CSE6243: Advanced Machine Learning **Fall 2023**

Lecture 4: Optimization: Conjugate and Gradient Descent

Lecturer: Bo Dai *Scribes: Abhinav Gullapalli, Mahdi Ghanei*

Note: *LaTeX template courtesy of UC Berkeley EECS Department.*

4.1 Recap

- How to judge whether an optimization is convex or not?

$$\min_{x\in\Omega} f(x) \text{ is convex if} \tag{4.1}$$

$$\Omega \text{ is a convex set } \textit{and } f(x) \text{ is a convex function.} \tag{4.2}$$

4.2 New Content

4.2.1 Zeroth Order Condition for Convex Functions

A function $f(\cdot) \in \Omega \to \mathbb{R}$ is a convex function *iff*

$$\begin{cases} \Omega \text{ is a convex set} \\ f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \quad \forall x, y \in \Omega,\ t \in [0,1] \end{cases} \tag{4.3}$$

4.2.2 First Order Condition for Convex Functions

A *differentiable* function $f(\cdot) \in \Omega \to \mathbb{R}$ is a convex function *iff*

$$\begin{cases} \Omega \ \text{ is a convex set} \\ f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \Omega \end{cases} \tag{4.4}$$

Proof: Bidirectionality of First-Order Condition

(if part: "$\Leftarrow$")

Given $x_0 = tx + (1-t)y, t \in [0,1]$, we have

$$\begin{cases} f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0) \\ f(y) \geq f(x_0) + \nabla f(x_0)^T (y - x_0) \end{cases} \tag{4.5}$$

We multiply top row by t and bottom row by $(1-t)$. Thus,

$$tf(x) + (1-t)f(y) \geq f(x_0) + \nabla f(x_0)^T(t(x-x_0) + (1-t)(y-x_0)) \tag{4.6}$$

Note: $t(x-x_0) + (1-t)(y-x_0) = 0$ when substituting $x_0 = tx + (1-t)y$, thus

$$tf(x) + (1-t)f(y) \geq f(tx + (1-t)y) \tag{4.7}$$

(only-if part: ”⇒”)

When $t = 0$, the inequality $f((1-t)y + tx) \leq tf(x) + (1-t)f(y)$ is trivially satisfied. Thus, we only consider below the case where $t \in (0, 1]$:

Given $f((1-t)y + tx) \leq tf(x) + (1-t)f(y)$, we have

$$\frac{f((1-t)y + tx)}{t} \leq \frac{tf(x) + (1-t)f(y)}{t} \tag{4.8}$$

$$\frac{f((1-t)y + tx)}{t} \leq f(x) - f(y) + \frac{f(y)}{t} \tag{4.9}$$

$$f(x) \geq f(y) + \frac{f((1-t)y + tx) - f(y)}{t} \tag{4.10}$$

As $t \to 0$, we have $f(x) \geq f(y) + \nabla f(y)^T(x-y)$ by applying Taylor Expansion

4.2.3 Second Order Condition for Convex Functions

A *twice-differentiable* $f(\cdot) \in \Omega \to \mathbb{R}$ is a convex function *iff*

$$\begin{cases} \Omega \text{ is a convex set} \\ \nabla^2 f(x) \geq 0, \quad \forall x \in \Omega \end{cases} \tag{4.11}$$

Note: $\nabla^2 f(x)$ is a positive semi-definite Hessian matrix.

Proof: Bidirectionality of Second Order Condition

(if part: ”⇐”)

Given $f(x+h) = f(x) + h^T \nabla f(x+h) + \frac{1}{2}h^T \nabla^2 f(x+h)h + O(||h||^3)$, we observe

$$\frac{1}{2}h^T \nabla^2 f(x+h)h + O(||h||^3) \geq 0 \tag{4.12}$$

Thus, $f(x+h) \geq f(x) + h^T \nabla f(x+h)$, $\forall h$.

Note: $O(||h||^3)$ is a residual term and dominated by the second-order term $\frac{1}{2}h^T \nabla^2 f(x+h)h$.

(only-if part: ”⇒”)

Given $f(x+h) = f(x) + h^T \nabla f(x+h) + \frac{1}{2}h^T \nabla^2 f(x+h)h + O(||h||^3)$, we have

$$0 \leq f(x+h) - f(x) - h^T \nabla f(x+h) = \frac{1}{2}h^T \nabla^2 f(x+h)h + O(||h||^3) \quad \forall h, x \tag{4.13}$$

Thus, $\nabla^2 f(x) \geq 0$.

4.2.4 Convexity of Composition of Functions

The function $g(x) = f(h(x))$ is convex when:

$$\begin{cases} f \text{ is convex and increasing AND } h \text{ is convex} \\ f \text{ is convex and decreasing AND } h \text{ is } \underline{\text{concave}} \end{cases} \tag{4.14}$$

Proof: This can be obtained via taking the derivative

$$g'(x) = f'(h(x))h'(x) \tag{4.15}$$

$$\Rightarrow g''(x) = f''(h(x))(h'(x))^2 + f'(h(x))h''(x) \tag{4.16}$$

■ The terms in the second derivative are positive, thus the g is convex.

Note: The above conditions are sufficient but not necessary for a function to be convex. Even if these conditions are not satisfied, it is possible for the function to be convex.

Example: Is the function $g(x) = \log \sum_{i=1}^{d} exp(a_i^T x + b_i)$ convex? Yes.

Proof: This can be proved by taking the derivative twice and using the second-order condition.

4.2.5 Gradient Descent

The objective is to solve the optimization $\min_x f(x)$ where $f(x)$ is a loss function. The algorithm is as follows:

1. initialize x_0
2. for $t = 1, ..., T$ do $x_{t+1} = x_t - \eta \nabla f(x_t)$

where η is step-size.

Observation 4.1 *x_{t+1} is the solution to surrogate loss function*

$$x_{t+1} = \arg\min_x f(x_t) + \nabla f(x_t)^T (x - x_t) + \frac{1}{2\eta} ||x - x_t||^2$$

Proof: We take the derivative and set it equal to zero:

$$\nabla f(x_t) + \frac{1}{\eta}(x - x_t) = 0 \tag{4.17}$$

$$\Rightarrow x_{t+1} = x_t - \eta \nabla f(x_t) \tag{4.18}$$

■

Observation 4.2 *$f(x_{t+1}) - f(x) \leq 0$ holds for L-smooth function.*

Proof:

$$\forall x, y,\ f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{L}{2} ||y - x||^2 \tag{4.19}$$

$$\Rightarrow f(x_{t+1}) - f(x_t) \leq \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} ||x_{t+1} - x||^2 \tag{4.20}$$

We want to prove the right-hand-side is ≤ 0.

Substituting for $x_{t+1} = x_t - \eta \nabla f(x_t)$, we get:

$$-\eta \nabla f(x_t)^T \nabla f(x_t) + \frac{L}{2} ||\eta \nabla f(x_t)||^2$$

note that $\frac{L}{2} ||\eta \nabla f(x_t)||^2 = \frac{L\eta^2}{2} \nabla f(x_t)^T \nabla f(x_t)$, hence

$$-\eta \nabla f(x_t)^T \nabla f(x_t) + \frac{L}{2} ||\eta \nabla f(x_t)||^2 = (\frac{L\eta^2}{2} - \eta) \nabla f(x_t)^T \nabla f(x_t)$$

For optimal η^* we have $L\eta^* - 1 = 0 \rightarrow \eta^* = \frac{1}{L}$ Thus,

$$-\eta^* \nabla f(x_t)||^2 = \frac{-1}{2L} \nabla f(x_t)^T \nabla f(x_t) \; ; \; \eta^* = \frac{1}{L}$$

From the above, we can get a bound on the improvement:

$$f(x_{t+1}) - f(x_t) \leq \frac{-1}{2L} ||\nabla f(x_t)||^2 \; ; \; \eta^* = \frac{1}{L}$$

■

Theorem 4.3 *Finding optimal x with Gradient Descent*

$$f(x_t) - min f(x) \leq \frac{2L||x_0 - x_*||^2}{t} \tag{4.21}$$

When gradient descent is applied for convex and L-smooth, this condition holds.

Note: $x_* = argmin f(x)$, and the optimal x could be a set, not necessarily a single value.

4.2.6 Stochastic Gradient Descent

Stochastic Gradient Descent is one type of Gradient Descent, which is often used in practice.

$\min_x f(x) = \sum_{i=1}^n f_i(x)$

$\nabla f(x) = \nabla \sum_{i=1}^n f_i(x) = \sum_{i=1}^n \nabla f_i(x)$

$\tilde{\nabla} f(x) = \sum_{i=1}^k \nabla f_i(x) \; \forall k << n$

Also,

$\mathbb{E}[\tilde{\nabla} f(x)] = \nabla f(x)$

$\mathbb{E}[||\tilde{\nabla} f(x)||^2] \leq C$

which means that it is that $\tilde{\nabla} f(x)$ is an unbiased estimator of $f(x)$.

$x_{t+1} = x_t - y \tilde{\nabla} f(x_t)$ gives stochastic gradient descent.

4.2.7 Convex Conjugate & Gradient Descent (Supplementary)

Definition 4.4 (Convex Conjugate)

$$f : \mathbb{R}^n \to \mathbb{R} \textit{is a conjugate function if} \tag{4.22}$$

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{y^T x - f(x)\} \quad \textbf{\textit{(Legendre-Fenchel Transformation)}} \tag{4.23}$$

which possesses the following properties:

Property 1 (Fenchel's Inequality)

$$f(x) + f^*(y) \geq x^T y, \quad \forall x, y \tag{4.24}$$

Property 2 $f^*(\cdot)$ is convex.

Hint to proof: consider pointwise max rule.

Definition 4.5 (BiConjugate Function) *We call $f^{**}(x)$ the biconjugate function if it is the conjugate of a conjugate function, i.e.,*

Given $f(x)$ and its conjugate $f^*(y) = \sup_y y^T x - f(x)$,

$$f^{**}(x) = \sup_x y^T x - f^*(y) \tag{4.25}$$

Theorem 4.6

$$f(x) \geq f^{**}(x) \tag{4.26}$$

Proof: By definition,

$$f^*(y) \geq y^T x - f(y), \quad \forall y \tag{4.27}$$

$$\Rightarrow f(x) \geq y^T x - f^*(y) \quad \text{i.e.,} \tag{4.28}$$

$$f(x) \geq \sup_y (y^T x - f^*(y)) = f^{**}(x) \tag{4.29}$$

■

Therefore, fˆ** is lower bound convex of $f(x)$.

Theorem 4.7 *if f is convex and closed,*

$$f^{**}(x) = f(x) \tag{4.30}$$