Lec. 19 (11/29/2022)

First Order Algorithms

$\underbrace{\text{First Order}}$ only requires first derivative/gradient of the objective function, i.e., ∇f

Gradient descent: (Most well-known 1^{st} order algorithm)

$\underbrace{\text{Cauchy (1820)}}$
~ 200 yrs old!!

$$\boxed{\min_{\underline{x} \in \mathcal{X}} f(\underline{x})}$$

Suppose, $\underbrace{\mathcal{X} \equiv \mathbb{R}^n}_{\text{unconstrained}}$

$$\underbrace{\underline{x}_{k+1} = \underline{x}_k - \eta_k \nabla f(\underline{x}_k),}_{\text{Iterative algorithm}}$$

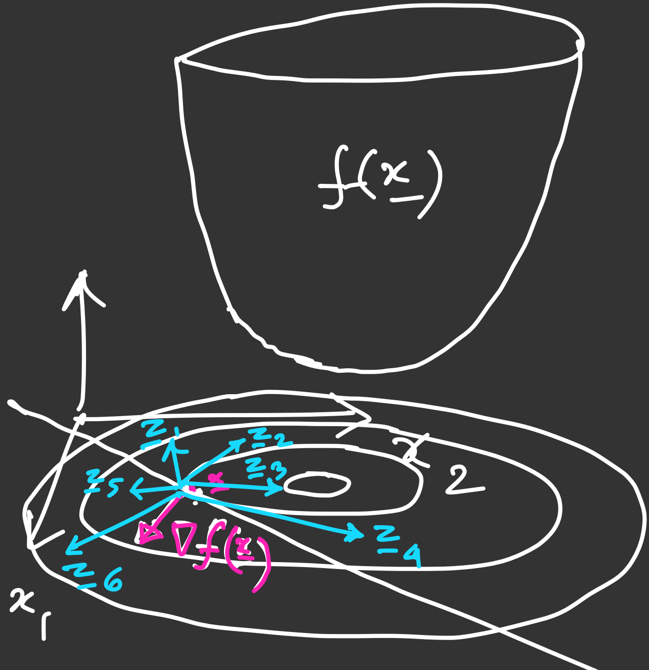
where $k = 0, 1, 2, \ldots$

k: iteration index

η_k: time step (may be constant $\eta_k \equiv \eta$)

Main idea: Want a descent:

$$\underbrace{f(\underline{x}_{k+1}) \leq f(\underline{x}_k) \quad \forall\, k = 0, 1, 2, \ldots}_{\text{An algorithm that does this is called "Descent method"}}$$



1^{st} order Taylor approximation of $f(\underline{x}+\underline{z})$ around $\underline{x}$ is

$$f(\underline{x}+\underline{z}) \approx \underbrace{\hat{f}(\underline{x}+\underline{z})}$$

$$= f(\underline{x}) + \underbrace{\langle \nabla f(\underline{x}), \underline{z} \rangle}_{D_{\underline{z}} f(\underline{x})}$$

Directional derivative of f @ $\underline{x}$ in the direction $\underline{z}$

$\therefore$ Any direction $\underline{z}$ that makes $D_{\underline{z}} f(\underline{x}) < 0$ is a descent direction $\underline{z}$

$\therefore$ Any $\underline{z}$ that makes an obtuse ($> 90°$) angle with $\nabla f(\underline{x})$ is a descent direction.

So, out of all descent directions $\underline{z}$, what is the $\underbrace{\text{"steepest descent"}}$ direction?

(makes $D_{\underline{z}} f(\underline{x})$ most negative)

$$D_{\underline{z}} f(\underline{x}) = \langle \nabla f(\underline{x}), \underline{z} \rangle$$

$\therefore$ Most negative value of $D_{\underline{z}} f(\underline{x})$ is achieved by $\underline{z} = -\nabla f(\underline{x})$

giving $\langle \nabla f(\underline{x}), -\nabla f(\underline{x}) \rangle = -\|\nabla f(\underline{x})\|_2^2$

How fast can any gradient-based algorithm / first order algorithm converge?

<u>Answer:</u> Nemirovskii - Yudin (1983)

optimal rate : $O(1/k^2)$

But gradient descent (particular 1^{st} order algorithm) does NOT achieve this optimal 1^{st} order rate. Only achieves $O(1/k)$.

Specifically, suppose $\eta_k \equiv \eta > 0$

$\boxed{\text{If}}$ ① f is convex

② f " C^1 ($\Leftrightarrow$ continuously differentiable)

③ f " Lipschitz continuous with Lipschitz constant L:

$$\Leftrightarrow \|\nabla f(\underline{x}) - \nabla f(\underline{y})\|_2 \leq L \|\underline{x} - \underline{y}\|_2$$

$\forall\, \underline{x}, \underline{y} \in \mathrm{dom}(f)$

$\boxed{\text{then}}$ with $\eta = \frac{1}{L}$, $\underbrace{\text{generated}}_{\text{using grad. descent}}$ sequence $\{\underline{x}_k\}$

next pg.

satisfies :

$$f(\underline{x}_k) - f(\underline{x}_{opt}) \leq \underbrace{\frac{L}{2k} \|\underline{x}_0 - \underline{x}_{opt}\|_2^2}_{O(1/k)}$$

If $f \in S^1_{L,m}$ $\Leftrightarrow$ ① $f \in C^1$

② f is convex

③ ∇f is Lipschitz with constant L

④ f is m-strongly convex

$\Leftrightarrow f(\underline{y}) \geq f(\underline{x}) + \langle \nabla f(\underline{x}), \underline{x} - \underline{y} \rangle + \frac{m}{2} \|\underline{x} - \underline{y}\|_2^2$

extra assumption

Then $\underbrace{\eta_{opt}}_{} = \frac{2}{L+m}$

optimal constant stepsize

Still $O(1/k)$ convergence

Gradient descent (1820, Cauchy)	Heavy Ball (1964, Polyak)	Nesterov's accelerated Grad. descent (1983)
$\underline{x}_{k+1} = \underline{x}_k - \eta_k \nabla f(\underline{x}_k)$	$\underline{y}_{k+1} = \underline{x}_k - \eta_k \nabla f(\underline{x}_k)$ $\underline{x}_{k+1} = \underline{y}_{k+1} + \alpha_k (\underline{x}_k - \underline{x}_{k-1})$ $\Updownarrow$ $\underline{x}_{k+1} = \underline{x}_k - \eta_k \nabla f(\underline{x}_k) + \underbrace{\alpha_k (\underline{x}_k - \underline{x}_{k-1})}_{\text{momentum term}}$	$\underline{y}_{k+1} = \underline{x}_k - \eta_k \nabla f(\underline{x}_k)$ $\underline{x}_{k+1} = \underline{y}_{k+1} + \alpha_k (\underline{y}_{k+1} - \underline{y}_k)$
Descent method $\mathcal{O}(1/k)$		

NOT descent method but may achieve $\mathcal{O}(1/k^2)$

Heavy ball is NOT globally convergent even for strongly convex function.

In the domain in which it converges, it can beat gradient descent by achieving $O(1/k^2)$.

Handling Constraints

$$\boxed{\min_{\underline{x} \in \mathcal{X}} f(\underline{x})}$$

① **Projected Gradient Descent:**

$$\underline{x}_{k+1} = \text{proj}_{\mathcal{X}}^{\|\cdot\|_2}\left(\underline{x}_k - \eta_k \nabla f(\underline{x}_k)\right)$$

$$\Updownarrow$$

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

$$\underline{x}_{k+1} = \text{proj}_{\mathcal{X}}^{\|\cdot\|_2}(\underline{y}_{k+1})$$

$$= \underset{\underline{x} \in \mathcal{X}}{\text{argmin}} \frac{1}{2}\|\underline{x} - \underline{y}_{k+1}\|_2^2$$

Next pg.

② Mirror Descent:

Constructs a "mirror function" $\underbrace{}$

strictly convex function ψ related to the geometry of the constraint set $\mathcal{X}$

$$\begin{cases} \nabla\psi(\underline{y}_{k+1}) = \nabla\psi(\underline{x}_k) - \eta_k \nabla f(\underline{x}_k) \\ \underline{x}_{k+1} = \text{proj}_{\mathcal{X}}^{D_\psi}(\underline{y}_{k+1}) \end{cases}$$

where D_ψ is the Bregman divergence induced by the mirror map/function ψ

Again,

$$\mathrm{proj}^{D_\psi}_{\mathcal{X}}(\underline{\eta}) := \underbrace{\operatorname*{argmin}_{\underline{\xi} \in \mathcal{X}} D_\psi(\underline{\xi}, \underline{\eta})}_{\text{projection of } \underline{\eta} \text{ onto the set } \mathcal{X} \text{ w.r.t. } D_\psi}$$

projection of $\underline{\eta}$ onto the set $\mathcal{X}$

w.r.t. D_ψ

not necessarily Euclidean distance