ECE 901: Large-scale Machine Learning and Optimization **Fall 2016**

Lecture 4 — 09/15

Lecturer: Dimitris Papailiopoulos *Scribe: Muxuan Liang*

4.1 Properties of Functions

4.1.1 Some useful properties of functions

Suppose that $f : \mathbf{X} \to \mathbb{R}$, $\nabla f(\mathbf{x})$ exists and is a column-based gradient vector. We will discuss about the following properties of a function

- L-Lipschitz: We say a function is L-Lipschitz if $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2$.
- β-smooth: We say a function is β-smooth if $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \beta\|\boldsymbol{x} - \boldsymbol{y}\|_2$.
- λ-strong convex: We say a function is λ- strong convex if $f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + < \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} > + \frac{\lambda}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. This implies that strong convex function is always bounded by a quadratic function from below.

4.1.2 Examples

Table 4.1.2 is a summary for some common functions and their properties regarding as convex, L-Lipschitz, and β-smooth.

	Convexity	L-Lipschitz	β-smooth
$\|x\|$	Convex	1	
x^2	Convex	$\max_{x \in D} 2x$	2
$log(1 + e^x)$	Convex	1	1/4
$\boldsymbol{w}^\top \boldsymbol{x} + b$	Convex	$\|\boldsymbol{w}\|_2$	0
$g(\boldsymbol{w}^\top \boldsymbol{x} + b)$	if g convex	$\|\boldsymbol{w}\|_2 L_g$, if g L_g-Lipschitz	$\|\boldsymbol{w}\|_2^2 \beta_g$, if g β_g-smooth

4.1.3 Some useful results

Suppose that we have a family of convex function $\mathcal{F}$, then $\sup_{f \in \mathcal{F}} f(x)$ is also a convex function. If a function f is L-Lipschitz, then $\|\nabla f(\boldsymbol{x})\|_2 \leq L$. If a function f is β-smooth, then $\|\nabla f(\boldsymbol{x})\|_2^2 \leq \beta f(\boldsymbol{x})$. If a function f is λ-strong convex, then $f(\boldsymbol{x}) - \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2$ is also convex.

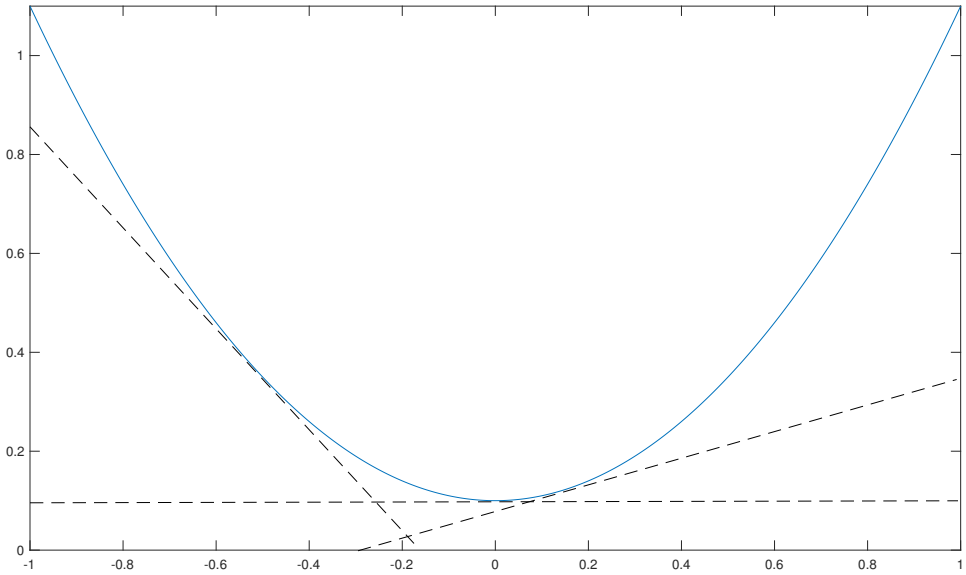

Figure 4.1. The figure show the linearization of a quadratic function. The dashed lines are linearization and solid curve is the quadratic function. The horizontal dashed line implies the gradient of the minimizer is 0.

4.2 Importance of convexity

Suppose $f(\boldsymbol{x})$ is convex and differentiable, for any point $\boldsymbol{x}_0$, the linearization of $f(\boldsymbol{x})$ is a minorization at point $\boldsymbol{x}_0$. To be specific,

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + < \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 >, \tag{4.1}$$

which implies that $\nabla f(\boldsymbol{x}^*) = 0$, where $f(\boldsymbol{x}^*) = \min f(\boldsymbol{x})$. Figure (**??**) show the relationship of the linearization and the minimizer.

4.3 Gradient Descent

Our goal is to get $\min_{\boldsymbol{x}} f(\boldsymbol{x})$ and minimizer $\boldsymbol{x}^*$. Suppose that we have a guess or last iteration of the minimizer $\boldsymbol{x}_k$, we can write our next guess or iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{u}_k$. We hope that when $k \to +\infty$, $\|\nabla f(\boldsymbol{x}_k)\| \to 0$. Our idea is to get $\boldsymbol{x}_{k+1}$ through the following optimization

$$\boldsymbol{x}_{k+1} = arg\min_{\boldsymbol{x}} \{f(\boldsymbol{x}_k) + < \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_k > + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2\}. \tag{4.2}$$

We denote the right-hand size of equation (4.2) as $g(\boldsymbol{x})$. The first two term is the linearization of $f(\boldsymbol{x})$ at $\boldsymbol{x}_k$, which gives the descent direction. The last term is the regularization to keep x close to $\boldsymbol{x}_k$. Note that equation (4.2) is a quadratic programming without

constraints. We can take the derivative and set it to be 0. Then we can get

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma \nabla f(\boldsymbol{x}_k). \tag{4.3}$$

4.3.1 Convergence of Gradient Descent

We will investigate the convergence result of the Gradient Descent method (4.3).

Theorem 4.1. *If we set* $\gamma = \frac{\Delta_1}{L\sqrt{T}}$*, where* L *is the Lipschitz constant of* $f(\boldsymbol{x})$ *and* $\Delta_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$*, then*

$$f(\frac{1}{T}\sum_{k=1}^{T} \boldsymbol{x}_k)) - f(\boldsymbol{x}^*) \leq \frac{\Delta_1 L}{\sqrt{T}}. \tag{4.4}$$

Proof: We start from the convexity of f. We know that

$$\begin{aligned} f(\boldsymbol{x}_k) - f(x^*) \leq & < \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^* > \\ = & < \frac{1}{\gamma}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^* > \\ = & \frac{1}{2\gamma}\left(\Delta_k^2 - \Delta_{k+1}^2 + \gamma^2 \|\nabla f(\boldsymbol{x}_k)\|_2^2\right). \end{aligned}$$

For f is L-Lipschitz, $\|\nabla f(\boldsymbol{x}_k)\|_2^2 \leq L^2$, then we have the following inequality

$$f(\boldsymbol{x}_k) - f(x^*) \leq \frac{1}{2\gamma}(\Delta_k^2 - \Delta_{k+1}^2) + \frac{\gamma}{2}L^2.$$

We take $k = 1, \ldots, T$ and add those inequality up. Then we can get

$$\sum_{k=1}^{T} f(\boldsymbol{x}_k) - f(x^*) \leq \frac{1}{2\gamma}\Delta_1^2 + \frac{\gamma}{2}TL^2.$$

By convexity and optimize over γ, when we set $\gamma = \frac{\Delta_1}{L\sqrt{T}}$, we have

$$f(\frac{1}{T}\sum_{k=1}^{T} \boldsymbol{x}_k)) - f(\boldsymbol{x}^*) \leq \frac{\Delta_1 L}{\sqrt{T}}.$$

□