Class #9

Tuesday, February 18, 2025 10:20 AM

When $f \in C^1_L(\mathbb{R}^n)$ and $t^{(k)} = \frac{1}{L}$

$$\|\nabla f(x^{(k)})\|_2^2 \longrightarrow 0 \quad \text{as} \quad k \to \infty$$

$$\nabla f(x^{(k)}) \longrightarrow 0$$

What about the case of variable step sizes?

① Decaying step size policy

② Step size is bounded below;

Let $\epsilon > 0$ be a fixed constant

$$\underline{\epsilon} \leq t^{(k)} \leq \frac{2-\epsilon}{L}$$

↳ Same proof works.

What about the rate of convergence?

$f \in C^1_L(\mathbb{R}^n)$; $t^{(k)} = \frac{1}{L}$

↳ From the previous lecture:

(★) $$\|\nabla f(x^{(k)})\|_2^2 \leq 2L\left[f(x^{(k)}) - f(x^{(k+1)})\right]$$

Sum (★) from $k=1$ to $k=K$

$$\sum_{k=1}^{K} \|\nabla f(x^{(k)})\|_2^2 \leq 2L \sum_{k=1}^{K} \left[\underbrace{f(x^{(k)}) - f(x^{(k+1)})}\right]$$

$$\leq 2L \underbrace{\sum_{k=0}^{K} \overbrace{\left[f(x^{(k)}) - f(x^{(k+1)})\right]}^{>0}}_{\text{Telescoping sum}}$$

$$\leq 2L \left(f(x^{(0)}) - \underbrace{f(x^{(K+1)})}_{\geq p^*}\right)$$

$$(\star\star) \quad \sum_{k=1}^{K} \|\nabla f(x^{(k)})\|_2^2 \leq 2L\left(f(x^{(0)}) - p^*\right)$$

$$\sum_{k=1}^{K} \|\nabla f(x^{(k)})\|_2^2 \geq K \min_{k \in \{1,2,\dots,K\}} \|\nabla f(x^{(k)})\|_2^2$$

$$K \min_{k\in\{1,\dots,K\}} \|\nabla f(x^{(k)})\|_2^2 \leq \underbrace{2L\left(f(x^{(0)}) - p^*\right)}_{\gamma > 0}$$

$$\min_{k\in\{1,\dots,K\}} \|\nabla f(x^{(k)})\|_2^2 \leq \frac{\gamma}{K}$$

Within K iterations, we will have at least one $x^{(k)}$ such that $\|\nabla f(x^{(k)})\|_2^2 \leq \boxed{\frac{\gamma}{K}} = O\left(\frac{1}{K}\right)$

Suppose we want $\|\nabla f(x^{(k)})\|_2^2 \leq \epsilon$ for ϵ very small

$$\Rightarrow \quad \frac{\gamma}{K} \leq \epsilon \quad \Rightarrow \quad K \geq \frac{\gamma}{\epsilon}$$

$$\Rightarrow \quad K = \Omega(\epsilon^{-1})$$

Say $\epsilon = 10^{-8}$

$\Rightarrow K = O(10^8)$ iterations

Similar results hold for step size choices for general descent methods, where the step size depends on the descent direction and is strictly lower bounded by $\epsilon > 0$.

Another Interpretation of Gradient Descent for $f \in C^1_L(\mathbb{R}^n)$

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|_2^2$$

Let us derive an iterative in which x is current iterate and $y = x^+$ is the next iterate

$$f(x^+) \leq f(x) + \nabla f(x)^T (x^+ - x) + \frac{L}{2}\|x^+ - x\|_2^2$$

we need x^+ such that $f(x^+)$ is as small as possible.

proximity term

Mirror Descent (when this is replaced by another $\|\cdot\|$)

we approach this by minimizing the upper bound

$$\tilde{f}(x^+) = f(x) + \nabla f(x)^T (x^+ - x) + \frac{L}{2}\underbrace{\|x^+ - x\|_2^2}_{(x^+-x)^T(x^+-x)} \quad \text{w.r.t. } x^+$$

$$= x^{+T}x^+ - x^{+T}x - x^{+T}x + x^Tx$$

$$= x^{+T}x^+ - 2x^{+T}x + x^Tx$$

$$\nabla_{x^+}\tilde{f}(x^+) = 0 + \nabla f(x) + \frac{L}{2}\left(2x^+ - 2x + 0\right)$$

$$= \nabla f(x) + L(x^+ - x) = 0$$

$$L(x^+ - x) = -\nabla f(x)$$

$$\boxed{x^+ = x - \frac{1}{L}\nabla f(x)}$$

Stepsize selection when $f \in C^1_L(\mathbb{R}^n)$ but L is not know or computing it is too expensive.

$\Downarrow$

Inexact line search.

Backtracking | Armijo rule | Armijo-Goldstein step

Another approach based on Wolfe conditions, but they are harder to compute and we won't study them.

$$\tilde{f}(t) = f(x^{(k)} + t\,\Delta x^{(k)}) \; ; \quad t \geq 0$$

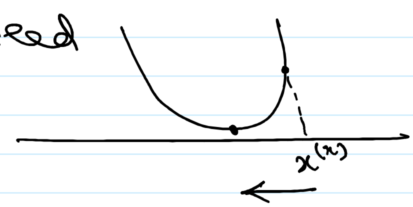
$$t = 0 \Rightarrow f(x^{(k)})$$

Inexact search line search requires finding a value of $t^{(k)}$ such that

$$\tilde{f}(t^{(k)}) = f(x^{(k)} + t^{(k)}\Delta x^{(k)}) \text{ is}$$

sufficiently smaller than $f(x^{(k)})$

but there has to be a guaranteed decrease.



Algorithm (Backtracking)

Input: Current iterate x
search direction Δx

Parameters $\alpha \in (0, 0.5) \rightarrow$ Sufficient decrease parameter
$\beta \in (0, 1)$

Initialize: $t \leftarrow 1$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$

$t \leftarrow \beta t$

Sufficient decrease condition

The algorithm ends when $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$

< 0

$\hookrightarrow$ It depends on α.

Backtracking

$t = 1$

$t = \beta$

$t = \beta^2$

$\beta \Rightarrow$ The gridding of $(0, 1]$

$\hookrightarrow$ larger β can slow down the line search

Smaller β can end up giving you a very

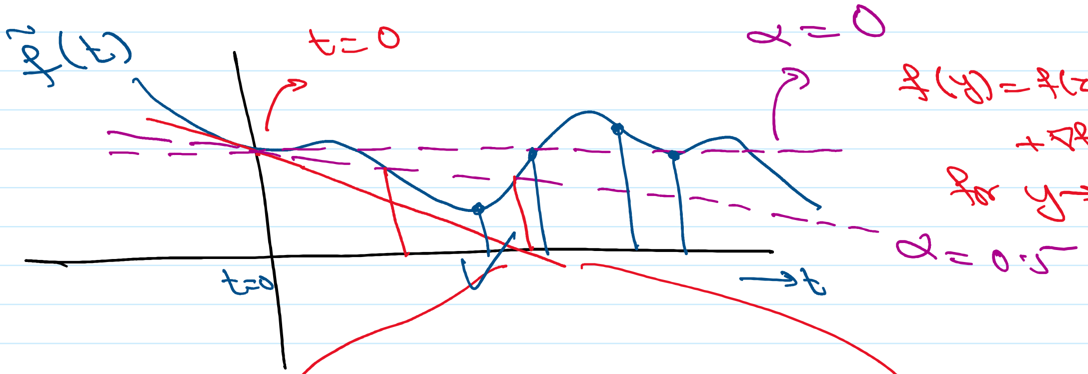
Small step size.

Geometric View of Backtracking

for t small enough

$$\tilde{f}(t) = f(x + t\Delta x) \simeq f(x) + t\nabla f(x)^T \Delta x$$

$$\tilde{f}'(t) = \nabla f(x)^T \Delta x$$



$$\tilde{f}(t) = \tilde{f}(0) + t\boxed{\tilde{f}'(0)}$$

$$= f(x) + t\nabla f(x)^T \Delta x$$

Slope $= \nabla f(x)^T \Delta x$

What is the approximation when slope is $\alpha \nabla f(x)^T \Delta x$

$$\hat{f}(t) = f(x) + \alpha t \nabla f(x)^T \Delta x$$

Switching to Newton's Method

Second Derivative of a function $f: \mathbb{R}^n \to \mathbb{R}$.

The second derivative of f, called the Hessian of f, at $x \in \text{int dom} f$, is denoted by $\nabla^2 f(x)$

f, at $x \in \operatorname{int} \operatorname{dom} f$, is denoted by $\nabla^2 f(x)$ and is defined as:

$n \times n$ matrix

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n$$

provided f is twice differentiable at x.

Gradient at $x \Rightarrow f(z) \approx f(x) + \nabla f(x)^T (z - x)$ as $z \to x$

Hessian, by definition, is a quadratic approximation of f at x

$$z \to x \qquad f(z) \approx \underbrace{f(x) + \nabla f(x)^T (z-x) + \frac{1}{2}(z-x)^T \nabla^2 f(x)(z-x)}_{\hat{f}(z)}$$

$$\lim_{\substack{z \in \operatorname{dom} f \\ z \neq x \\ z \to x}} \frac{|f(z) - \hat{f}(z)|}{\|z - x\|_2^2} = 0$$

Note: $D \nabla f(x) = \nabla^2 f(x)$

$\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n$

Hessian is the derivative of the gradient

Hessian is the derivative of the

gradient