Mittagsseminar

The Heavy Ball Method

Sebastian Stich

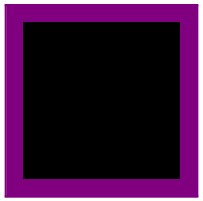
ETH Zürich

April 3, 2012

Table of contents

Black-box optimization

Given: $f : E \mapsto \mathbb{R}$
Goal: $\min_{\mathbf{x} \in E} f(\mathbf{x})$

$\mathbf{x} \rightarrow$  $\rightarrow \begin{cases} f(\mathbf{x}) \\ \nabla f(\mathbf{x}) \end{cases}$

- **Problem class** $f \in \mathcal{C}$
- **Oracle** access to $(f(\mathbf{x}), \nabla f(\mathbf{x}))$
- **Complexity:** number of oracle calls sufficient to solve any problem of the class
- **Solution:** $\mathbf{y} : f(\mathbf{y}) - \min_{\mathbf{x} \in E} f(\mathbf{x}) \leq \epsilon$

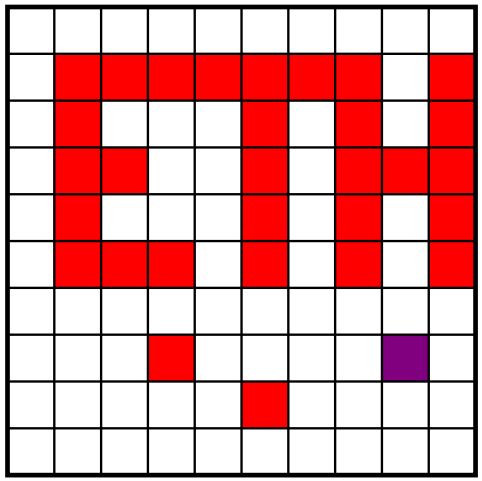
Example: Global Minimization

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L \|\mathbf{x} - \mathbf{y}\|_\infty$$
$$\forall \|\mathbf{x}\|_\infty, \|\mathbf{y}\|_\infty \leq 1$$

Oracle: $f(\mathbf{x})$

How difficult?
$\rightarrow$ **resisting oracle**

$$N(\epsilon) \geq \left(\frac{L}{2\epsilon}\right)^n$$

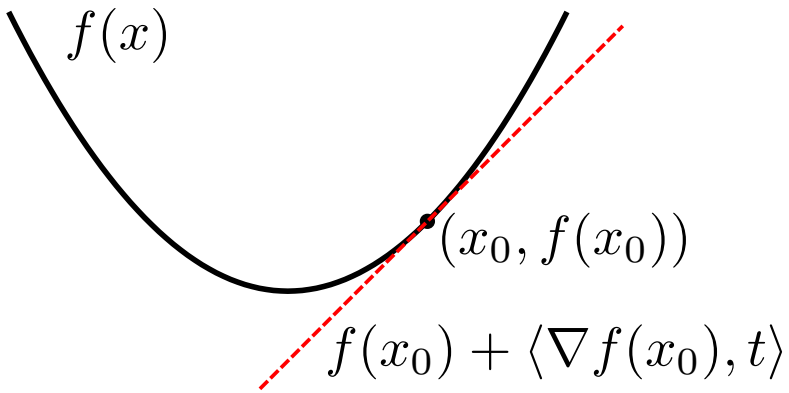


$f(\blacksquare) := 0 \quad f(\blacksquare) < 0$

Convex functions

first-order condition:

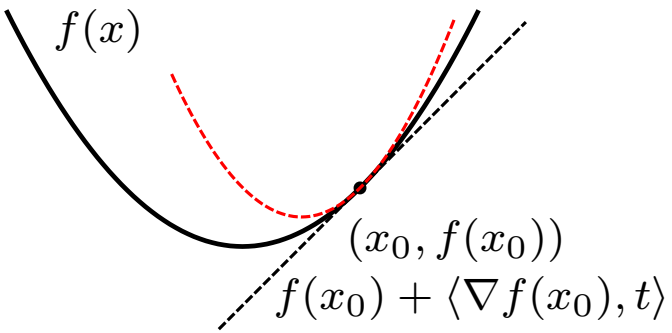
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \quad \forall \mathbf{x}, \mathbf{y} \in E$$



Convex functions II

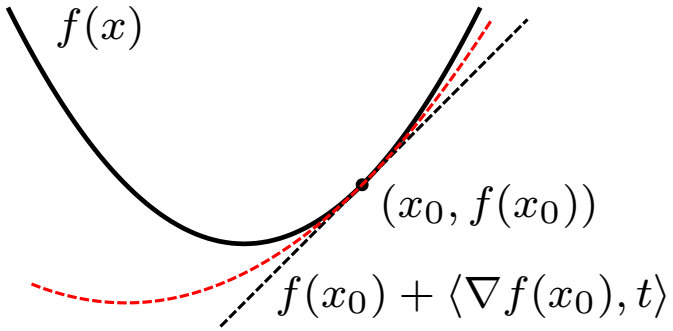
Quadratic upper bound:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2$$



Quadratic lower bound: (strongly convex)

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2$$



- We call $\kappa := L/\mu$ condition number; $\mu I_n \preceq \nabla^2 f(\mathbf{x}) \preceq L I_n$
- Only (!) for strongly convex: $\|\mathbf{x} - \mathbf{x}^*\|^2 \leq \frac{2}{\mu}(f(\mathbf{x}) - f(\mathbf{x}^*))$

Convex Optimization is hard

Your favorite hard problem (not SAT this time...):

PARTITION: *(weakly NP-hard)*

$$\sum_{i=1}^n a_i x_i = 0\,, \qquad x_i \in \{-1, 1\}\,,\ i = 1, \ldots, n$$

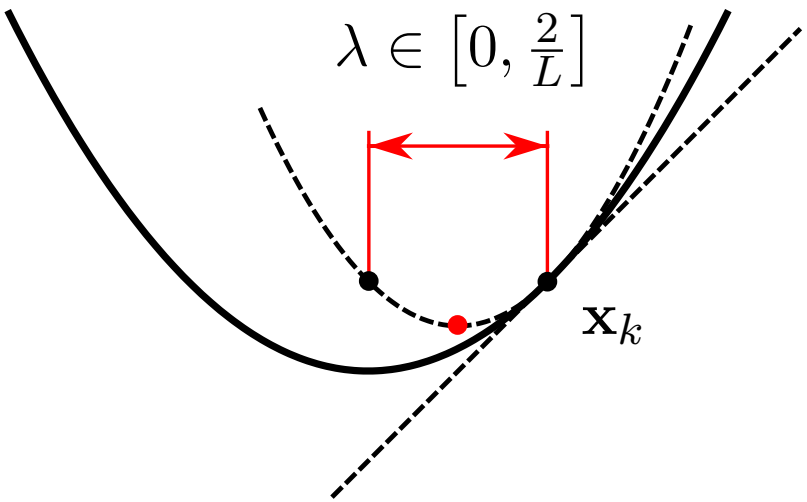
Convex polynomial:

$$f(\mathbf{x}) = \underbrace{\sum_{i=1}^n \mathbf{x}_i^4 - \frac{1}{n}\left(\sum_{i=1}^n \mathbf{x}_i^2\right)^2}_{\mathbf{x}_i^2 = \mathbf{x}_1^2} + \underbrace{\left(\sum_{i=1}^n a_i \mathbf{x}_i\right)^4}_{=0} + \underbrace{(1-\mathbf{x}_1)^4}_{\mathbf{x}_1 = 1} \geq 0$$

$$f(\mathbf{x}) = 0 \Leftrightarrow \sum_{i=1}^n a_i \mathbf{x}_i = 0$$

The Gradient Method

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \lambda \nabla f(\mathbf{x}_k)$$



For strongly convex functions: $\lambda \in \left[\frac{1}{L}, \frac{2}{L}\right]$.

Convergence

$$\underbrace{\text{Taylor: } \nabla f(\mathbf{x}_k) = \nabla^2 f(\mathbf{z})(\mathbf{x}_k - \mathbf{x}^*)}$$

$$\begin{aligned}\|\mathbf{x}_k - \lambda \nabla f(\mathbf{x}_k) - \mathbf{x}^*\| &= \left\| I_n \left[\mathbf{x}_k - \mathbf{x}^*\right] - \lambda \nabla f(\mathbf{x}_k) \right\| \\ &\leq \sup_{\mathbf{z}} \left\| I_n - \lambda \nabla^2 f(\mathbf{z}) \right\| \cdot \|\mathbf{x}_k - \mathbf{x}^*\|\end{aligned}$$

Eigenvalues of $I_n - \lambda \nabla^2 f(\mathbf{z})$ are all $\leq \max\left\{\left|1 - \lambda L\right|, \left|1 - \lambda \mu\right|\right\}$

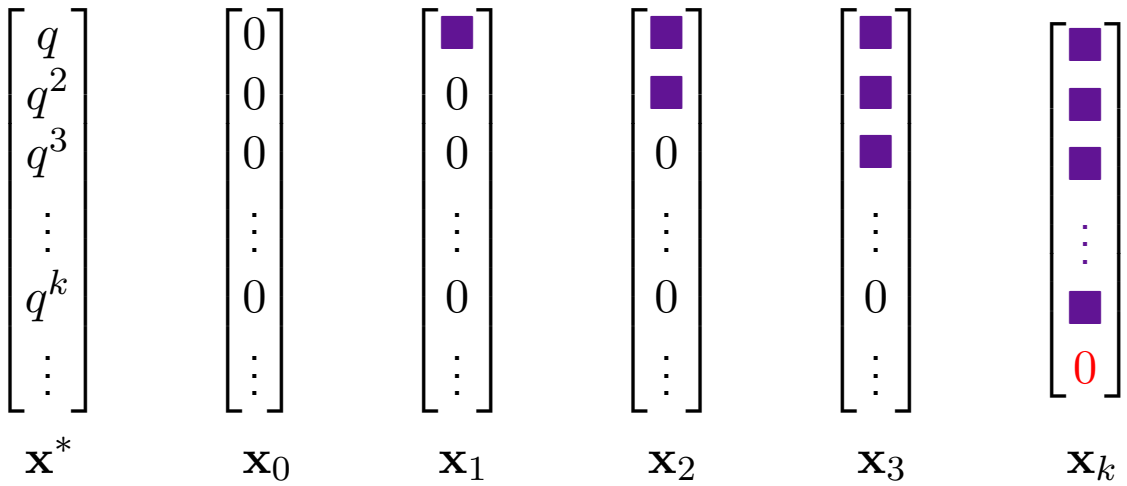
Optimal: $\lambda = \frac{2}{L+\mu}$ and

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|$$

Lower Bound on the Convergence

Assume: $\mathbf{x}_k \in \operatorname{span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_{k-1})\}$

We can find a function $f : l_2 \mapsto \mathbb{R}$ with:



$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \geq \sum_{i=k+1}^{\infty} q^{2i} = q^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Realized by a quadratic function

$$f(\mathbf{x}) = \frac{L-\mu}{8}\left(\langle A\mathbf{x}, \mathbf{x}\rangle - 2\mathbf{x}_1\right) + \frac{\mu}{2}\|\mathbf{x}\|^2$$

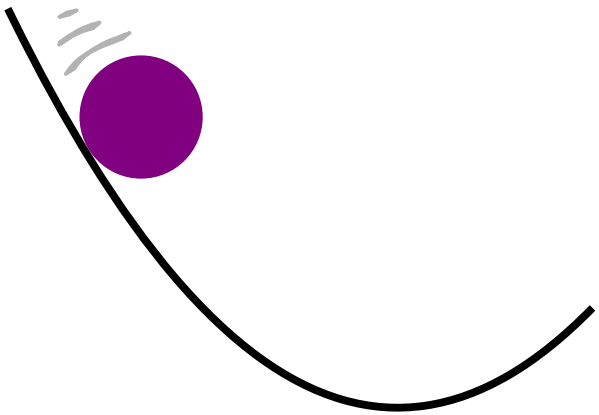
$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & \ddots \\ 0 & 0 & \ddots & \ddots \end{bmatrix}$$

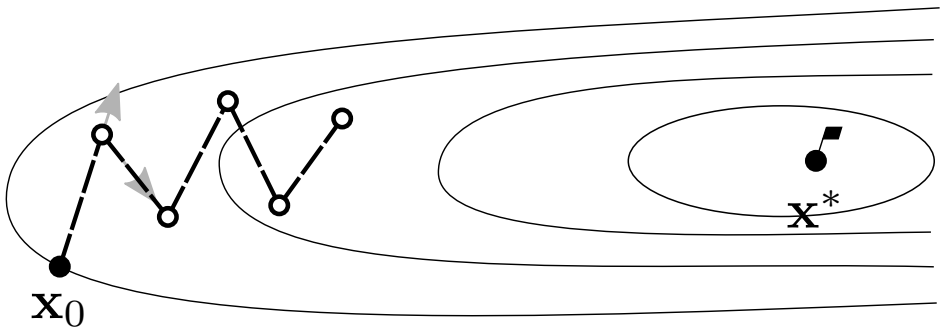
What Physics tells us..

$$\ddot{\mathbf{x}} = F \quad (= -\nabla f(\mathbf{x}))$$

Discretization:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \nabla f(\mathbf{x}_k) + (\mathbf{x}_k - \mathbf{x}_{k-1})$$



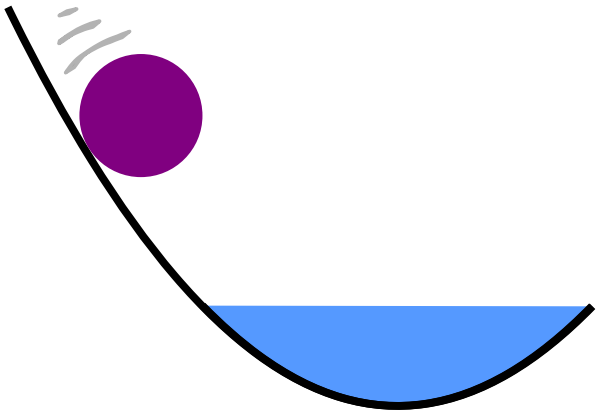


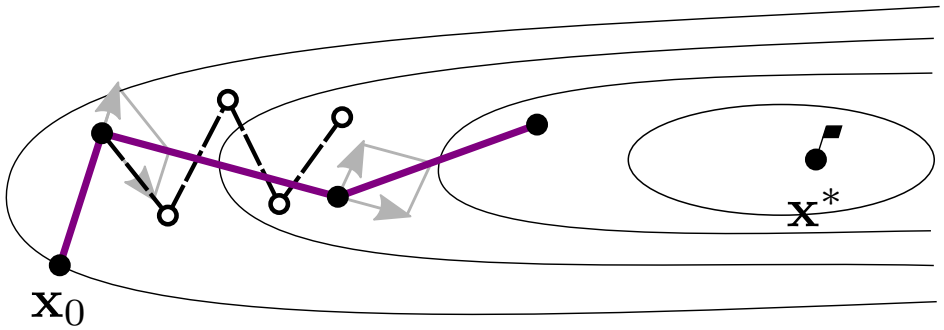
Damping

$$\ddot{\mathbf{x}} = F - b\dot{\mathbf{x}}$$

Discretization:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \nabla f(\mathbf{x}_k) + \beta(\mathbf{x}_k - \mathbf{x}_{k-1})$$





Polyak's Heavy Ball Method

$$\ddot{\mathbf{x}} = -a\nabla f(\mathbf{x}) - b\dot{\mathbf{x}}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\nabla f(\mathbf{x}) + \beta(\mathbf{x}_k - \mathbf{x}_{k-1})$$

Theorem

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|$$

For $\alpha = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$, $\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$.

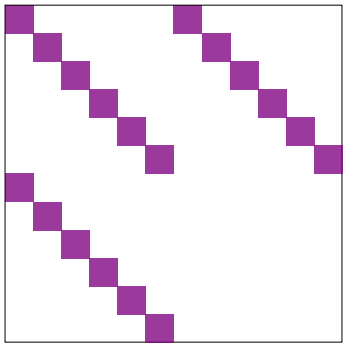
Lyapunov type analysis - idea

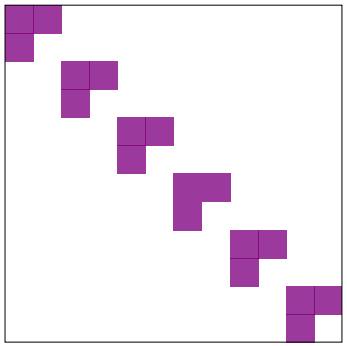
Write

$$\left\| \begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} \right\|^2 \leq \|A\|^2 \left\| \begin{bmatrix} \mathbf{x}_k - \mathbf{x}^* \\ \mathbf{x}_{k-1} - \mathbf{x}^* \end{bmatrix} \right\|^2$$

and show that all eigenvalues $|\lambda_i(A)| \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)$.

Matrix A has a nice structure:





We need only to compute eigenvalues of 2x2 matrices!

Lyapunov type analysis

$$\left\| \begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} (1+\beta) I_n & -\beta I_n \\ I_n & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_k - \mathbf{x}^* \\ \mathbf{x}_{k-1} - \mathbf{x}^* \end{bmatrix} - \alpha \begin{bmatrix} \nabla f(\mathbf{x}_k) \\ \mathbf{0} \end{bmatrix} \right\|^2$$

$$\leq \sup_{\mathbf{z}} \left\| \begin{bmatrix} (1+\beta) I_n - \alpha \nabla^2 f(\mathbf{z}) & -\beta I_n \\ I_n & 0 \end{bmatrix} \right\|^2 \left\| \begin{bmatrix} \mathbf{x}_k - \mathbf{x}^* \\ \mathbf{x}_{k-1} - \mathbf{x}^* \end{bmatrix} \right\|^2$$

2x2 matrices:

$$\begin{bmatrix} 1 + \beta - \alpha \lambda_i(\nabla^2 f(\mathbf{z})) & -\beta \\ 1 & 0 \end{bmatrix}$$

with $\mu \leq \lambda_i(\nabla^2 f(\mathbf{z})) \leq L$.

Limitations

- Constraint Handling
  - Possible! (with different methods)
- The parameters must be known...
  - Estimate them 'on the run' (not so easy)
- In general not monotone
  - again: resolved in more advanced schemes

Thank you

References

B. Polyak: Introduction to Optimization. Optimization Software - Inc, Publications Division, New York 1987. *available online*

Y. Nesterov, Introductory Lectures on Convex Optimization, *Kluwer*, Boston 2004.