Super-acceleration with cyclical step-sizes

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

Google Research

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

aka gradient descent with momentum

Two parameters; step-size $h > 0$ and momentum $m \in (0, 1)$

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

Optimal among gradient-based methods on quadratics.

Stochastic variant popular in deep learning.
Cyclical HeavyBall

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

$$x_{t+1} = x_t - h_t \nabla f(x_t) + m(x_t - x_{t-1})$$

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

Pervasive (TF, PyTorch, optax, etc.)

No analysis that explains why/when it works.
Benchmarks

Today's topic

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

\[
\text{Slope} = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}
\]
Optimization and Polynomials

Cyclical HeavyBall

Simulations & Open problems
Polynomials and optimization

Some problems can be posed in the space of polynomials.

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

Consider Gradient Descent on

\[ f(x) = \frac{1}{2}(x - x^*)H(x - x^*) \]

Then at iteration \( t \) we have

\[
\begin{align*}
x_{t+1} - x^* &= x_t - x^* - \gamma H(x_t - x^*) \\
&= (I - \gamma H)(x_t - x^*) \\
&= \ldots \\
&= (I - \gamma H)^{t+1}(x_0 - x^*)
\end{align*}
\]

Polynomial in \( H \)
Real-valued polynomials

Taking norms on the previous expression

\[ \| \mathbf{x}_{t+1} - \mathbf{x}^* \|_2 \leq \|(I - \frac{2}{L+\mu} \mathbf{H})^{t+1} \|_2 \| \mathbf{x}_0 - \mathbf{x}^* \|_2 \]

Cauchy-Schwarz

Matrix 2-norm

\[ \leq \max_{\lambda \in [\mu, L]} \left| \left(1 - \frac{2}{L+\mu} \lambda \right)^{t+1} \right| \| \mathbf{x}_0 - \mathbf{x}^* \|_2 \]

Convergence rate \( \left( \frac{L - \mu}{L + \mu} \right)^t \)

\( \lambda_{\max}, \lambda_{\min} = \) largest and smallest eigenvalue of \( \mathbf{H} \)

The residual polynomial \( P_t^{GD} \), with \( t = 2 \)

\( P_t(\lambda) \)

Gradient Descent \( P_t^{GD} \)

\( \max_{\lambda \in [\mu, L]} |P_t^{GD}(\lambda)| \)
Gradient-based Methods and Polynomials

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

$$\|x_t - x^*\| \leq \max_{\lambda \in [\mu, L]} |P_t(\lambda)| \|x_0 - x^*\|. \quad (17)$$

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

The HeavyBall update

\[ x_{t+1} = x_t + m(x_t - x_{t-1}) - h \nabla f(x_t) \]

Gives the residual polynomial

\[ P_t(\lambda) = m^{t/2} \left( \frac{2m}{1 + m} T_t(\sigma(\lambda)) - \frac{m - 1}{1 + m} U_t(\sigma(\lambda)) \right) \]

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

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

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

Chebyshev polynomials of degree 2
**Link function**

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

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

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

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

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

Asymptotic convergence rate:

$$\sqrt{m} = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$$
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

$$x_{t+1} = x_t - h_t \nabla f(x_t) + m(x_t - x_{t-1})$$
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:

\[
P_{2t+2}(\lambda) = ((1 + m + h_0 \lambda)(1 + m + h_1 \lambda) - 2m)P_{2t}(\lambda) - m^2 P_{2t-2}(\lambda)
\]
The residual polynomial for the cyclical HeavyBall method at even iterations is

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

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 \( \zeta \)
Complex Chebyshev polynomials

Image of link function can now be real or imaginary

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

Pre-image no longer interval.

union of two intervals

**Robust region:** If

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

Then

\[ \| x_t - x_* \| = O(\sqrt{m^t}) \]
A finer model for the Hessian eigenvalues

Consider eigenvalues in union of two disjoint intervals

\[ \Lambda = [\mu_1, L_1] \cup [\mu_2, L_2], \quad L_1 - \mu_1 = L_2 - \mu_2. \]

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.
Eigengaps Everywhere

(a) MNIST, train

(b) MNIST, test

(e) CIFAR10, train

(f) CIFAR10, test

(Papyan 2020)
Minimize $m$ s.t. $[\mu, L] \subseteq \sigma^{-1}([-1, 1])$

**Robust region**

**Solution**

$$m = \left( \frac{\sqrt{\rho^2 - R^2} - \sqrt{\rho^2 - 1}}{\sqrt{1 - R^2}} \right)^2$$

with

$$\rho \overset{\text{def}}{=} \frac{L + \mu_1}{L - \mu_1}$$

- $R=0$ we recover Polyak HeavyBall
- Decreasing in $R$
Optimal step-sizes

\[ h_t = \frac{1+m}{L_1} \text{ if } t \text{ is odd and } h_t = \frac{1+m}{\mu_2} \text{ otherwise} \]
Convergence Rates

Asymptotic rate = \( \sqrt{m} = \frac{\sqrt{\rho^2 - R^2} - \sqrt{\rho^2 - 1}}{\sqrt{1 - R^2}} \)

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

\( \sqrt{m} \approx \sqrt{1 - R^2} r^{\text{Polyak}} \)
Benchmarks

Quadratic Loss (MNIST)

Logistic Loss (MNIST)

Quadratic Loss (Synthetic)

Logistic Loss (Synthetic)

\[ \text{eigenvalue density} \]

\( R = 0.77 \)

\[ \text{eigenvalue magnitude} \]

\[ R = 0.76 \]

\[ R = 0.74 \]

\[ R = 0.71 \]

\[ \| \nabla f(x) \| \]

\[ \text{Iterations} \]

\[ 10^{-12}, 10^{-8}, 10^{-4}, 10^0 \]

\[ \text{Polyak heavy ball} \]

\[ \text{Cyclical heavy ball with optimal parameters (K=2)} \]
Beyond cycles of length 2

Link functions are optimal if $2\zeta - 1$ hit $\pm 1$ at edges and $\notin [-1, 1]$ outside.

- Cycle 1 is optimal.
- Cycle 2 and 3 are optimal.
- Cycle 3 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?