OPTIMIZATION (SI 416) – LECTURE 4

Harsha Hutridurga

IIT Bombay

♣ $f : \mathbb{R}^n \to \mathbb{R}$ is said to be STRONGLY CONVEX if $\exists\, \lambda > 0$ such that

$$f(x) - \lambda \|x\|^2 \quad \text{is convex.}$$

$$\boxed{\begin{array}{ccc} f \text{ is} & & \text{for every } x \in \mathbb{R}^n, \\ \text{strongly convex} & \Longrightarrow & \left\langle \nabla^2 f(x) p, p \right\rangle \geq 2\lambda \|p\|^2 \quad \text{for all } p \in \mathbb{R}^n \end{array}}$$

$$\boxed{\begin{array}{ccc} f \text{ is} & & \text{for every } x, y \in \mathbb{R}^n, \\ \text{strongly convex} & \Longrightarrow & f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \lambda \|y - x\|^2 \end{array}}$$

♣ f is strongly implies $\Longrightarrow$ for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$ we have

$$\boxed{f(\alpha x + (1-\alpha) y) \quad \leq \quad \alpha f(x) + (1-\alpha) f(y) - \lambda \alpha (1-\alpha) \|x - y\|^2}$$

$$\boxed{\begin{array}{ccccc} f \text{ is} & & f \text{ is} & & f \text{ is} \\ \text{strongly convex} & \Longrightarrow & \text{strictly convex} & \Longrightarrow & \text{convex} \end{array}}$$

Theorem

Every convex function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous on $\mathbb{R}^n$.

- ♣ Continuous functions on compact sets attain extremal values
- ♣ More precisely, for a continuous function f there exist points x_* and $\hat{x}$ in K s.t.

$$f(x_*) = \min_{x \in K} f(x) \quad \text{and} \quad f(\hat{x}) = \max_{x \in K} f(x)$$

where K is a compact set in $\mathbb{R}^n$

- Recall that for a strongly convex function f, we have for $\alpha \in [0,1]$,
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \lambda\alpha(1-\alpha)\|x-y\|^2$$

- By picking $x \gg 1$, $y = 0$ and $\alpha = \frac{1}{\|x\|}$ in the above inequality yields
$$f\left(\frac{x}{\|x\|}\right) \leq \frac{1}{\|x\|} f(x) + \left(1 - \frac{1}{\|x\|}\right) f(0) - \frac{\lambda}{\|x\|}\left(1 - \frac{1}{\|x\|}\right)\|x\|^2$$

- Rearranging the above inequality yields
$$f(x) \geq \|x\| \left(f\left(\frac{x}{\|x\|}\right) - \left(1 - \frac{1}{\|x\|}\right) f(0) + \lambda(\|x\| - 1) \right)$$

- Observe that the following quantities take finite values, thanks to continuity:
$$f\left(\frac{x}{\|x\|}\right) \quad \text{and} \quad f(0)$$

- Recall that we had

$$f(x) \geq \|x\| \left(f\left(\frac{x}{\|x\|}\right) - \left(1 - \frac{1}{\|x\|}\right) f(0) + \lambda(\|x\| - 1) \right)$$

- Hence $f(x)$ goes to infinity as $\|x\|$ goes to infinity
- Therefore we can take a ball of large enough radius such that f attains its minimum in that ball
- Continuity of f tells us that there is at least a point in the closure of that ball where the minimum is attained
- As every strongly convex function is also strictly convex, there should be a unique point of minimum

Theorem

If $f : \mathbb{R}^n \to \mathbb{R}$ is a strongly convex function, then there exists a unique point $x_* \in \mathbb{R}^n$ such that

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

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex continuously differentiable function
- This is equivalent to having

$$f(y) \geq f(x) + \langle \nabla f(x), (y-x) \rangle \qquad \text{for all } x, y \in \mathbb{R}^n$$

- Interchanging the roles of x and y in the above inequality:

$$f(x) \geq f(y) + \langle \nabla f(y), (x-y) \rangle$$

- Adding the above inequalities results in

$$\begin{aligned} 0 &\geq \langle \nabla f(x), (y-x) \rangle + \langle \nabla f(y), (x-y) \rangle \\ &\qquad \Longrightarrow \left\langle \Big(\nabla f(x) - \nabla f(y) \Big), (x-y) \right\rangle \geq 0 \end{aligned}$$

$$\boxed{\begin{array}{ccc} f \text{ is} & & \text{for every } x, y \in \mathbb{R}^n, \\ \text{convex} & \Longrightarrow & \left\langle \Big(\nabla f(x) - \nabla f(y) \Big), (x-y) \right\rangle \geq 0 \end{array}}$$

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable s.t.

$$\left\langle \left(\nabla f(x) - \nabla f(y) \right), (x - y) \right\rangle \geq 0 \quad \text{for all } x, y \in \mathbb{R}^n.$$

- Let x and y be arbitrarily fixed. Define $g : [0, \infty) \to \mathbb{R}$ as follows:

$$g(\alpha) := f(x + \alpha(y - x)) \quad \text{for } \alpha \in [0, \infty)$$

- Thanks to the above monotonicity property of the gradient,

$$\left\langle \left(\nabla f(x + \alpha(y - x)) - \nabla f(x) \right), \left(\alpha(y - x) \right) \right\rangle \geq 0$$

- As $\alpha \geq 0$, we deduce that

$$\langle \nabla f(x + \alpha(y - x)), (y - x) \rangle \geq \langle \nabla f(x), (y - x) \rangle \implies g'(\alpha) \geq g'(0)$$

- Note further that $g(0) = f(x)$ and $g(1) = f(y)$

- Employing fundamental theorem of calculus, we have

$$g(1) - g(0) = \int_0^1 g'(\alpha)\,\mathrm{d}\alpha \geq g'(0)$$

  where we have used the observation from earlier
- Writing the above inequality in terms of f, we obtain

$$f(y) - f(x) \geq \langle \nabla f(x), (y - x) \rangle$$

- As x and y were arbitrary, we get convexity of f

$$\boxed{\begin{array}{ccc} f \text{ is} & & \text{for every } x, y \in \mathbb{R}^n, \\ \text{convex} & \Longleftarrow & \left\langle \Big(\nabla f(x) - \nabla f(y)\Big), (x - y) \right\rangle \geq 0 \end{array}}$$

♣ Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex, i.e.

$$g(x) := f(x) - \lambda\|x\|^2 \quad \text{is convex for some } \lambda > 0$$

♣ Hence the gradient of g must be monotone, i.e.

$$\left\langle \Big(\nabla g(x) - \nabla g(y)\Big), (x-y) \right\rangle \geq 0 \qquad \text{for all } x, y \in \mathbb{R}^n$$

♣ Writing the above inequality in terms of f, we obtain

$$\begin{aligned} &\left\langle \Big(\nabla f(x) - 2\lambda x - \nabla f(y) + 2\lambda y\Big), (x-y) \right\rangle \geq 0 \\ &\qquad\qquad \Longrightarrow \left\langle \Big(\nabla f(x) - \nabla f(y)\Big), (x-y) \right\rangle \geq 2\lambda\|x-y\|^2 \end{aligned}$$

$$\boxed{\begin{array}{ccc} f \text{ is} & & \text{for every } x, y \in \mathbb{R}^n, \\ \text{strongly convex} & \Longleftrightarrow & \left\langle \Big(\nabla f(x) - \nabla f(y)\Big), (x-y) \right\rangle \geq 2\lambda\|x-y\|^2 \end{array}}$$

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex
- This is equivalent to having a constant $\lambda > 0$ s.t.

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \lambda \|y - x\|^2 =: Q(x,y) \qquad \text{for all } x, y \in \mathbb{R}^n$$

- Recall that f has a unique global minimizer, say x_*
- Let us minimize the above inequality in the y variable:

$$\min_{y \in \mathbb{R}^n} f(y) \geq \min_{y \in \mathbb{R}^n} Q(x,y)$$

$$\text{Note that} \qquad \min_{y \in \mathbb{R}^n} f(y) = f(x_*)$$

- Observe that $Q(x,y)$ is a quadratic function in y variable:

$$Q(x,y) = f(x) + \langle \nabla f(x), y - x \rangle + \lambda \langle y - x, y - x \rangle$$

- Remark that $Q(x,y)$ has a unique minimizer, say y_*

- The first order necessary condition says

$$\nabla_y Q(x, y_*) = 0 \implies \nabla f(x) + 2\lambda(y_* - x) = 0$$
$$\implies y_* = x - \frac{1}{2\lambda}\nabla f(x)$$

- Note that

$$Q(x, y_*) = f(x) - \frac{1}{2\lambda}\left\|\nabla f(x)\right\|^2 + \frac{1}{4\lambda}\left\|\nabla f(x)\right\|^2 = f(x) - \frac{1}{4\lambda}\left\|\nabla f(x)\right\|^2$$

- Hence we obtain

$$f(x_*) \geq f(x) - \frac{1}{4\lambda}\left\|\nabla f(x)\right\|^2 \implies f(x) - f(x_*) \leq \frac{1}{4\lambda}\left\|\nabla f(x)\right\|^2$$

- This is referred to as the PL condition
- It essentially says: f cannot grow too fast near its minimizer
- It is named after Polyak and Lojasiewicz

Definition

Let $\beta > 0$. A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be β-smooth if

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\| \qquad \textit{for all } x, y \in \mathbb{R}^n$$

- The gradient ∇f is also said to be Lipschitz continuous
- Let x and y be arbitrarily fixed
- Note that we have

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + \alpha(y - x)), (y - x) \rangle \, \mathrm{d}\alpha$$

- We rewrite the above equality as

$$f(y) - f(x) - \langle \nabla f(x), (y - x) \rangle = \int_0^1 \Big(\nabla f(x + \alpha(y - x)) - \nabla f(x) \Big) \cdot (y - x) \, \mathrm{d}\alpha$$

♣ Employing Cauchy-Schwarz inequality, we obtain

$$
\begin{aligned}
f(y) - f(x) - \nabla f(x) \cdot (y - x) &= \int_0^1 \Big(\nabla f(x + \alpha(y - x)) - \nabla f(x)\Big) \cdot (y - x \\
&\le \int_0^1 \|\nabla f(x + \alpha(y - x)) - \nabla f(x)\| \, \|y - x\| \, \mathrm{d}\alpha \\
&\le \int_0^1 \beta\alpha \|y - x\|^2 \, \mathrm{d}\alpha
\end{aligned}
$$

where we have used the β-smoothness property.

♣ Hence we deduce

$$
f(y) \le f(x) + \langle \nabla f(x), (y - x) \rangle + \frac{\beta}{2}\|y - x\|^2
$$

$$
\boxed{\begin{array}{ccc} f \text{ is} & & \text{for every } x, y \in \mathbb{R}^n, \\ \beta\text{-smooth} & \Longrightarrow & f(y) \le f(x) + \langle \nabla f(x), (y - x) \rangle + \frac{\beta}{2}\|y - x\|^2 \end{array}}
$$

STRONGLY CONVEX AND β-SMOOTH

- For a function $f : \mathbb{R}^n \to \mathbb{R}$ which is both strongly convex and β-smooth, we have for all $x, y \in \mathbb{R}^n$,

$$f(x) + \langle \nabla f(x), y - x \rangle + \lambda \|y - x\|^2$$

$$\leq f(y)$$

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

- i.e. it is sandwiched between two quadratic functions

- ♣ Let $f : \mathbb{R}^n \to \mathbb{R}$ be β-smooth
- ♣ Pick a point $x^{(0)} \in \mathbb{R}^n$ and define a new point $x^{(1)}$ as follows

$$x^{(1)} := x^{(0)} - \delta \nabla f(x^{(0)}) \qquad \text{for some } \delta > 0.$$

- ♣ Employ the earlier inequality for β-smooth functions with points $x^{(0)}$ and $x^{(1)}$:

$$\begin{aligned} f(x^{(1)}) &\le f(x^{(0)}) + \left\langle \nabla f(x^{(0)}), (-\delta \nabla f(x^{(0)})) \right\rangle + \frac{\beta\delta^2}{2} \left\| \nabla f(x^{(0)}) \right\|^2 \\ &= f(x^{(0)}) - \delta \left\| \nabla f(x^{(0)}) \right\|^2 + \frac{\beta\delta^2}{2} \left\| \nabla f(x^{(0)}) \right\|^2 \end{aligned}$$

- ♣ Suppose that we take $\delta < \frac{1}{\beta}$. Then

$$f(x^{(1)}) \le f(x^{(0)}) - \frac{\delta}{2} \left\| \nabla f(x^{(0)}) \right\|^2 \le f(x^{(0)})$$

- ♣ This hints at a recipe for building a minimizing sequence

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be β-smooth
- Take $0 < \delta \leq \frac{1}{\beta}$
- Pick a point $x^{(0)} \in \mathbb{R}^n$
- Define a sequence of points $x^{(n)}$ iteratively as follows:

$$x^{(n+1)} := x^{(n)} - \delta \nabla f(x^{(n)}) \qquad \text{for } n = 0, 1, 2, \ldots$$

- $x^{(n)}$ is referred to as the n^{th} iterate
- The above algorithm for generating $x^{(n)}$ is called gradient descent
- Exploiting the β-smoothness property, we deduce

$$f(x^{(n+1)}) \leq f(x^{(n)}) - \frac{\delta}{2} \left\| \nabla f(x^{(n)}) \right\|^2 \leq f(x^{(n)}) \qquad \text{for } n = 0, 1, 2, \ldots$$

- So far, we have used only β-smoothness of f
- Let us now assume that f is strongly convex as well
- We have from earlier that

$$f(x^{(n+1)}) \leq f(x^{(n)}) - \frac{\delta}{2} \left\| \nabla f(x^{(n)}) \right\|^2 \qquad \text{for } n = 0, 1, 2, \ldots$$

- Recall: for a convex function f and for any $x, y \in \mathbb{R}^n$, we have

$$f(y) \geq f(x) + \langle \nabla f(x), (y - x) \rangle$$

- Strong convexity of f gives a unique minimizer x_* of f
- Applying the above inequality for $y = x_*$ and $x = x^{(n)}$ yields

$$f(x_*) \geq f(x^{(n)}) + \left\langle \nabla f(x^{(n)}), (x_* - x^{(n)}) \right\rangle$$

$$\implies f(x^{(n)}) \leq f(x_*) - \left\langle \nabla f(x^{(n)}), (x_* - x^{(n)}) \right\rangle$$

- Hence we obtain

$$f(x^{(n+1)}) \leq f(x_*) - \left\langle \nabla f(x^{(n)}), (x_* - x^{(n)}) \right\rangle - \frac{\delta}{2} \left\| \nabla f(x^{(n)}) \right\|^2$$

- Observe that

$$- \left\langle \nabla f(x^{(n)}), (x_* - x^{(n)}) \right\rangle - \frac{\delta}{2} \left\| \nabla f(x^{(n)}) \right\|^2 = \frac{1}{2\delta} \left(\left\| x^{(n)} - x_* \right\|^2 - \left\| x^{(n)} - x_* - \delta \nabla f(x^{(n)}) \right\|^2 \right)$$

- Hence we deduce that

$$f(x^{(n+1)}) \leq f(x_*) + \frac{1}{2\delta} \left(\left\| x^{(n)} - x_* \right\|^2 - \left\| x^{(n+1)} - x_* \right\|^2 \right)$$

- We can rewrite the above inequality as

$$f(x^{(n+1)}) - f(x_*) \leq \frac{1}{2\delta} \left(\left\| x^{(n)} - x_* \right\|^2 - \left\| x^{(n+1)} - x_* \right\|^2 \right)$$

- Recall: for a strongly convex function f and for any $x, y \in \mathbb{R}^n$,

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \lambda \|y - x\|^2$$

- Applying the above inequality for $y = x^{(n+1)}$ and $x = x_*$ yields

$$\begin{aligned} f(x^{(n+1)}) &\geq f(x_*) + \nabla f(x_*) \cdot (x^{(n+1)} - x_*) + \lambda \left\| x^{(n+1)} - x_* \right\|^2 \\ &= f(x_*) + \lambda \left\| x^{(n+1)} - x_* \right\|^2 \end{aligned}$$

where we have used the fact that $\nabla f(x_*) = 0$

- Hence we deduce that

$$\lambda \left\| x^{(n+1)} - x_* \right\|^2 \leq \frac{1}{2\delta} \Big(\left\| x^{(n)} - x_* \right\|^2 - \left\| x^{(n+1)} - x_* \right\|^2 \Big)$$

- We can rewrite the above inequality as

$$(1 + 2\delta\lambda) \left\| x^{(n+1)} - x_* \right\|^2 \leq \left\| x^{(n)} - x_* \right\|^2 \qquad \text{for } n = 0, 1, \dots$$

- Recall that we have

$$\left\| x^{(n+1)} - x_* \right\|^2 \le \frac{1}{1+2\delta\lambda} \left\| x^{(n)} - x_* \right\|^2 \qquad \text{for } n = 0, 1, \dots$$

- This translates to

$$\left\| x^{(n)} - x_* \right\|^2 \le \left(\frac{1}{1+2\delta\lambda} \right)^n \left\| x^{(0)} - x_* \right\|^2 \qquad \text{for } n = 0, 1, \dots$$

- Let us denote the error committed at the n^{th} iteration as e_n, i.e.

$$e_n := \left\| x^{(n)} - x_* \right\|^2 \qquad \text{for } n = 0, 1, 2, \dots$$

- Note that we have

$$e_n \le \left(\frac{1}{1+2\delta\lambda} \right)^n e_0 \qquad \text{for } n = 0, 1, 2, \dots$$

- Suppose we have a tolerance of $\varepsilon > 0$, i.e we are looking for $x^{(n)}$ such that the error falls below ε, i.e.

$$e_n \le \varepsilon$$

- Note that this can be ensured by taking n such that

$$\left(\frac{1}{1+2\delta\lambda}\right)^n e_0 \le \varepsilon$$

- This is the same as having

$$n \ln\left(\frac{1}{1+2\delta\lambda}\right) \le \ln\left(\frac{\varepsilon}{e_0}\right) \implies n \ge \frac{1}{\ln(1+2\delta\lambda)} \ln\left(\frac{e_0}{\varepsilon}\right)$$

- This suggests that we should take the number of iterations

$$n = \mathcal{O}(\ln(\varepsilon^{-1}))$$

END OF LECTURE 4

THANK YOU FOR YOUR ATTENTION