Heavy-Ball Method

$$\min_{x \in \mathbb{R}^n} f(x)$$

f - convex, Lipschitz gradient.

Algorithm (Heavy-Ball Method)

Init: $x_0 \in \mathbb{R}^n$, $L > 0$, momentum parameter $0 \le \beta \le 1$
Fix $K \ge 1$ the number of iters. $s_0 = 0 \in \mathbb{R}^n$

For $k = 0 \dots K-1$:

1. Aggregate: $s_{k+1} = \beta s_k + f'(x_k)$
2. Perform a step: $x_{k+1} = x_k - \frac{1}{L} s_{k+1}$.

Return a point $\bar{x}$ with the best desired acc. measure.

The Gradient Method $\beta := 0$.

- $f(x_k) - f^* \le \dfrac{2LR^2}{k+4}$, $\quad R = \|x_0 - x^*\|$.

- $\min\limits_{0 \le i \le k-1} \|f'(x_i)\| \le \dfrac{4LR}{k}$, $\quad k \ge 1$.

- Strongly convex, smooth: $\mu I \preceq f''(x) \preceq LI \quad \forall x$

  Linear rate:

$$f(x_k) - f^* \le \exp\left(-k \cdot \frac{\mu}{L}\right) \cdot (f(x_0) - f^*)$$

To find: $f(x_k) - f^* \le \varepsilon \Rightarrow K = \frac{L}{\mu} \ln \frac{f(x_0) - f^*}{\varepsilon}$.

- $\beta = 1$.

- Form average points $\bar{x}_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i$

- Take the minimal gradient:

$$g_k = \min\{\|f'(\bar{x}_1)\|, \dots, \|f'(\bar{x}_k)\|\}.$$

- The rate of HB (f is convex quadratic):

$$g_k^2 \le \frac{12L(f(x_0)-f^*)}{k(k+1)} \qquad g_k = O\left(\frac{1}{k}\right)$$

- Assume $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$, $\mu = \lambda_{min}(A) > 0$.

$$f(x) - f^* \le \frac{1}{2\mu}\|f'(x)\|^2 \quad \forall x$$

$$f(\bar{x}_k^*) - f^* \le \frac{6L}{\mu}\cdot\frac{1}{k^2}(f(x_0)-f^*)$$

We can use restarts

$$K = \sqrt{\frac{12L}{\mu}}. \quad \text{Then}$$

$$f(\bar{x}_k^*) - f^* \le \frac{1}{2}(f(x_0)-f^*)$$

$\Rightarrow$ We need only $T = \log_2 \frac{f(x_0)-f^*}{\varepsilon}$ to get $f(\bar{x}) - f^* \le \varepsilon$.

Theorem The complexity of the HB method with restarts is

$$\sqrt{\frac{12L}{\mu}}\cdot\log_2\frac{f(x_0)-f^*}{\varepsilon}$$

oracle calls (matrix-vector products).

- $\beta = 0$ Gradient Method $\hat{O}(\frac{L}{\mu})$
- $\beta = 1$ with restarts : $\hat{O}(\sqrt{\frac{L}{\mu}}) \to$ optimal
- $\beta \approx 1 - \sqrt{\frac{\mu}{L}}$
- In practice : $\beta \approx 0.99$
- Only for Quadratic Functions.
  $\downarrow$
  Best Method: Conjugate Gradient Method.
- Nesterov's Fast Gradient Method

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

$$f'(x) = Ax - b \quad , \quad f'(x^*) = Ax^* - b = 0$$

$$f'(x) = A(x - x^*)$$

$$x_{n+1} = x_n - \alpha f'(x_n) + \beta(x_n - x_{n-1})$$

$$\underbrace{x_{n+1} - x^*}_{r_{n+1}} = \underbrace{x_n - x^*}_{r_n} - \alpha A\underbrace{(x_n - x^*)}_{r_n} + \beta(\underbrace{x_n - x^*}_{r_n} \underbrace{- x_{n-1} + x^*}_{-r_{n-1}})$$

$$\mathbb{R}^{2n} \ni \begin{bmatrix} r_{n+1} \\ r_n \end{bmatrix} = \overbrace{\left[\begin{array}{c|c} I(1+\beta) - \alpha A & -\beta I \\ \hline I & 0 \end{array}\right]}^{C} \begin{bmatrix} r_n \\ r_{n-1} \end{bmatrix}$$

$$\left\| \begin{bmatrix} r_{n+1} \\ r_n \end{bmatrix} \right\| \leqslant \left\| C \cdot \begin{bmatrix} r_n \\ r_{n-1} \end{bmatrix} \right\|$$

$$\boxed{\begin{array}{ll} \beta = 0 & \|C\| \leq (1 - \frac{\mu}{L}) \\ \alpha = \frac{1}{L} & \end{array}}$$

$\to$ min , $\alpha, \beta \Rightarrow$ get the optimal rate.

Lower Bounds for Smooth Convex Optimization

GM: $f(x_k) - f^* \leq O\left(\frac{LR^2}{k}\right)$

Theorem let $L > 0$. For any first-order optimization algorithm running for $K \geq 1$, $\exists$ convex function $f: \mathbb{R}^n \to \mathbb{R}$ with $\boxed{n \geq 2k+1}$, it has Lipschitz grad. with const $L > 0$ such that:

$$f(x_k) - f^* \geq \frac{3LR^2}{16(k+1)^2}.$$

Simplify everything:

- $x_0 = 0$. $\quad \varphi(x) = f(x - x_0)$
- $L = \text{const.}$ $(L = 4)$
- Linear Span Methods

Assume $\quad x_{k+1} \in L_{k+1} = \text{span}\{f'(x_0), \dots, f'(x_k)\}$.

$L_0 \subseteq L_1 \subseteq L_2 \subseteq L_3 \subseteq \dots$ Krylov Subspaces

$L_0 = \{0\}$

$L_1 = \text{span}\{f'(x_0)\}$

$L_2 = \text{span}\{f'(x_0), f'(x_1)\}, \dots$

Fix $k \geq 1$.

$$f_k(x) = \frac{1}{2}\left[\sum_{i=1}^{k-1} \left(x^{(i)} - x^{(i+1)}\right)^2 + \sum_{i=k}^{n} \left(x^{(i)}\right)^2 \right] - \langle b, x \rangle$$

$$= \frac{1}{2}\Big[\left(x^{(1)} - x^{(2)}\right)^2 + \left(x^{(2)} - x^{(3)}\right)^2 + \dots + \left(x^{(k-1)} - x^{(k)}\right)^2 +$$

$$+ \left(x^{(k)}\right)^2 + \dots + \left(x^{(n)}\right)^2 \Big] - \langle b, x \rangle$$

$$= \frac{1}{2}\|C_k x\|^2 - \langle b, x \rangle = \frac{1}{2}\langle \underbrace{C_k^T C_k}_{A_k} x, x \rangle - \langle b, x \rangle$$

$$C_k = \begin{bmatrix} D_k & 0 \\ 0 & I_{n-k} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

$$D_k = \begin{bmatrix} 1 & -1 & & & & \\ & 1 & -1 & & 0 & \\ & & 1 & -1 & & \\ & & & 1 & -1 & \\ & 0 & & & 1 & -1 \\ & & & & & 1 \end{bmatrix} \in \mathbb{R}^{k \times k}$$

$$D_k^T D_k$$

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

$$A_k = \left[\begin{array}{cccccc|c} 1 & -1 & 0 & 0 & 0 & 0 & \\ -1 & 2 & -1 & & 0 & & \\ 0 & -1 & 2 & -1 & & & 0 \\ 0 & & -1 & 2 & -1 & & \\ 0 & 0 & & -1 & 2 & & \\ & & & & & 1 & \\ \hline & & 0 & & & & I \end{array}\right]$$

$$f_k(x) = \frac{1}{2}\langle A_k x, x \rangle - \langle b, x \rangle$$

• Simple enough, $b = e_1$

• Graphs

Laplacian of this graph $L = D - G$

$$\langle f_n''(x)h,h\rangle = \sum_{i=1}^{n-1}(h^{(i)}-h^{(i+1)})^2 + \sum_{i=k}^{n}(h^{(i)})^2 \geq 0$$

$$\leq \sum_{i=1}^{n-1} 2(h^{(i)})^2 + 2(h^{(i+1)})^2 + \sum_{i=k}^{n}(h^{(i)})^2 \leq$$

$$\leq 4\cdot\|h\|_2^2 \quad \Rightarrow \boxed{L=4}$$

$$x_n^* = \operatorname*{argmin}_{x\in\mathbb{R}^n} f_n(x)$$

$b = e_1$

$$\begin{cases} x^{(1)} - x^{(2)} - 1 = 0 & i=1 \\ 2x^{(i)} - x^{(i-1)} - x^{(i+1)} = 0 & 2\leq i\leq k \\ 2x^{(k)} - x^{(k-1)} = 0 & i=k \\ x^{(i)} = 0 & k<i\leq n \end{cases}$$

$$\boxed{\begin{aligned} (x_n^*)^{(1)} &= k \\ (x_n^*)^{(2)} &= k-1 \\ &\dots \\ (x_n^*)^{(k)} &= 1 \\ (x_n^*)^{(i)} &= 0 \end{aligned}}$$

For $k<i\leq n$.

$$f_n^* = \frac{1}{2}\langle A_n x_n^*, x_n^*\rangle - \langle e_1, x_k^*\rangle = -\frac{1}{2}\langle e_1, x_n^*\rangle = -\frac{k}{2}.$$

$$\|x_0 - x_n^*\|^2 = \|x_n^*\|^2 = \sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6} \leq \frac{(k+1)^3}{6}.$$

Proposition Assume $\{x_k\}$ that satisfies: $x_k \in L_k$. Then

$$L_k \subseteq \mathbb{R}^{n,k} = \{ x \in \mathbb{R}^n \mid x^{(i)} = 0 \quad i \geq k+1 \}.$$

Proof. $L_0 = \{0\} \in \mathbb{R}^{n,0}$

$L_1 = \operatorname{span}\{ f'(x_0) \} = \operatorname{span}\{ -e_1 \} \in \mathbb{R}^{n,1}$

.....

By induction, $\underset{\mathbb{R}^{n,k+1}}{f'(x)} = A_k x - b$, $x \in \mathbb{R}^{n,k}$ for tridiag. matrix.

Proposition: $f_k(x)$, $f_{k+p}(x)$, $p \geq 0$

$$f_k(x) \equiv f_{k+p}(x) \quad \forall x \in \mathbb{R}^{n,k}.$$

<u>To prove the lower bound</u>: $f(x) \equiv f_{2k+1}(x)$

$$f(x_k) = f_{2k+1}(x_k) = f_k(x_k) \geq f_k^* = -\frac{k}{2}.$$

But

$$f^* = f_{2k+1}^*(x_{2k+1}^*) = -\frac{2k+1}{2}. \quad \blacksquare$$