$\lambda_k = \Lambda k^{-\nu}(1+o(1)), \quad k \to \infty \quad \text{(eigenvalue decay)}$$
$$\sum_{k:\lambda_k<\lambda} \lambda_k \langle \mathbf{w}_*, \mathbf{e}_k\rangle^2 = Q\lambda^{\zeta}, \quad \lambda \to 0 \quad \text{(source condition)}$$

Theorem.

1. **[Divergent phase]** If $\nu < 1$, then $\sup_t L_t = \infty$
2. **[Signal-dominated phase]** If $\xi_U > \xi_V$, then
$$L_t = \frac{\alpha_{\text{eff}}^{-\zeta}}{1-U_\Sigma} Q\Gamma(\zeta+1)2^{-\zeta-1}(1+o(1))t^{-\zeta}$$
3. **[Noise-dominated phase]** If $\xi_V > \xi_U$, then
$$L_t = \frac{\alpha_{\text{eff}}^{1/\nu} V_\Sigma}{|B|(1-U_\Sigma)^2} \frac{\Lambda^{1/\nu}\Gamma(2-1/\nu)}{\nu 2^{3-1/\nu}}(1+o(1))t^{\frac{1}{\nu}-2}.$$



In signal-dominated regime: accelerate = increase α_{eff} while keeping $U_\Sigma < 1$

Memory $M = 1$: stationary algorithms

General stationary memory-1 SGD:

$$\begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_t \\ \mathbf{u}_{t+1} \end{pmatrix} = \begin{pmatrix} -\alpha & b \\ c & D \end{pmatrix} \begin{pmatrix} \nabla L_{B_t}(\mathbf{w}_t + a\mathbf{u}_t) \\ \mathbf{u}_t \end{pmatrix}$$

Corresponding sequential stationary memory-1 GD:

$$\mathbf{w}_{t+2} = p_0\mathbf{w}_t + p_1\mathbf{w}_{t+1} + q_0\nabla L(\mathbf{w}_t) + q_1\nabla L(\mathbf{w}_{t+1}), \quad p_0 + p_1 = 1$$

Can be characterized by the triplet $(\delta, \alpha_{\text{eff}}, q_0)$, where $\delta = 2 - p_1$ and $\alpha_{\text{eff}} = -\frac{q_0+q_1}{\delta}$

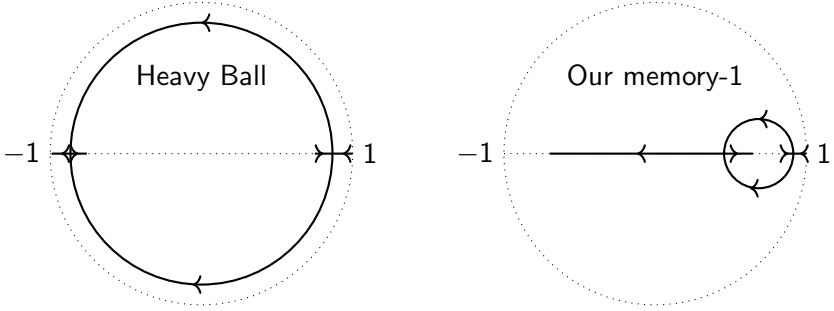
Heavy Ball: $q_0 = 0$

Theorem.

1. If $\text{Tr}(H) < \infty$, then α_{eff} can be made arbitrarily large while keeping $U_\Sigma < 1$
2. In the particular case of power law $\lambda_k \sim \Lambda k^{-\nu}, \nu < 1$, this can be done by choosing $\alpha_{\text{eff}} = \delta^{-h}$, $q_0 = \delta^g$ with any $0 \le g < 1$ and $h \le h_{\max} = (1 - \frac{1}{\nu})(1 - g)$.

Accelerated Heavy Ball vs. our memory-1

Eigenvalues of the transfer matrix S_λ as λ increases from 0 to 1:



Non-stationary accelerated schedule (heuristic)

Idea: gradually adjust $(\delta, \alpha_{\text{eff}}, q_0)$ with t within the acceleration region

Under an approximation of commuting $S_{\lambda,t}$, the optimal schedule is

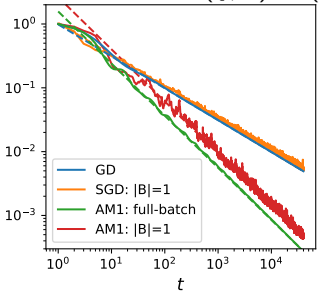
$$(\delta, \alpha_{\text{eff}}, q_0)_t \sim (t^{-1}, t^{1-\frac{1}{\nu}}, const > 0)$$

giving

$$L_t \sim t^{-\zeta(2-\frac{1}{\nu})}$$

Confirmation by experiment:

Gaussian data with $(\zeta, \nu) = (0.5, 3)$



MNIST