# Super-acceleration with cyclical step-sizes

Baptiste Goujaud, Damien Scieur, Aymeric Dieuleveut, Adrien Taylor, Fabian Pedregosa





Google Research All-Russian seminar on optimization, 2022-03-16

Preprint: https://arxiv.org/pdf/2106.09687.pdf

# HeavyBall

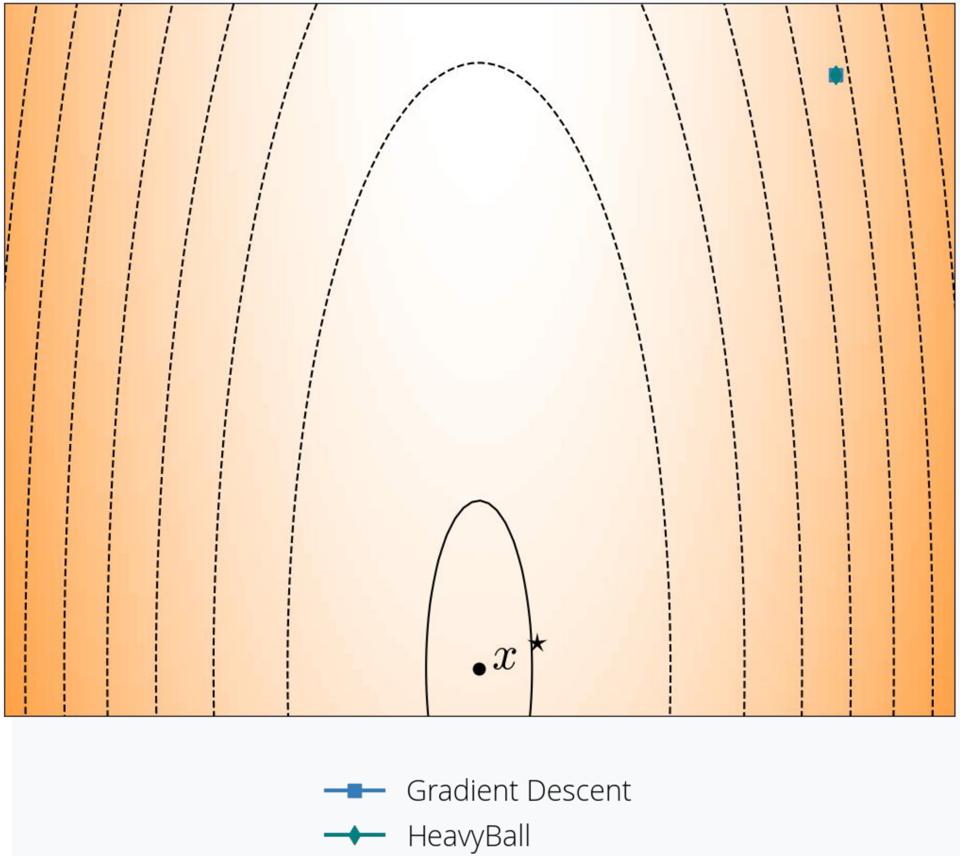
**aka gradient descent with momentum**

Two parameters; step-size **h** > 0 and momentum **m** ∈ (0, 1)

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + m(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) - h\nabla f(\boldsymbol{x}_t)$$

Optimal among gradient-based methods on quadratics.

Stochastic variant popular in deep learning.




Google Research

# Cyclical HeavyBall

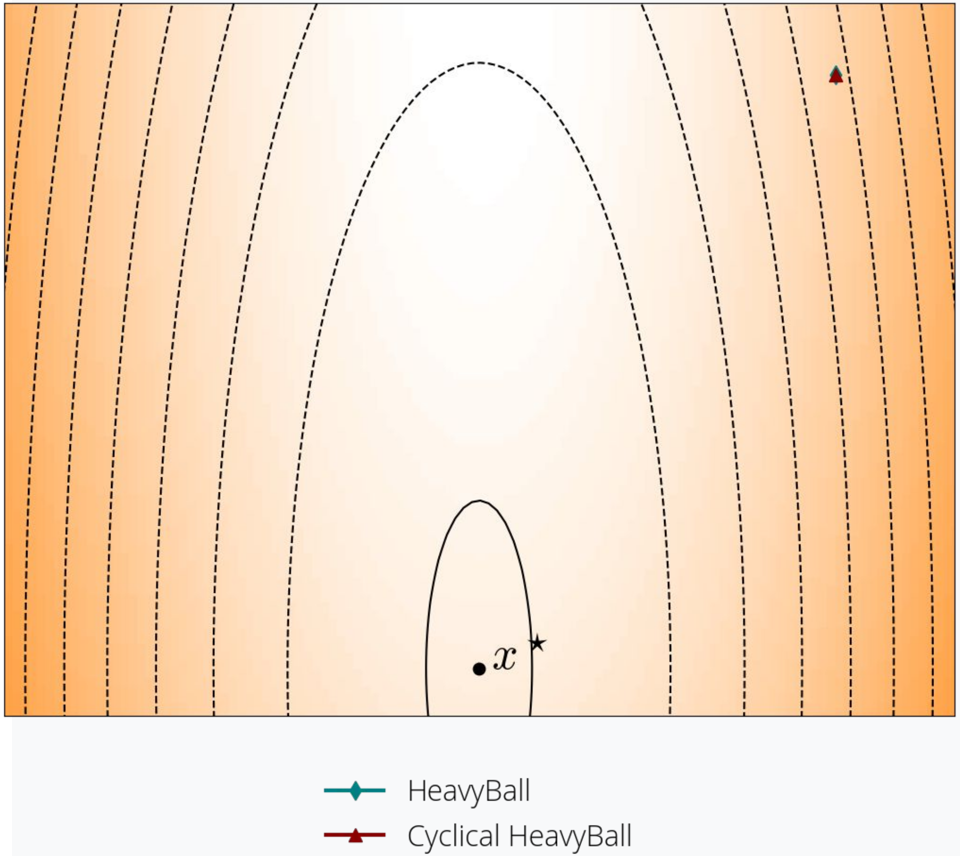
Alternates between two step-sizes $h_0$ and $h_1$

$$\text{Set } h_t = h_0 \text{ if } t \text{ is odd and } h_t = h_1 \text{ otherwise}$$
$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - h_t \nabla f(\boldsymbol{x}_t) + m(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})$$

Reported faster convergence (Loshchilov and Hutter, 2017; Smith, 2017)

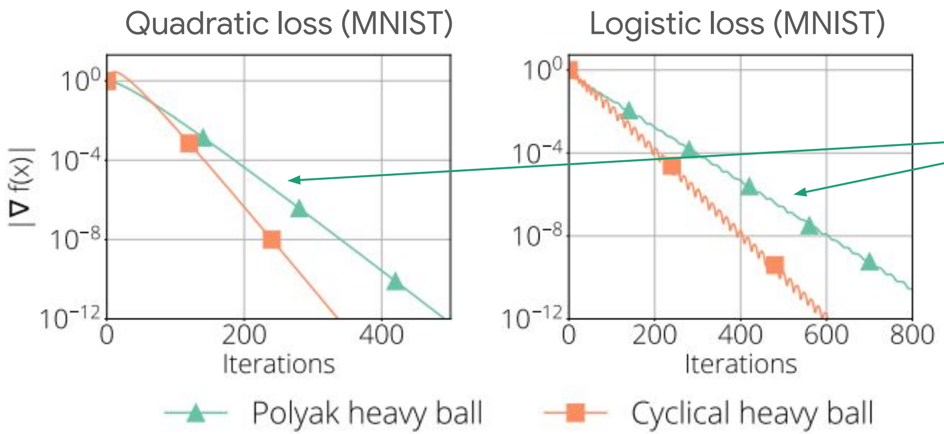
Pervasive (TF, PyTorch, optax, etc.)

No analysis that explains why/when it works.




Google Research

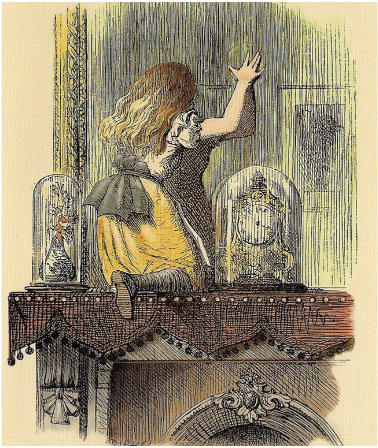
# Benchmarks

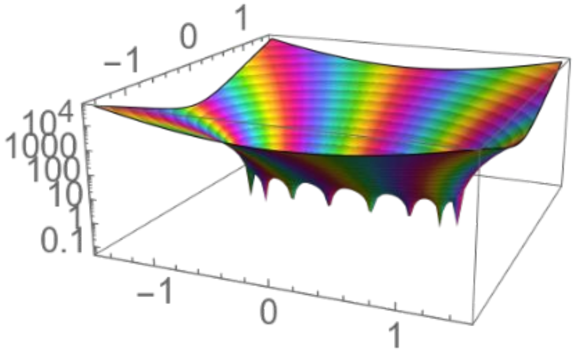


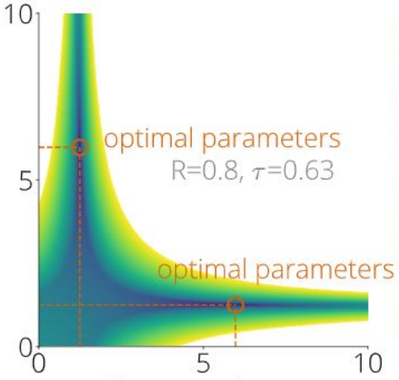

Slope = $\dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$

What is the slope of cyclical heavy ball?
What are the optimal parameters?

Today's topic






Optimization and Polynomials

Cyclical HeavyBall

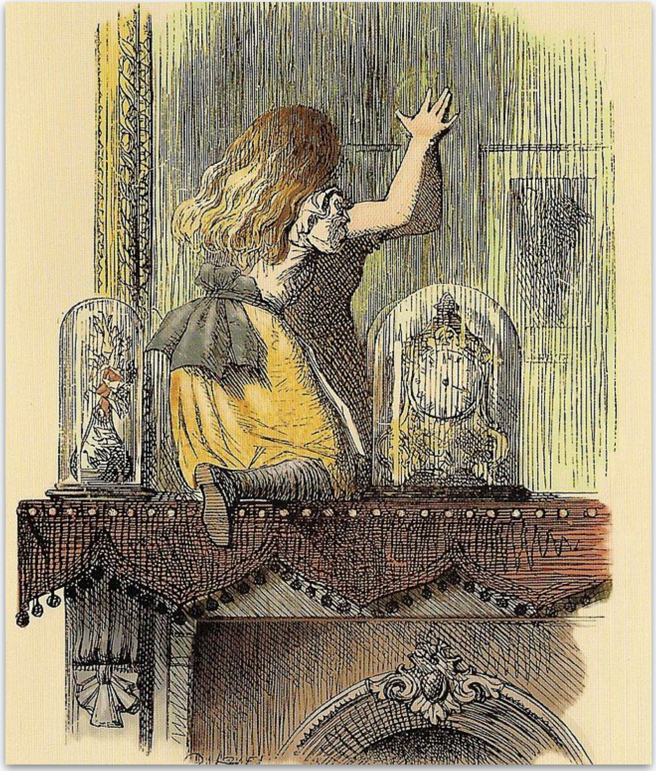
Simulations & Open problems

Google Research

# Polynomials and optimization

Some problems can be posed in the space of polynomials.

Exploited in early numerical analysis [Hestenes and Stiefel (1952), Rutishauser (1959)]



Google Research

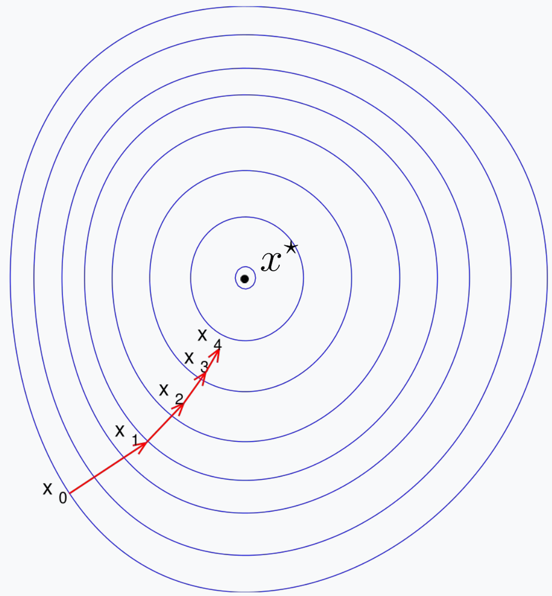
# Polynomials and Optimization

Consider Gradient Descent on

$$f(x) = \frac{1}{2}(x - x^{\star})H(x - x_{\star})$$

Then at iteration *t* we have

$$\begin{aligned} x_{t+1} - x_{\star} &= x_t - x_{\star} - \gamma H(x_t - x_{\star}) \\ &= (I - \gamma H)(x_t - x_{\star}) \\ &= \dots \\ &= \underbrace{(I - \gamma H)^{t+1}}_{\text{Polynomial in } H}(x_0 - x_{\star}) \end{aligned}$$



# Real-valued polynomials

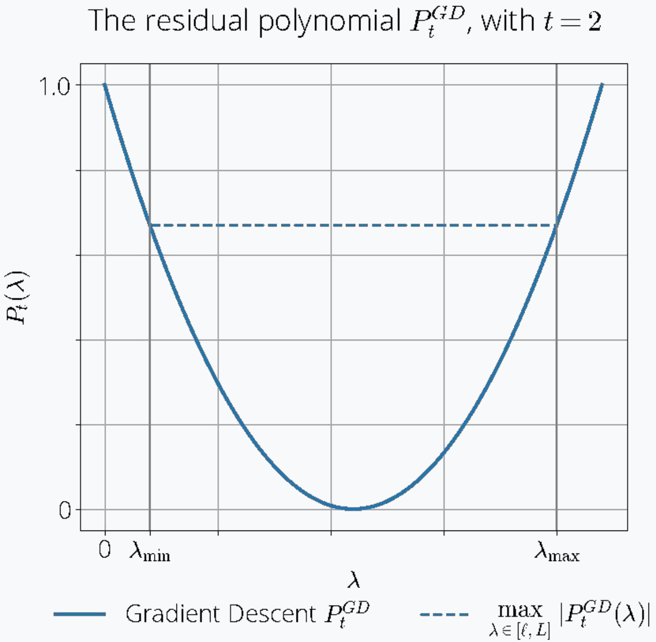
Taking norms on the previous expression

Cauchy-Schwarz

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|_2 \leq \|(\boldsymbol{I} - \tfrac{2}{L+\mu}\boldsymbol{H})^{t+1}\|_2 \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2$$

Matrix 2-norm

$$\leq \underbrace{\max_{\lambda \in [\mu, L]} |(1 - \tfrac{2}{L+\mu}\lambda)^{t+1}|}_{\text{Convergence rate } \left(\frac{L-\mu}{L+\mu}\right)^t} \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2$$




⚠ L, $\mu$ = largest and smallest eigenvalue of H

# Gradient-based Methods and Polynomials

**Corollary (Convergence rate)** *Let $\mu$ and $L$ be the smallest and largest eigenvalue of $\boldsymbol{H}$ respectively. Then for any gradient-based method with residual polynomial $P_t$, we have*

$$\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \leq \underbrace{\max_{\lambda \in [\mu, L]}}_{\text{conditioning}} \underbrace{|P_t(\lambda)|}_{\text{algorithm}} \underbrace{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|}_{\text{initialization}} . \tag{17}$$

- Problem difficulty enters through $[\mu, L]$, interval that contains Hessian eigenvalues.
- Algorithm enters through polynomial $P_t$. This polynomial verifies $P_t(0)=1$

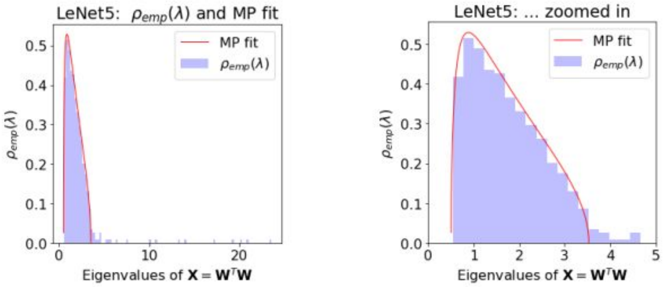
Google Research

# Parenthesis: average-case analysis

Same technique can be used to derive average-case analysis

$$\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 = \overbrace{R^2}^{\text{initialization}} \int \underbrace{P_t^2}_{\text{algorithm}} \overbrace{\mathrm{d}\mu}^{\text{problem}} .$$

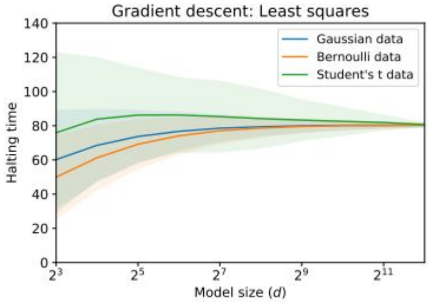
$\mathrm{d}\mu$ = density of Hessian eigenvalues




[Martin and Mahoney, 2018]

Average-case acceleration [P. and Scieur, 2020]

Concentration [Paquette et al. 2022]




Google Research

# HeavyBall

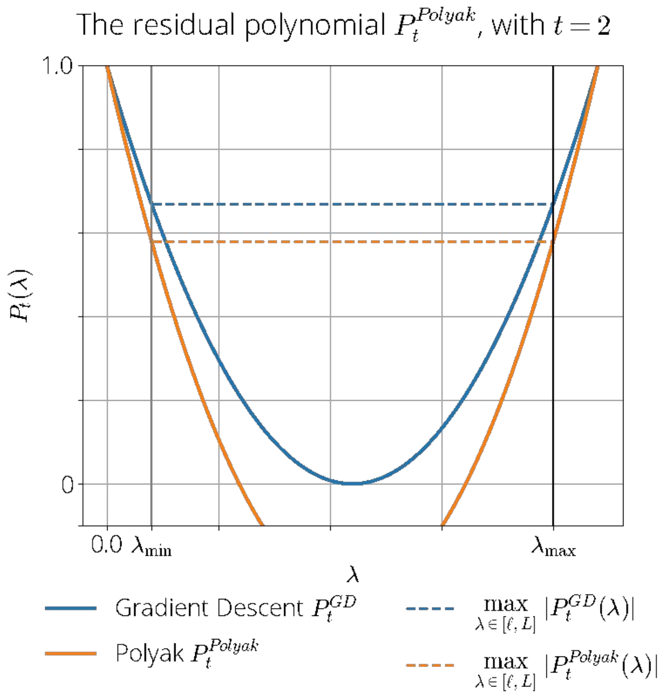
The HeavyBall update

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + m(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) - h\nabla f(\boldsymbol{x}_t)$$

Gives the residual polynomial

$$P_t(\lambda) = m^{t/2}\left(\frac{2m}{1+m}\underbrace{T_t(\sigma(\lambda))}_{\text{Chebyshev 1st kind}} - \frac{m-1}{1+m}\underbrace{U_t(\sigma(\lambda))}_{\text{Chebyshev 2nd kind}}\right)$$

with $\sigma(\lambda) = \frac{1}{2\sqrt{m}}(1 + m - h\,\lambda)$




Google Research

# The two faces of Chebyshev polynomials

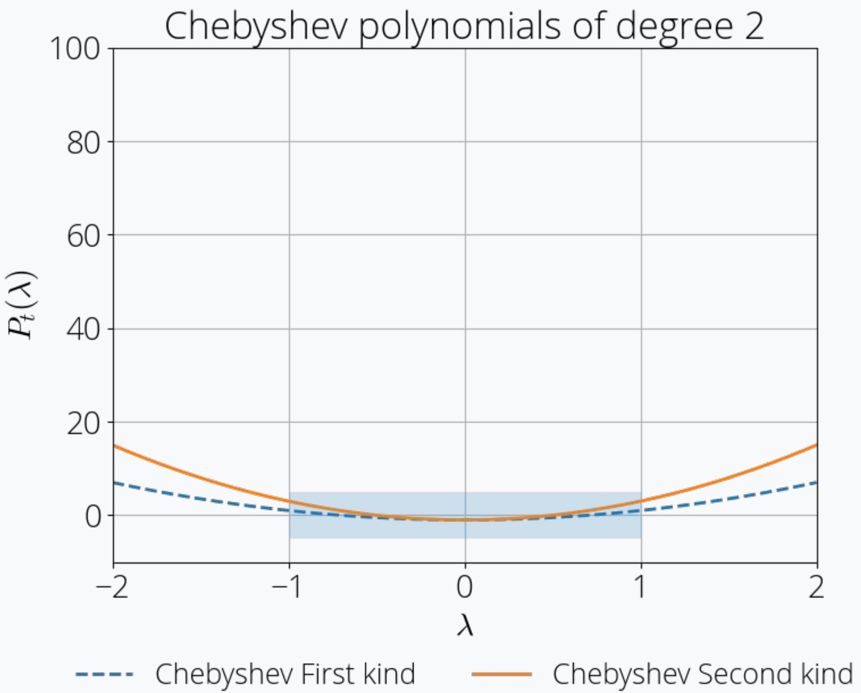
In the [-1, 1] interval, Chebyshev polynomials are linearly bounded.

$$|T_t(\xi)| \leq 1 \quad \text{and} \quad |U_t(\xi)| \leq t+1$$

Outside, they grow exponentially.

$$T_t(\xi) = \frac{1}{2}\left(\xi - \sqrt{\xi^2 - 1}\right)^t + \frac{1}{2}\left(\xi + \sqrt{\xi^2 - 1}\right)^t$$

$$U_t(\xi) = \frac{(\xi + \sqrt{\xi^2 - 1})^{t+1} - (\xi - \sqrt{\xi^2 - 1})^{t+1}}{2\sqrt{\xi^2 - 1}}\,.$$




Google Research

# Link function

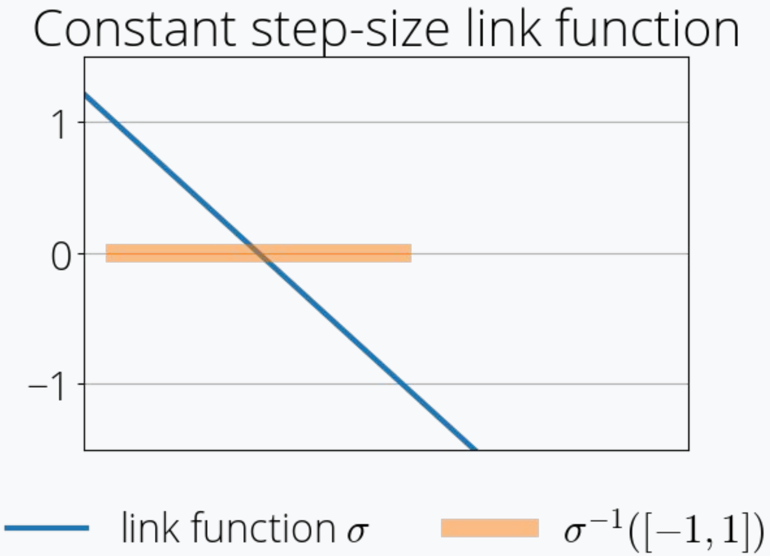
$$\sigma(\lambda) = \frac{1}{2\sqrt{m}}(1 + m - h\,\lambda)$$

Pre-image is also an interval:

$$\sigma^{-1}([-1,1]) = \left[\frac{(1-\sqrt{m})^2}{h}, \frac{(1+\sqrt{m})^2}{h}\right]$$

**Robust region**: Parameters for which

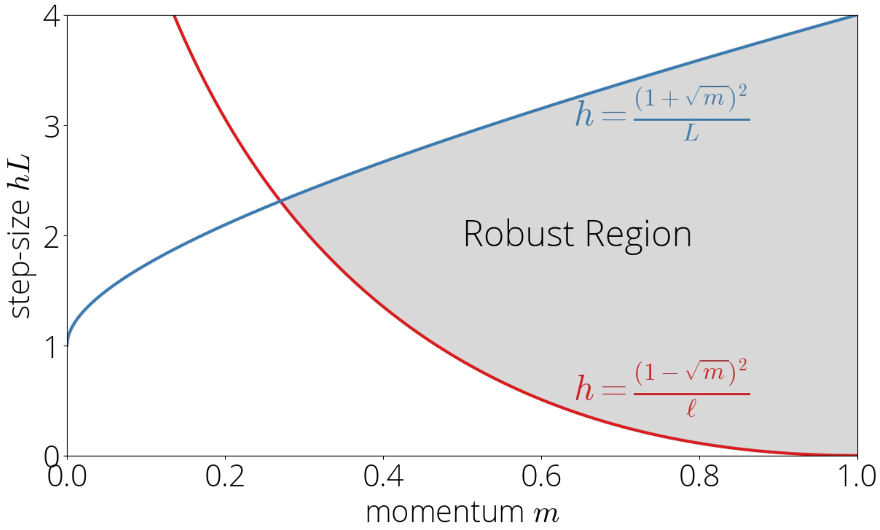
$[\mu, L] \subseteq \sigma^{-1}([-1,1])$

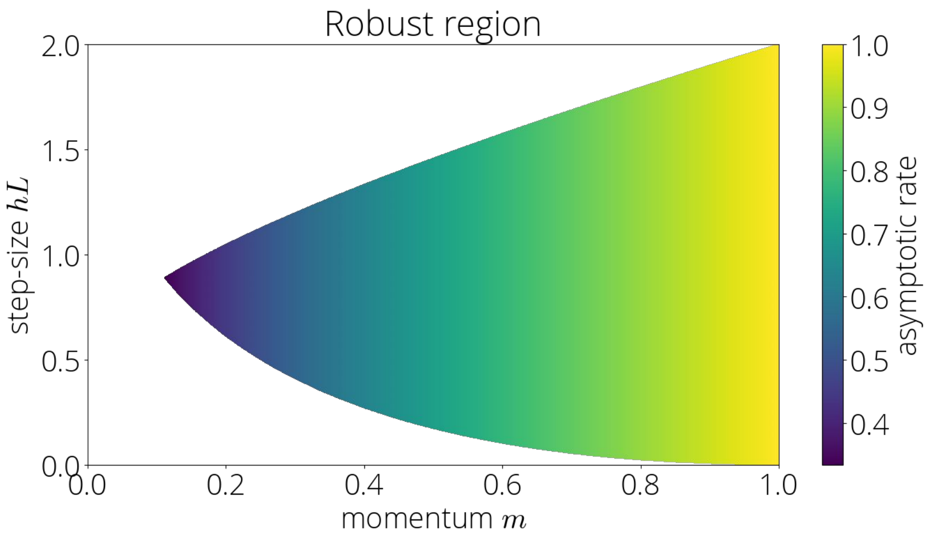



Google Research

The asymptotic rate in the robust region is $\sqrt{m}$.

$$\equiv \|x_t - x_\star\| = \mathcal{O}(\sqrt{m}^t)$$



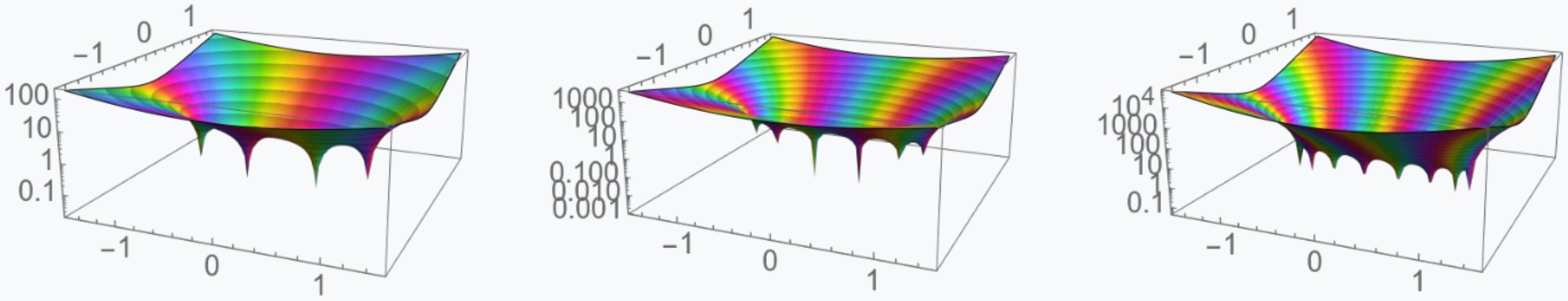
# Optimal parameters (aka Polyak HeavyBall)

Minimizing ***m*** in the robust region results in (worst-case) optimal params

$$m = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2 \qquad h = \left(\frac{2}{\sqrt{L} + \sqrt{\mu}}\right)^2$$

Asymptotic convergence rate: $\sqrt{m} = \dfrac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$

# 2. Cyclical HeavyBall



# Cyclical HeavyBall

Alternates between 2 step-sizes

Set $h_t = h_0$ if $t$ is odd and $h_t = h_1$ otherwise

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - h_t \nabla f(\boldsymbol{x}_t) + m(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})$$

Google Research

# Analysis of Cyclical HeavyBall

- Coefficients in recurrence now depends on t

$$P_{2t+1}(\lambda) = (1 + m + h_1\lambda)P_{2t}(\lambda) - mP_{2t-1}(\lambda)$$
$$P_{2t+2}(\lambda) = (1 + m + h_0\lambda)P_{2t+1}(\lambda) - mP_{2t}(\lambda)$$

- Known in the OP field as "orthogonal polynomials with varying coefficients" [Chihara (1968), Van Assche (1985)]

- Analyzed by chaining iterations:

$$P_{2t+2}(\lambda) = ((1 + m + h_0\lambda)(1 + m + h_1\lambda) - 2m)P_{2t}(\lambda) - m^2 P_{2t-2}(\lambda)$$


TS Chihara

Google Research

# Cyclical step-sizes

The residual polynomial for the cyclical HeavyBall method at even iterations is

$$P_{2t}(\lambda) = m^t \left( \frac{2m}{1+m} T_{2t}(\zeta(\lambda)) + \frac{1-m}{1+m} U_{2t}(\zeta(\lambda)) \right) , \qquad (6)$$

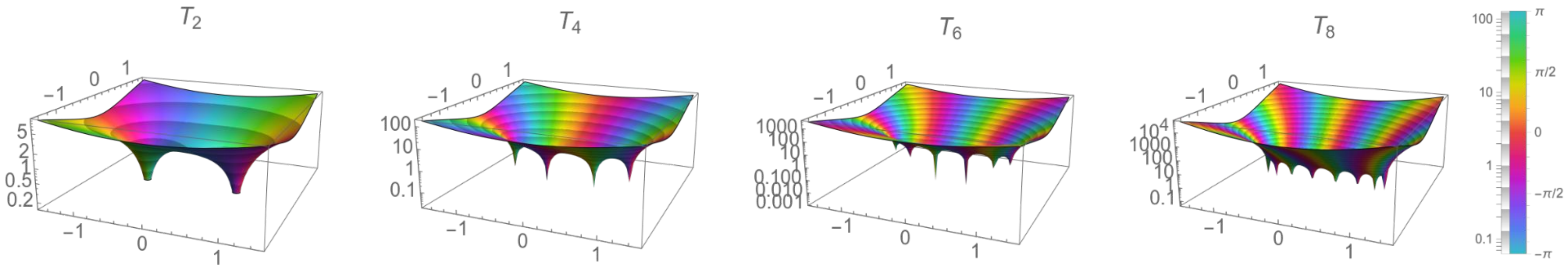
$$\textit{with } \zeta(\lambda) = \frac{1+m}{2\sqrt{m}} \sqrt{(1 - \frac{h_0}{1+m}\lambda)(1 - \frac{h_1}{1+m}\lambda)}.$$

★ Same than HeavyBall except for link function $\zeta$

Google Research

# Complex Chebyshev polynomials

Image of link function can now be real or imaginary




Chebyshev polynomials grow exponentially in $\mathbb{C}\backslash[-1, 1]$

# Link function

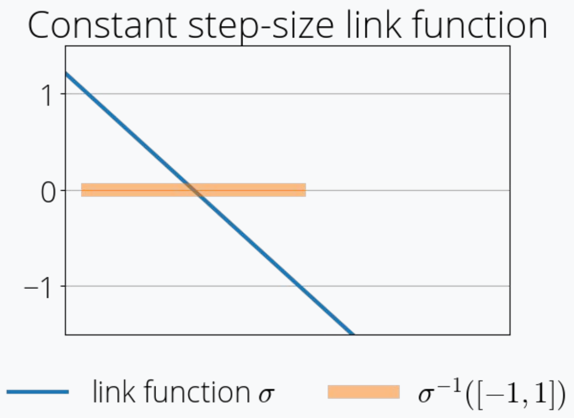
Pre-image no longer interval.

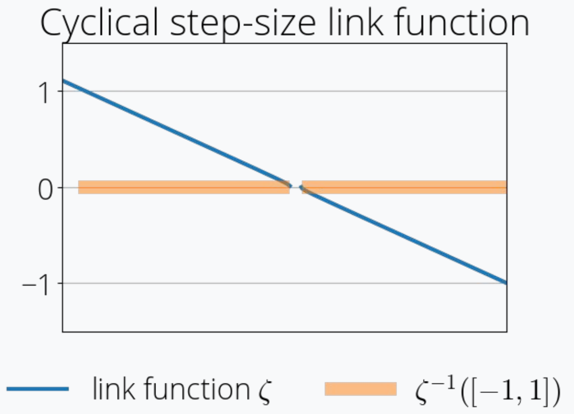
union of two intervals

**Robust region**: If

$$[\mu, L] \subseteq \sigma^{-1}([-1, 1])$$

Then $\|x_t - x_\star\| = \mathcal{O}(\sqrt{m}^t)$




Google Research

# A finer model for the Hessian eigenvalues

Consider eigenvalues in union of two disjoint intervals

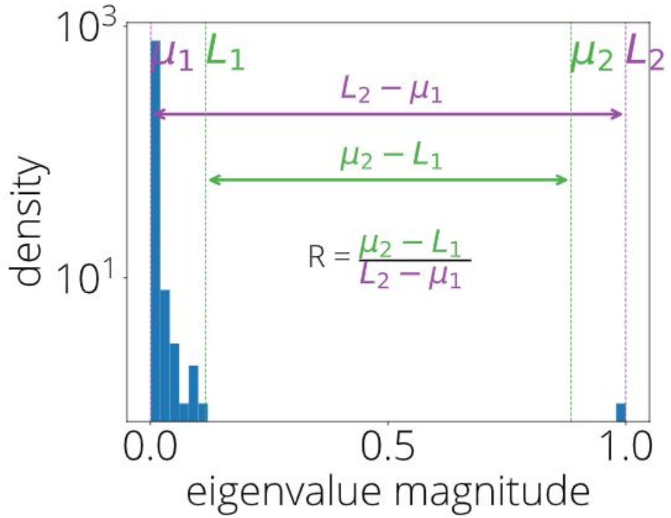
$$\Lambda = [\mu_1, L_1] \cup [\mu_2, L_2], \underbrace{L_1 - \mu_1 = L_2 - \mu_2}_{\text{same size}}.$$

The ratio *R* will play an important role:

$$R \triangleq \frac{\mu_2 - L_1}{L_2 - \mu_1}$$

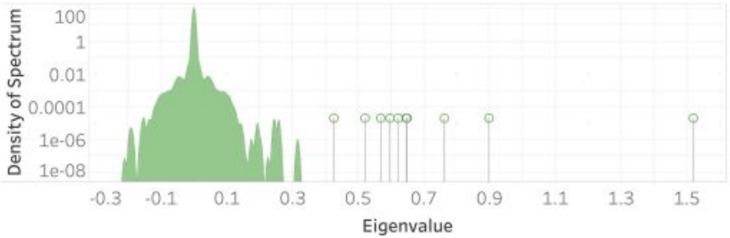
*R* = 0, one interval
R=1, all eigenvalues are at extremes.

MNIST

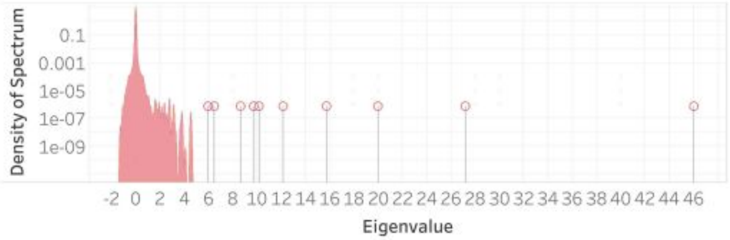



Google Research

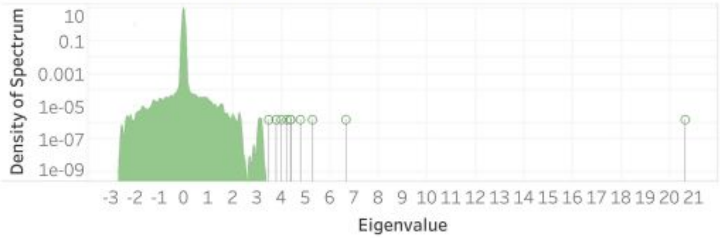
# Eigengaps Everywhere



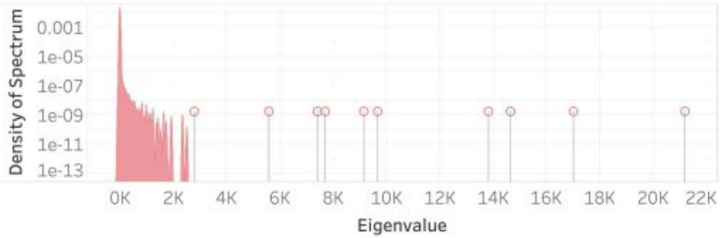

(a) MNIST, train




(b) MNIST, test




(e) CIFAR10, train




(f) CIFAR10, test

(Papyan 2020)

Google Research

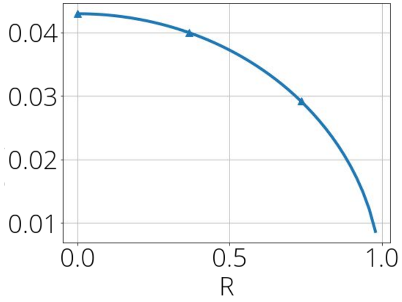
# Optimal parameters

Minimize $m$ s.t. $\underbrace{[\mu, L] \subseteq \sigma^{-1}([-1,1])}_{\text{robust region}}$

**Solution**

$$m = \left( \frac{\sqrt{\rho^2 - R^2} - \sqrt{\rho^2 - 1}}{\sqrt{1 - R^2}} \right)^2 \quad \text{with} \quad \rho \stackrel{\text{def}}{=} \frac{L_2 + \mu_1}{L_2 - \mu_1}$$

- R=0 we recover Polyak HeavyBall
- Decreasing in R

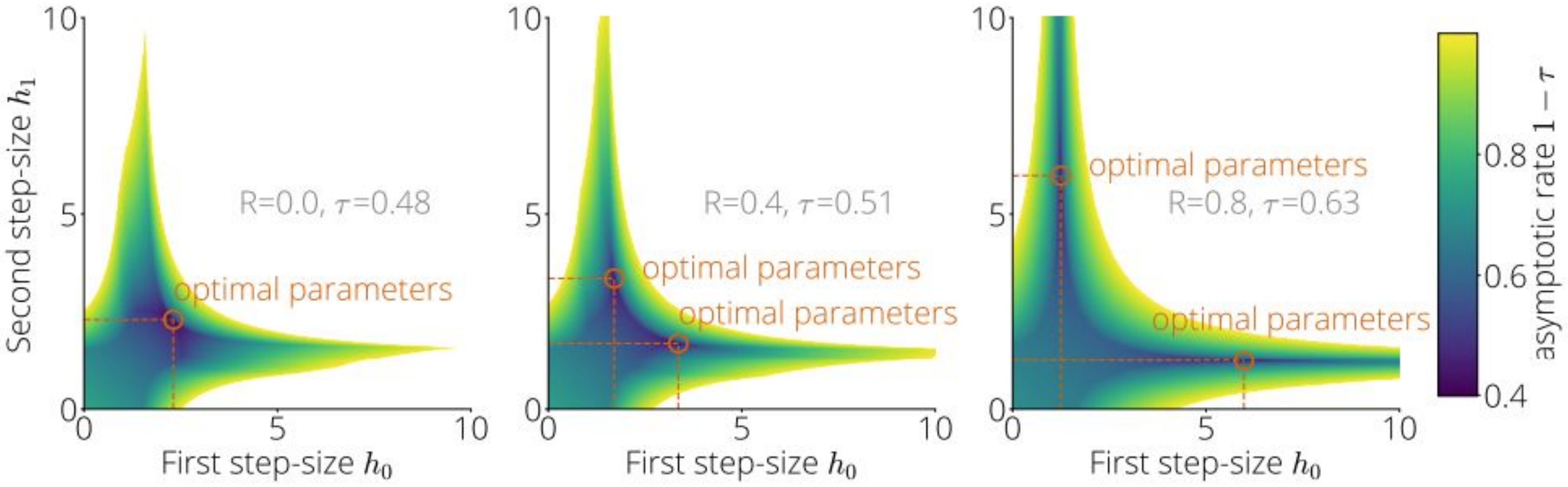


# Optimal step-sizes

$$h_t = \frac{1+m}{L_1} \text{ if } t \text{ is odd and } h_t = \frac{1+m}{\mu_2} \text{ otherwise}$$

Gap is small

Gap is large



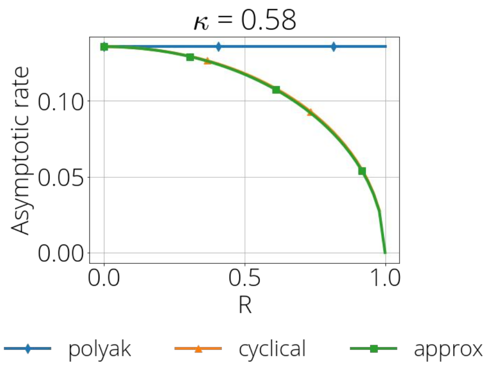

Google Research

# Convergence Rates

Asymptotic rate = $\sqrt{m} = \dfrac{\sqrt{\rho^2 - R^2} - \sqrt{\rho^2 - 1}}{\sqrt{1 - R^2}}$

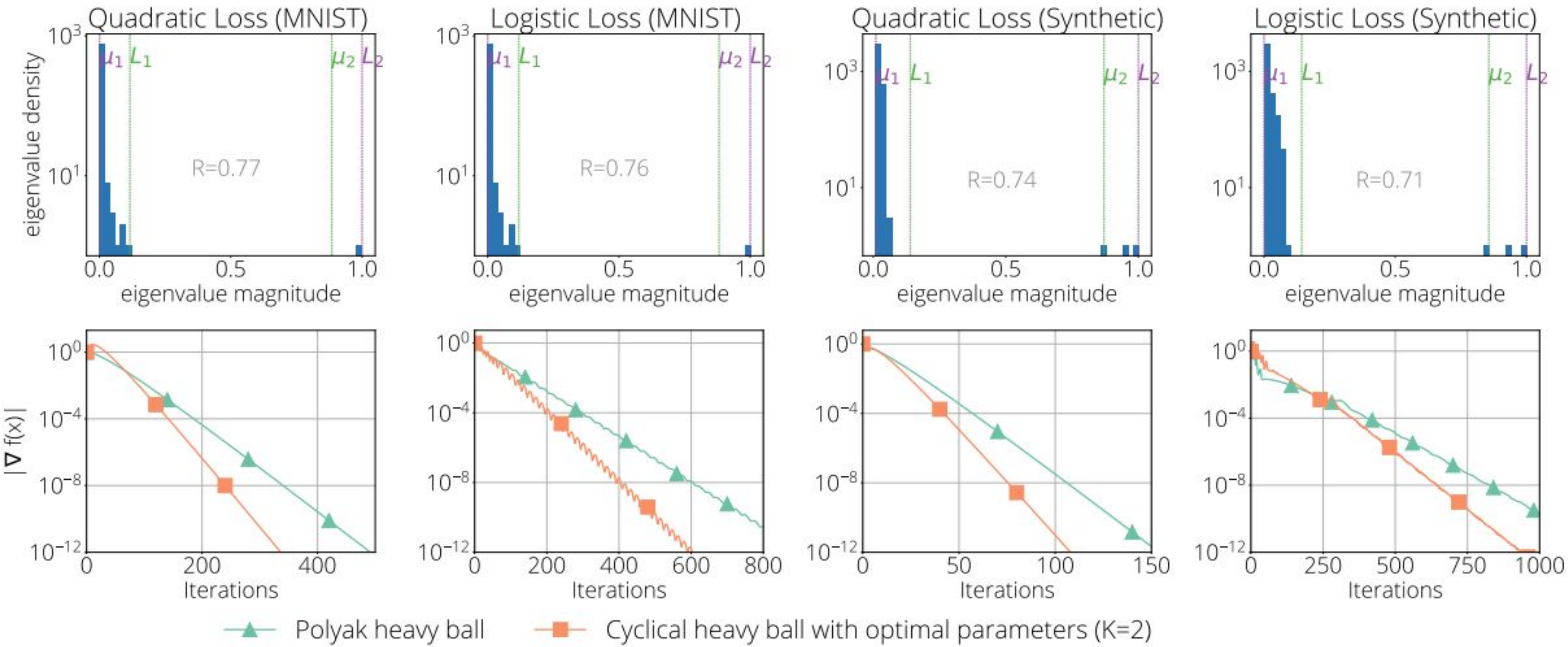
For ill-conditioned problems ($\mu \ll$ L),

$$\sqrt{m} \approx \sqrt{1 - R^2}\, r^{\mathrm{Polyak}}$$




Google Research

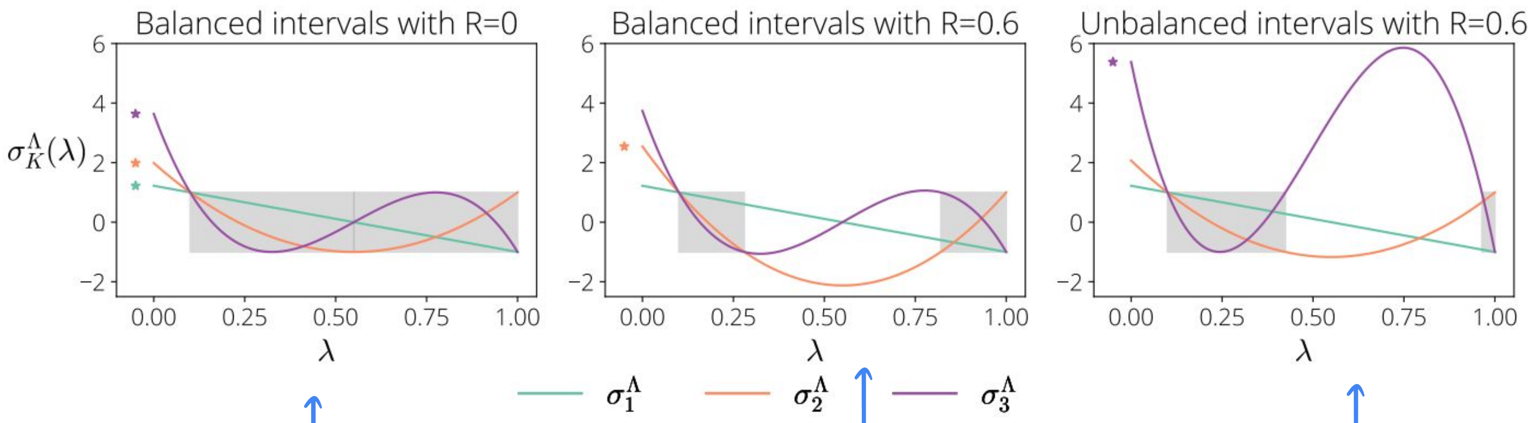
# Benchmarks


Quadratic Loss (MNIST)
Logistic Loss (MNIST)
Quadratic Loss (Synthetic)
Logistic Loss (Synthetic)
eigenvalue density
$\mu_1$ $L_1$ $\mu_2$ $L_2$
R=0.77
R=0.76
R=0.74
R=0.71
eigenvalue magnitude
$\|\nabla f(x)\|$
Iterations
Polyak heavy ball
Cyclical heavy ball with optimal parameters (K=2)

Google Research

# Beyond cycles of length 2

Link functionsare optimal if 2ζ-1 hit ±1 at edges and ∉[-1, 1] outside




1 cycle is optimal, 2/3 cycles are optimal, 3 cycles is optimal

# Conclusions

Cyclical Heavy Ball converges faster in the presence of spectral gap.

Assuming knowledge of this gap, converges at a rate $\approx \sqrt{1 - R^2}\, r^{\text{Polyak}}$

Speedup observed also on non-quadratic objectives.

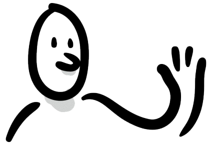
# Open Problems

More complex Hessian support: closed form for larger cycles.

Interpolating step-sizes

How to estimate the eigen-gap?

Stochastic algorithm? Non-quadratic objectives?



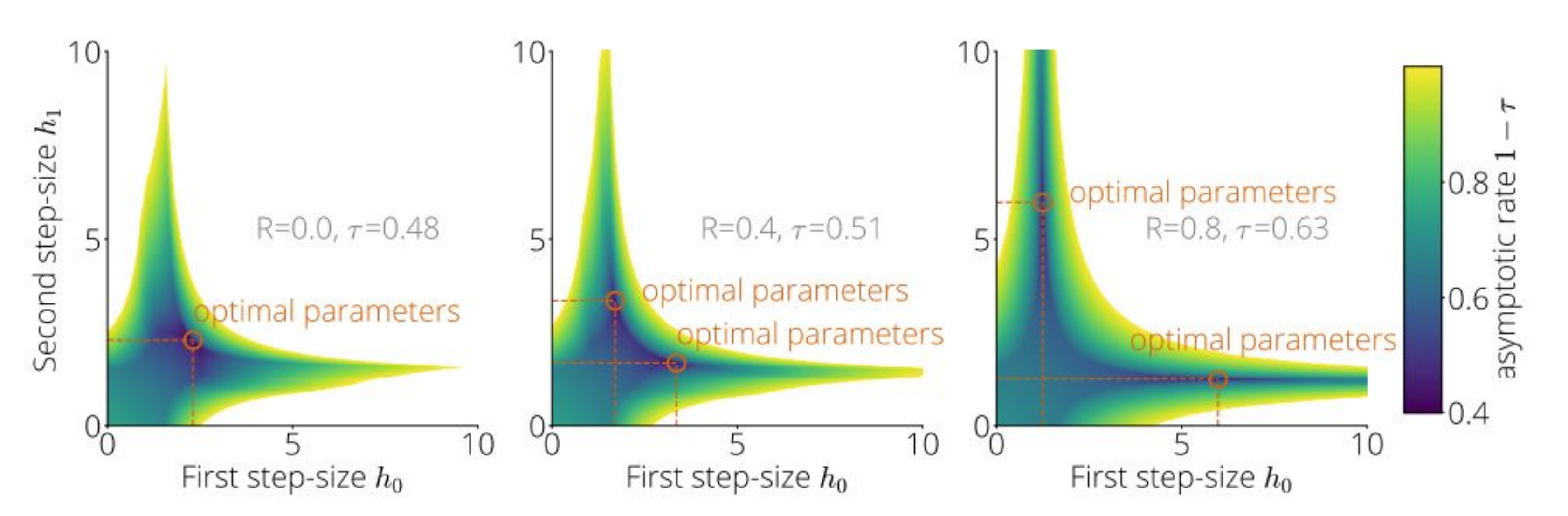
# Local convergence for non-quadratics

**Theorem 5.1** (Local convergence). *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a (potentially non-quadratic) twice continuously differentiable function, $x_*$ a local minimizer, and $H$ be the Hessian of $f$ at $x_*$ with $Sp(H) \subseteq \Lambda$. Let $x_t$ denote the result of running Algorithm 1 with parameters $h_1, h_2, \cdots, h_K, m$, and let $1 - \tau$ be the linear convergence rate on the quadratic objective (OPT). Then we have*

$$\forall \varepsilon > 0, \exists \text{ open set } V_\varepsilon : x_0, x_* \in V_\varepsilon \implies \|x_t - x_*\| = O((1 - \tau + \varepsilon)^t)\|x_0 - x_*\|. \tag{24}$$

Potential extension: A Modular Analysis of Provable Acceleration via Polyak's Momentum: Training a Wide ReLU Network and a Deep Linear Network, Jun-Kun Wang, Chi-Heng Lin, Jacob Abernethy

# 3. Acceleration & Simulations




Google Research