Universal Average-Case Optimality of Polyak Momentum

Damien Scieur
(Samsung SAIT AI Lab, Montréal)

SAMSUNG
Advanced Institute of Technology AI Lab Montreal

Fabian Pedregosa
(Google Research)

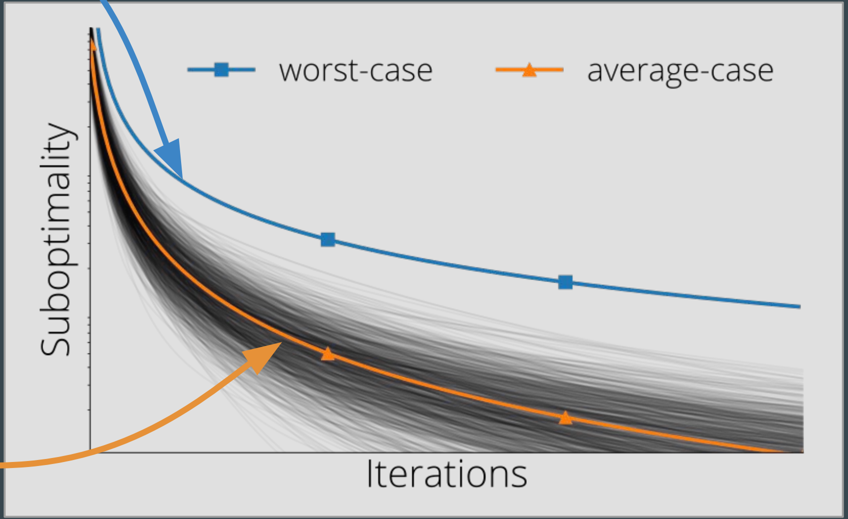
Google Research

Worst Case V.S. Average Case

Worst case

✓ Complexity bounds for any input.

✗ Not representative of typical runtime.



Average case

✓ Representative of the typical behaviour.

??? Complexity averaged over all problem instances.

Optimal Average Case Methods

Best method: minimizes P_t in

$$\overbrace{R^2}^{\text{initialization}} \int_{\mathbb{R}} \underbrace{P_t^2}_{\text{algorithm}} \overbrace{\mathrm{d}\mu}^{\text{problem}}$$

Theorem Pedregosa, Scieur (ICML 2020)

Any optimal average-case method has the form

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + (1 - a_t)(\boldsymbol{x}_{t-2} - \boldsymbol{x}_{t-1}) + b_t \nabla f(\boldsymbol{x}_{t-1})$$

Momentum Gradient step-size

Optimal Average Case Methods: Applications

- Neural networks share similar training dynamics of a quadratic problem.
  (Jacot et al. 2018, Novak et al. 2018, Arora et al. 2019, Chizat and Bach, 2019)

- Design of accelerated gossip algorithms: optimal method w.r.t. Jacobi measure
  Berthier, Bach, Gaillard, (arxiv 1805.08531).

- Random matrix sketching for solving least-squares
  Lacotte, Pilanci (ICML 2020)

- Possible to design optimal methods for any distribution
  Pedregosa, Scieur (ICML 2020)

Noticed some regularity when #iterations goes to infinity

Conjecture

All average-case optimal methods converge to Polyak momentum

(in #iterations, whatever the expected spectral density)

Asymptotic Rate of Optimal Methods (Scieur & Pedregosa, ICML 2020)

Assume we use an **average-case optimal method** w.r.t. the density function μ, **strictly positive** on the interval $[\ell, L]$. Then, *for all such densities* μ,

$$\lim_{t \to \infty} \sqrt[t]{\mathbb{E}\left[\frac{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2}{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}\right]} = \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^2 .$$

Expectation over problem instances

Relative rate of convergence

Rate of Polyak Momentum

Asymptotic Recurrence (Scieur & Pedregosa, ICML 2020)

Assume we use an **average-case optimal method** w.r.t. the density function μ, **strictly positive** on the interval $[\ell, L]$. Then, *for all such densities* μ,

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + (1 - a_t)(\boldsymbol{x}_{t-2} - \boldsymbol{x}_{t-1}) + b_t \nabla f(\boldsymbol{x}_{t-1})$$

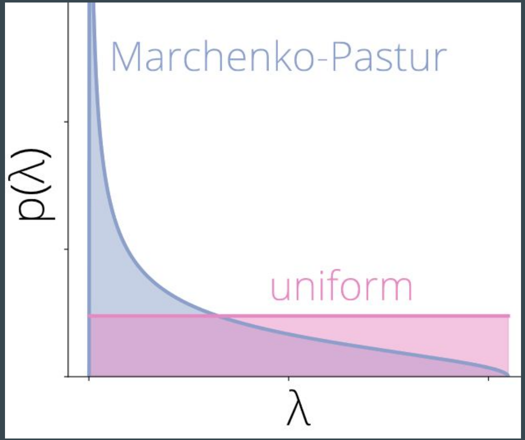
$$\lim_{t \to \infty} (1 - a_t) = -\left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^2$$

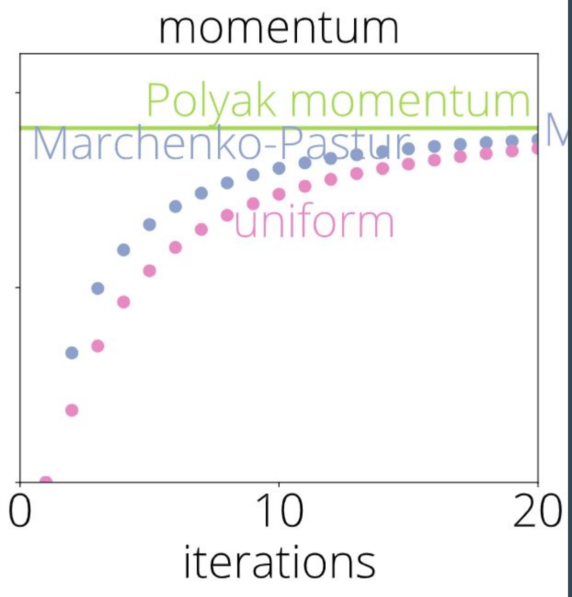
Polyak's optimal momentum

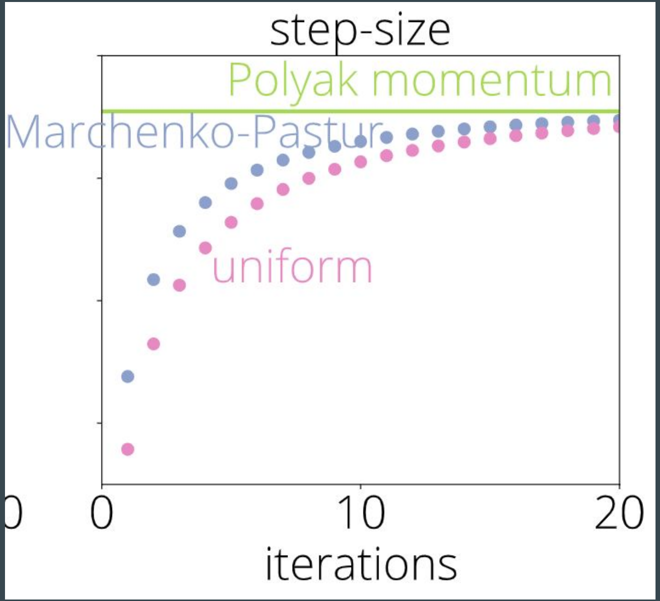
$$\lim_{t \to \infty} b_t = -\left(\frac{2}{\sqrt{L} + \sqrt{\ell}}\right)^2$$

Polyak's optimal step-size

Numerical evidences







Take-home message

Polyak momentum is *provably* always a good choice.

- Easier to design than optimal method
- Strong theory explaining its good empirical performance
- Possible loss: constant number of iterations