Class #15

Tuesday, March 11, 2025 10:23 AM

Implications of Strongly Convex $f \in C^2(\mathbb{R}^n)$

twice continuously differentiable

① The Suboptimality gap $f(x) - p^*$ can be bounded in terms of the norm of the gradient $\|\nabla f(x)\|_2$.

Since $f \in C^2(\mathbb{R}^n)$ and is strongly convex with parameter m

$$f(y) \geq \underbrace{f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2}_{\tilde{f}(y)}$$

$\frac{m}{2}\|y-x\|_2^2 \to \frac{m}{2}(y-x)^T(y-x)$

$\forall y \in \operatorname{dom} f$

$\Rightarrow$ Quadratic function which is strongly convex itself.

It has a unique minimizer.

$\hookrightarrow \nabla\tilde{f}(y^*) = 0 \iff y^*$ is the unique minimizer

$$\nabla\tilde{f}(y) = \nabla f(x) + m(y-x)$$

$$\nabla\tilde{f}(y^*) = 0 \iff \boxed{y^* = x - \frac{1}{m}\nabla f(x)}$$

Since $f(y) \geq \tilde{f}(y) \quad \forall y \in \operatorname{dom} f$

$$\Rightarrow f(y) \geq \tilde{f}(y) \geq \tilde{f}(y^*)$$

$$\Rightarrow f(y) \geq \tilde{f}(y^*)$$

$$\Rightarrow f(y) \geq f(x) - \frac{1}{m}\|\nabla f(x)\|_2^2 + \frac{1}{2m}\|\nabla f(x)\|_2^2$$

$$\Rightarrow f(y) \geq f(x) - \frac{1}{\quad}\|\nabla f(x)\|_2^2$$

$$\Rightarrow \quad f(y) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2$$

Since this is true for all $y \in \operatorname{dom} f$

Put $y = x^*$

$$\Rightarrow \quad \underbrace{f(x^*)}_{p^*} \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2$$

$$\Rightarrow \quad \boxed{f(x) - p^* \leq \frac{1}{2m} \|\nabla f(x)\|_2^2}$$

<u>Question</u>: We want $f(x) - p^* \leq \epsilon$. When should we stop gradient descent?

<u>Ans</u>: If we have $\frac{1}{2m} \|\nabla f(x)\|_2^2 \leq \epsilon$

then $f(x) - p^* \leq \epsilon$

$$\Rightarrow \quad \|\nabla f(x)\|_2^2 \leq 2m\epsilon$$

$$\boxed{\|\nabla f(x)\|_2 \leq \sqrt{2m\epsilon}}$$

<u>e.g.</u> $\epsilon = 10^{-8}$

$\Rightarrow$ Stop GD when $\|\nabla f(x)\|_2 \leq \sqrt{2m}\, 10^{-4}$.

<u>Challenge</u>: Requires knowledge of m.

<u>Still</u>: Requiring $\|\nabla f(x)\|_2$ to be small enough is a good stopping criterion.

<u>What about the distance of x from x^*</u>?

$$f(y) \geq \tilde{f}(y) \quad \forall \; y \in \operatorname{dom} f$$

Take $y = x^*$

Take $y = x^*$

$$f(x^*) \geq f(x) + \underbrace{\nabla f(x)^T (x^* - x)}_{\geq -\left|\nabla f(x)^T (x^* - x)\right|} + \frac{m}{2}\|x^* - x\|_2^2$$

$$\geq f(x) \underbrace{- \left|\nabla f(x)^T (x^* - x)\right|}_{\geq -\|\nabla f(x)\|_2 \|x^* - x\|_2} + \frac{m}{2}\|x^* - x\|_2^2$$

$$\Rightarrow \quad f(x^*) \geq f(x) - \underbrace{\|\nabla f(x)\|_2 \|x^* - x\|_2 + \frac{m}{2}\|x^* - x\|_2^2}$$

since $f(x^*) \leq f(x)$

$$\|\nabla f(x)\|_2 \|x^* - x\|_2 - \frac{m}{2}\|x^* - x\|_2^2 \geq f(x) - f(x^*) \geq 0$$

$$\Leftrightarrow \quad \frac{m}{2}\|x^* - x\|_2^{\cancel{2}} \leq \|\nabla f(x)\|_2 \cancel{\|x^* - x\|_2}$$

$$\Leftrightarrow \quad \boxed{\|x^* - x\| \leq \frac{2}{m}\|\nabla f(x)\|_2}$$

Looking at previous example, which set

$$\|\nabla f(x)\|_2 \leq \sqrt{2m\epsilon}$$

$$\Rightarrow \quad \|x^* - x\|_2 \leq \frac{2}{m} \times \sqrt{2m\epsilon} = \frac{2\sqrt{2}}{\sqrt{m}}\sqrt{\epsilon}.$$

Condition number of an optimization Problem

and regularity of objective function around

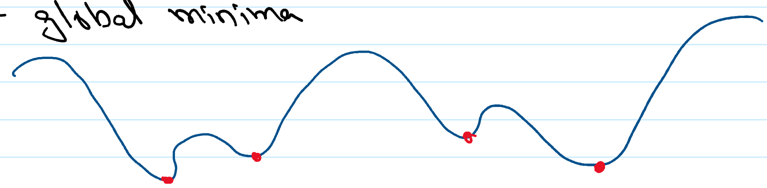
Condition number of an optimization problem and regularity of objective functions around local minima

Let $f \in C^2(\mathbb{R}^n)$

↳ Twice continuously differentiable

Let x^* be a local minimum of f

$f(x^*) = p^*$ → not global minima



Let's look at second-order approximation of f around this x^*.

$= 0$ b/c $\nabla f(x^*) = 0$

$$f(y) \approx f(x^*) + \cancel{\nabla f(x^*)^T (y - x^*)} + \frac{1}{2}(y-x^*)^T \nabla^2 f(x^*)(y-x^*)$$

$\forall\ y \in$ small neighborhood of x^*

↳ i.e. $\exists\ \epsilon > 0\ !\ y \in \underbrace{\{x : \|x - x^*\|_2 \le \epsilon\}}_{B_\epsilon(x^*)}$

$$\Rightarrow f(y) \approx \underbrace{p^* + \frac{1}{2}(y-x^*)^T \nabla^2 f(x^*)(y-x^*)}_{\text{Quadratic function in } y}$$

↳ plus if $\nabla^2 f(x^*) \succeq mI$ then it is strongly convex quadratic

<u>Conclusion:</u> Every "nice" local minimum of an $f \in C^2(\mathbb{R}^n)$ looks like a strongly convex quadratic function in a small enough neighborhood of the local minimum.

How difficult is it to go to the local minimum? (or global minimum in a strongly convex problem).

The difficulty of an optimization problem is determined by the condition number of the Hessian (in local neighborhood or more globally for strongly convex functions).

<u>Recall</u>: If f is m-strongly convex and $f \in C^2(\mathbb{R}^n)$

① $\nabla^2 f(x) \succeq mI$

$$\Leftrightarrow \quad \lambda_{min}(\nabla^2 f(x)) \geq m \quad \forall\, x \in \text{dom} f$$

② $\nabla^2 f(x) \preceq MI$ for some $M \geq m$

$$\Leftrightarrow \quad \lambda_{max}(\nabla^2 f(x)) \leq M \quad \forall\, x \in \{x \in \text{dom} f : f(x) \leq f(x^{(0)})\}$$

In the case of nonconvex functions, we replace these with statements in the neighborhood of the local minimum.

$$\boxed{\frac{M}{m} = K \geq \frac{\lambda_{max}(\nabla^2 f(x))}{\lambda_{min}(\nabla^2 f(x))}} \quad \forall\, x \in \text{ in a neighborhood around } x^*$$

condition number of optimization problem.

<u>Recall</u>:

$$f(y) \approx p^* + \frac{1}{2}(y - x^*)^T \nabla^2 f(x^*)(y - x^*)$$

What do the sublevel sets of $f(y)$ look like

What do the Sublevel Sets of $f(y)$ look like around x^* for some $\alpha \geq p^*$

(what about $\alpha < p^*$? $\Rightarrow$ Empty set)

$$C_\alpha := \{ y : f(y) \leq \alpha \}$$

$$f(y) \leq \alpha \iff p^* + \frac{1}{2}(y-x^*)^T \nabla^2 f(x^*)(y-x^*) \leq \alpha$$

$$\iff (y-x^*)^T \nabla^2 f(x^*)(y-x^*) \leq 2(\alpha - p^*)$$

$$\iff C_\alpha := \{ y : \underbrace{(y-x^*)^T \nabla^2 f(x^*)(y-x^*)}_{z^T Q z \text{ with } Q \text{ pos. def.}} \leq 2(\alpha - p^*) \}$$

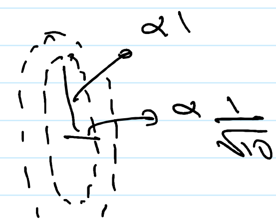
$\rightarrow$ This is the equation an ellipsoid in $\mathbb{R}^n$.

e.g., If $Q = I \Rightarrow z^T z \leq \tilde{\alpha}$

$\iff \|z\|_2^2 \leq \tilde{\alpha}$

When $Q \neq I \Rightarrow$ Ellipsoid has axes aligned with the eigenvectors of Q and the axis lengths are proportional to $\frac{1}{\sqrt{\text{eigenvalue}}}$.

e.g., $z^T \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix} z \Rightarrow$



The difficulty of an optimization problem is determined by the shape of the sublevel sets. The more 'squished' the ellipsoid is, the more challenging it is for the algorithm to converge.

↳ This is determined by $K = \frac{M}{m}$

$K = 1 \Rightarrow$ we have spheres

$K \uparrow \infty \Rightarrow$ The minor axis is collapsing.

BV: 9.1.2