Lecture 10 Accelerated Gradient Methods E1 260

- Nesterov's acceleration
  - Convergence for L-smooth convex

$$O\left(1/\sqrt{\varepsilon}\right)$$

  - Interpretation via Second-order ODE

Nesterov's method:

Yuri Nesterov 1983

$$y_{t+1} = x_t - \eta \nabla f(x_t)$$

$$x_{t+1} = y_{t+1} + \frac{t}{t+3}\left(y_{t+1} - y_t\right)$$

$$y_{t+1} = x_t - \eta_t \nabla f(x_t)$$: normal

$$z_{t+1} = z_t - \eta_t \cdot \frac{t+1}{2} \nabla f(x_t)$$: aggressive

$$x_{t+1} = \frac{t+1}{t+3} y_{t+1} + \frac{2}{t+3} z_{t+1}$$: average

($\frac{2}{t+3}$: lower weight)

Recall :

For $f : \mathbb{R}^d \to \mathbb{R}$ that is L smooth and convex, gradient descent yields

$$f(\underline{x}_T) - f(\underline{x}^*) \leq \frac{L}{2T} \| \underline{x}_0 - \underline{x}^* \|^2 \quad , T \geq 0$$

Iteration complexity : $O(1/\varepsilon)$

$\varepsilon \sim 10^{-6}$

$T_{GD} \geq 10^{6}$

What does Nestrov's accelerated gradient yield?

Iteration complexity of $O(1/\sqrt{\varepsilon})$

$T_{NGD} \geq 10^{3}$

Let us define the energy function (potential or Lyapunov function) and assign to each time 't':

$$\phi(t) = t(t+1)\left(f(\underline{y}_t) - f(\underline{x}^*)\right) + 2L\,\|\underline{z}_t - \underline{x}^*\|_2^2$$

($\underline{z}_t$: aggressive step)

If $\boxed{\phi(t+1) \leq \phi(t)}$, for each t, we have

$$\underbrace{T(T+1)\left(f(\underline{y}_T) - f(\underline{x}^*)\right) + 2L\,\|\underline{z}_T - \underline{x}^*\|_2^2}_{\phi(T)} \leq \underbrace{2L\,\|\underline{z}_0 - \underline{x}^*\|^2}_{\phi(0)}$$

$$\Rightarrow \quad f(\underline{y}_T) - f(\underline{x}^*) \leq \frac{2L\,\|\underline{z}_0 - \underline{x}^*\|^2}{T(T+1)}$$

$$\Rightarrow \quad T \approx O\left(\frac{1}{\sqrt{\varepsilon}}\right)$$

Recall from the vanilla analysis and for L-Smooth functions:

with $\eta_t = \eta = \frac{t+1}{2L}$ and $\underline{g}_t = \nabla f(\underline{x}_t)$

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

- $$\underline{g}_t^T(\underline{z}_t - \underline{x}^*) = \frac{t+1}{4L}\|\underline{g}_t\|^2 + \frac{L}{t+1}\left(\|\underline{z}_t - \underline{x}^*\|^2 - \|\underline{z}_{t+1} - \underline{x}^*\|^2\right)$$

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

- $$f(\underline{y}_{t+1}) \leq f(\underline{x}_t) - \frac{1}{2L}\|\underline{g}_t\|_2^2 \quad ; \eta = \frac{1}{L}$$

- Convexity: $f(\underline{x}_t) - f(\underline{w}) \leq \underline{g}_t^T(\underline{x}_t - \underline{w})$

For the potential function:

$$\phi(t+1) = t(t+1)\left(f(\underline{y}_{t+1}) - f(\underline{x}^*)\right) + 2(t+1)\left[f(\underline{y}_{t+1}) - f(\underline{x}^*)\right] + 2L\,\|\underline{z}_{t+1} - \underline{x}^*\|^2$$

$$\phi(t) = t(t+1)\left(f(\underline{y}_t) - f(\underline{x}^*)\right) + 2L\,\|\underline{z}_t - \underline{x}^*\|^2$$

$$\Delta = \frac{\phi(t+1) - \phi(t)}{t+1}$$

$$\Delta = t\left[f(\underline{y}_{t+1}) - f(\underline{y}_t)\right] + 2\left[f(\underline{y}_{t+1}) - f(\underline{x}^*)\right] + \frac{2L}{t+1}\left[\|\underline{z}_{t+1} - \underline{x}^*\|^2 - \|\underline{z}_t - \underline{x}^*\|^2\right]$$

$$\Delta = t\left[f(\underline{y}_{t+1}) - f(\underline{y}_t)\right] + 2\left[f(\underline{y}_{t+1}) - f(\underline{x}^*)\right] + \frac{t+1}{2L}\|\underline{g}_t\|^2 - 2\underline{g}_t^T(\underline{z}_t - \underline{x}^*)$$

$$\Delta \le t\left[f(\underline{x}_t) - f(\underline{y}_t)\right] + 2\left[f(\underline{x}_t) - f(\underline{x}^*)\right] - \frac{1}{2L}\|\underline{g}_t\|^2 - 2\underline{g}_t^T(\underline{z}_t - \underline{x}^*)$$

$$\le t\left[f(\underline{x}_t) - f(\underline{y}_t)\right] + 2\left[f(\underline{x}_t) - f(\underline{x}^*)\right] - 2\underline{g}_t^T(\underline{z}_t - \underline{x}^*)$$

$$\le t\,\underline{g}_t^T(\underline{x}_t - \underline{y}_t) + 2\,\underline{g}_t^T(\underline{x}_t - \underline{x}^*) - 2\underline{g}_t^T(\underline{z}_t - \underline{x}^*)$$

$$= \underline{g}_t^T\left[\underbrace{(t+2)\underline{x}_t - t\underline{y}_t - 2\underline{z}_t}_{=0}\right]$$

$= 0 \Leftarrow \underline{x}_{t+1} = \frac{t+1}{t+3}\underline{y}_{t+1} + \frac{2}{t+3}\underline{z}_{t+1}$: average

$\Rightarrow \Delta \le 0 \Rightarrow \phi(t+1) \le \phi(t)$ for each t. ∎

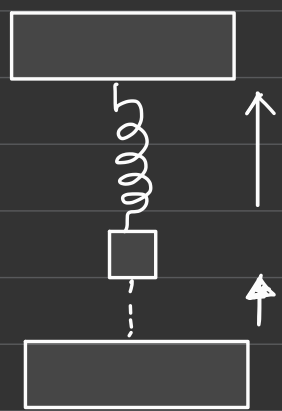
Interpretation using differential equations:

Second-order ODE:

$$\ddot{y}(\tau) + \frac{3}{\tau}\,\dot{y}(\tau) + \nabla f(y(\tau)) = 0$$

damping co-efficient

Lyapunov



$F = -\nabla f(x)$ (spring force)

$-\frac{\tau}{3}\dot{x}$ (time-varying damping)

$$\frac{t}{t+3} \approx 1 - 3/t$$

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

$$\underline{x}_{t+1} = \underline{y}_{t+1} + \frac{t}{t+3}\left(\underline{y}_{t+1} - \underline{y}_t\right)$$

$$y_{t+1} = y_t + \frac{t-1}{t+2}(y_t - y_{t-1}) - \eta \nabla f(x_t)$$

$$\Rightarrow \frac{\underline{y}_{t+1} - \underline{y}_t}{\sqrt{\eta}} = \frac{t-1}{t+2} \frac{\underline{y}_t - \underline{y}_{t-1}}{\sqrt{\eta}} - \sqrt{\eta} \nabla f(\underline{x}_t)$$

Let $t = \tau/\sqrt{\eta}$, $Y(\tau) \approx y_\tau/\sqrt{\eta} = \underline{y}_t$

and

$$Y(\tau + \sqrt{\eta}) \approx \underline{y}_{t+1}$$

Then, using Taylor expansion:

$$\frac{y_{t+1} - y_t}{\sqrt{\eta}} = \left(Y(\tau + \sqrt{\eta}) - Y(\tau)\right) \cdot 1/\sqrt{\eta}$$

$$\approx \dot{Y}(\tau) + \frac{1}{2} \ddot{Y}(\tau) \sqrt{\eta}$$

Similarly, $\frac{y_t - y_{t-1}}{\sqrt{\eta}} \approx \dot{Y}(\tau) - \frac{1}{2} \ddot{Y}(\tau) \sqrt{\eta}$

So Nestrov's acceleration

$$\dot{Y}(\tau) + \frac{\sqrt{\eta}}{2}\ddot{Y}(\tau) \approx \left(1 - \frac{3\sqrt{\eta}}{\tau}\right)\left[\dot{Y}(\tau) - \frac{\sqrt{\eta}}{2}\ddot{Y}(\tau)\right] - \sqrt{\eta}\,\nabla f(Y(\tau))$$

$$\Rightarrow \quad \ddot{Y}(\tau) + \frac{3}{\tau}\dot{Y}(\tau) + \nabla f(Y(\tau)) \approx 0$$

For this second-order ODE

$$f(Y(\tau)) - f_{opt} \leq O\left(1/\tau^2\right)$$

Actually, 3 is the smallest constant that guarantees $O(1/\tau^2)$

Let f is $L-$ smooth and $\mu-$ strongly convex, then Nestrov's accelerated gradient descent satisfies

$$f(\underline{y}_T) - f(\underline{x}^*) \leq \frac{L+\mu}{2} \exp\left(-\frac{T}{\sqrt{K}}\right) \|\underline{x}_0 - \underline{x}\|_2^2$$

with

$$t = \frac{3}{2}\left(\sqrt{K} - 1\right) \qquad K = \frac{L}{\mu}$$

$$\eta = \frac{1}{L}$$