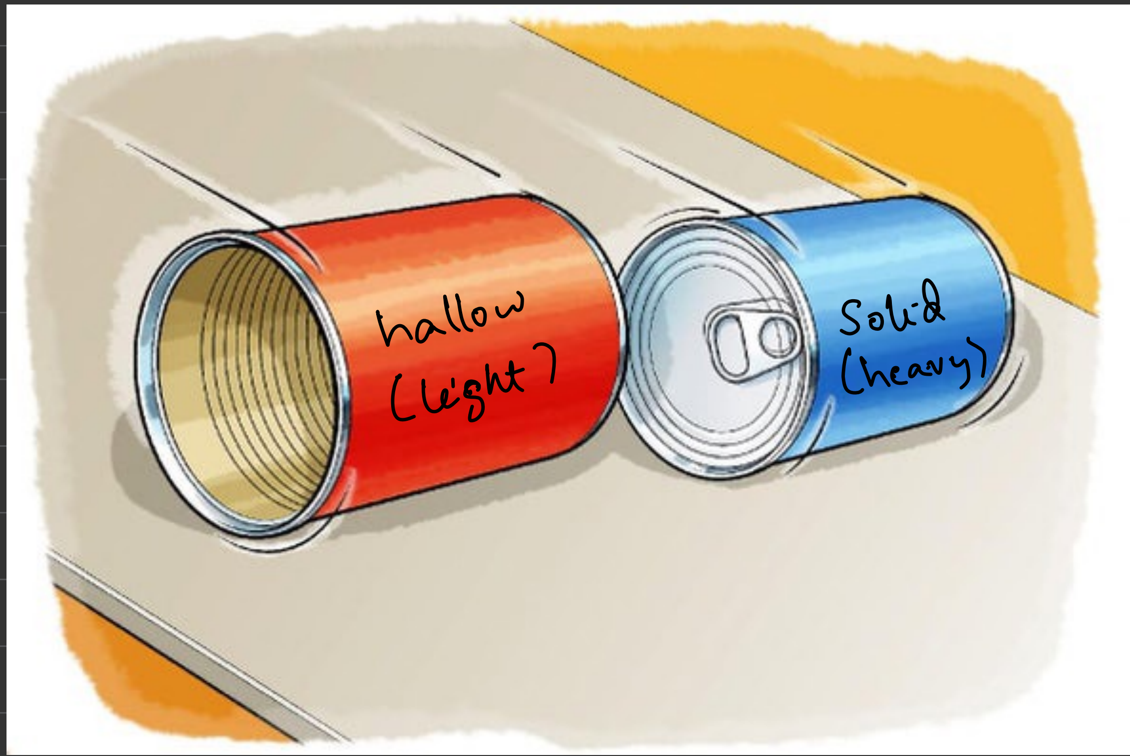
Lecture 9 Accelerated Gradient methods EI 260

- Heavy-ball method
  - Quadratic function
  - iteration complexity $O\left(\sqrt{\kappa}\log\left(\frac{1}{\varepsilon}\right)\right)$
- Nestrov's acceleration
  (just the idea today)



Which object goes down the ramp faster?

- added inertia acts as a smoother and an accelerator

Heavy-ball method or Polyak momentum

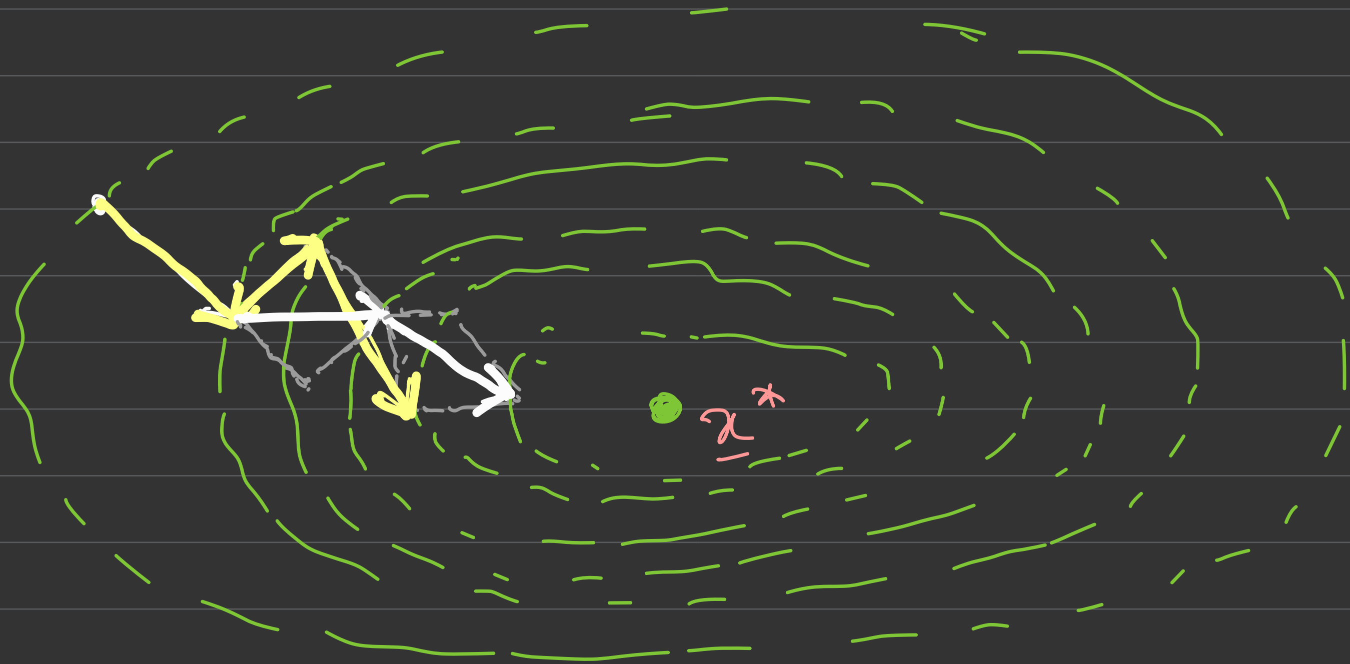
B.T. Polyak (1964)

" Some methods of speeding up convergence of iteration methods."

$$\underset{\underline{x} \in \mathbb{R}^n}{\text{minimize}} \quad f(\underline{x})$$

$$\boxed{\underline{x}_{t+1} = \underline{x}_t - \eta_t \nabla f(\underline{x}_t) + \underbrace{\theta_t (\underline{x}_t - \underline{x}_{t-1})}_{\text{momentum term}}}$$

$$= (1+\theta_t)\underline{x}_t - \eta_t \nabla f(\underline{x}_t) - \theta_t \underline{x}_{t-1}$$



gradient descent

heavy-ball method

Quadratic problem:

$$\underset{\underline{x}}{\text{minimize}} \quad \frac{1}{2} (\underline{x} - \underline{x}^*)^T Q (\underline{x} - \underline{x}^*)$$

Recall $\nabla f(\underline{x}) = Q(\underline{x} - \underline{x}^*)$

Gradient descent:

$$\| \underline{x}_t - \underline{x}^* \|_2 \leq \left(\frac{K-1}{K+1} \right)^t \| \underline{x}_0 - \underline{x}^* \|_2$$

iteration complexity: $O\left(K \log \left(1/\varepsilon \right) \right)$

Heavy-ball State-space model:

$$\begin{bmatrix} \underline{x}_{t+1} \\ \underline{x}_t \end{bmatrix} = \begin{bmatrix} (1+\theta_t) I & -\theta_t I \\ I & 0 \end{bmatrix} \begin{bmatrix} \underline{x}_t \\ \underline{x}_{t-1} \end{bmatrix} - \begin{bmatrix} \eta_t \nabla f(\underline{x}_t) \\ \underline{0} \end{bmatrix}$$

$$\begin{bmatrix} \underline{x}_{t+1} - \underline{x}^* \\ \underline{x}_t - \underline{x}^* \end{bmatrix} = \begin{bmatrix} (1+\theta_t) I & -\theta_t I \\ I & 0 \end{bmatrix} \begin{bmatrix} \underline{x}_t - \underline{x}^* \\ \underline{x}_{t-1} - \underline{x}^* \end{bmatrix} - \begin{bmatrix} \eta_t \nabla f(\underline{x}_t) \\ \underline{0} \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} (1+\theta_t) I - \eta_t Q & -\theta_t I \\ I & 0 \end{bmatrix}} \begin{bmatrix} \underline{x}_t - \underline{x}^* \\ \underline{x}_{t-1} - \underline{x}^* \end{bmatrix}$$

Convergence depends on the spectrum of : H_t

Find η_t and θ_t to control the spectrum of H_t

- Suppose λ_i is the i th eigenvalue of Q
- Spectral radius of H_t

$$\rho(H_t) = \rho\left(\begin{bmatrix} (1+\theta_t) I - \eta_t \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} & -\theta_t I \\ I & 0 \end{bmatrix}\right)$$

$$\leq \max_{1 \leq i \leq n} \rho\left(\underbrace{\begin{bmatrix} 1+\theta_t - \eta_t \lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}}_{\text{two eigenvalues are}}\right)$$

roots of $\quad s^2 - (1+\theta_t - \eta_t \lambda_i) s + \theta_t = 0$

- If $(1+\theta_t - \eta_t \lambda_i)^2 - 4\theta_t \leq 0$, then roots have the same magnitude $\sqrt{\theta_t}$ (imaginary roots)

- $(1+\theta_t - \eta_t \lambda_i)^2 \leq 4\theta_t$ is satisfied if

$$\theta_t \in \left[(1-\sqrt{\eta_t \lambda_i})^2, (1+\sqrt{\eta_t \lambda_i})^2 \right]$$

Suppose f is a L-smooth and μ strongly convex quadratic function

$$0 \leq \mu I \leq Q \leq L I$$

$$\theta_t = \max \left\{ (1-\sqrt{\eta_t L})^2, (1-\sqrt{\eta_t \mu})^2 \right\}$$

still yields $(1+\theta_t - \eta_t \lambda_i)^2 \leq 4\theta_t$

$$\Rightarrow \quad e(H_t) \leq \sqrt{\theta_t}$$

Now,

$$1 - \sqrt{\eta_t L} = -\left(1 - \sqrt{\eta_t \mu}\right) \Rightarrow \eta_t = \frac{4}{\left(\sqrt{L} + \sqrt{\mu}\right)^2}$$

$$\Rightarrow \theta_t = \max\left\{\left(1 - \frac{2\sqrt{L}}{\sqrt{L} + \sqrt{\mu}}\right)^2, \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2\right\}$$

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

$$\Rightarrow e(H_t) \leq \frac{\sqrt{K} - 1}{\sqrt{K} + 1}$$

Theorem:

Suppose f is a L-smooth and μ-strongly convex quadratic function. Then $\eta_t = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and

$$\theta_t = \max\left\{ |1 - \sqrt{\eta_t L}|, |1 - \sqrt{\eta_t \mu}| \right\}^2 \quad \text{with } K = \frac{L}{\mu}$$

results in

$$\boxed{\left\| \begin{bmatrix} \underline{x}_{t+1} - \underline{x}^* \\ \underline{x}_t - \underline{x}^* \end{bmatrix} \right\|_2 \leq \left(\frac{\sqrt{K} - 1}{\sqrt{K} + 1} \right)^t \left\| \begin{bmatrix} \underline{x}_1 - \underline{x}^* \\ \underline{x}_0 - \underline{x}^* \end{bmatrix} \right\|_2}$$

Recall $\exp(-x) \geq 1 - x$

$$\exp\left(-\frac{2}{\sqrt{K+1}}\right) \geq 1 - \frac{2}{\sqrt{K+1}}$$

$$\Rightarrow \quad \left\| \begin{bmatrix} \underline{x}_{t+1} - \underline{x}^* \\ \underline{x}_t - \underline{x}^* \end{bmatrix} \right\|_2^2 \leq \exp\left(-\frac{4t}{\sqrt{K+1}}\right) \left\| \begin{bmatrix} \underline{x}_1 - \underline{x}^* \\ \underline{x}_0 - \underline{x}^* \end{bmatrix} \right\|_2^2$$

Iteration complexity : $O\left(\sqrt{K} \log\left(1/\varepsilon\right)\right)$

Suppose condition number is $K = 100$

$$10 \log\left(1/\varepsilon\right) \lesssim 100 \log\left(1/\varepsilon\right)$$

How about more general convex functions?

Nestrov's method:

Yuri Nestrov 1983

$$\boxed{\begin{aligned} \underline{y}_{t+1} &= \underline{x}_t - \eta \nabla f(\underline{x}_t) \\ \underline{x}_{t+1} &= \underline{y}_{t+1} + \frac{t}{t+3}\left(\underline{y}_{t+1} - \underline{y}_t\right) \end{aligned}}$$

$$\underline{y}_{t+1} = \underline{x}_t - \eta_t \nabla f(\underline{x}_t) \quad : \text{normal}$$

$$\underline{z}_{t+1} = \underline{z}_t - \eta_t \cdot \frac{t+1}{2} \nabla f(\underline{x}_t) \quad : \text{aggressive}$$

$$\underline{x}_{t+1} = \frac{t+1}{t+3}\underline{y}_{t+1} + \underbrace{\frac{2}{t+3}}_{\text{lower weight}}\underline{z}_{t+1} \quad : \text{average}$$