Recitation: Convexity

Shiyun

October 9, 2020

1 Convexity, Strong convexity, Smoothness

1.1 Convex set

Definition 1 *A set C is convex if for all $x, y \in C$, the segment connecting x and y is contained in C, that is $seg(x, y) \subset C$.*

Remark 2 *The set $x : f_i(x) \le b_i, i = 1, ..., m$ is convex if all f_i are convex functions.*

Definition 3 *A function f is convex if dom f is a convex set and Jensen's inequality holds:*

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y) \quad \text{for all } x, y \in \operatorname{dom} f, \theta \in [0,1].$$

When the domain of f is not convex, then the function f is no longer convex in its domain.

Theorem 4 (Separating hyperplane theorem) *Every two disjoint convex sets C and D can be separated by a hyperplane, which means there exists $a \neq 0, b$ such that*

$$a^\top x \le b \text{ for } x \in C, \quad a^\top x \ge b \text{ for } x \in D$$

.

1.2 Function convexity

Definition 5 *An operator $\mathcal{F} : H \to H$ is monotone if $\langle \mathcal{F}(x) - \mathcal{F}(y), x - y\rangle \ge 0$.*

Theorem 6 *A differentiable function f is convex iff ∇f is monotone.*

Proof:.

1.3 Function smoothness

Definition 7 *A function f is called L-smooth if the gradient of f is Lipschitz continuous with parameter $L > 0$ such that*

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2 \quad \text{for all } x, y \in \operatorname{dom} f$$

Remark 8 *If f is twice differentiable then*

$$\begin{aligned} f \text{ is L-smooth} &\Longleftrightarrow \nabla^2 f(x) \preceq LI \\ &\Longleftrightarrow \frac{L}{2}\|x\|_2^2 - f(x) \text{ is convex} \end{aligned}$$

Remark 9 *Descent Lemma / Quadratic upper bound.*
By definition, f is L-smooth implies

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \le L\|x - y\|^2, \quad \text{for all } x, y \in \operatorname{dom} f. \tag{1}$$

If dom f is convex, the inequality 1 is equivalent to

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|_2^2.$$

Proof:.

Lemma 10 *A consequence of the descent lemma is that*

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f(x^*) \leq \frac{L}{2}\|x - x^*\|^2$$

.

Proof:.

Remark 11
Lipschitz continuity:

$$|f(x) - f(y)|_2 \leq L\|x - y\|_2$$

Lipschitz continuity of gradients: The function $f(x)$ is differentiable and its gradient is also Lipschitz continuous.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|_2$$

.

co-coercivity of gradients:

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2, \text{ for all } x, y.$$

Lipschitz continuity of $\nabla f \iff$ co-coercivity of ∇f.
Proof:.

1.4 Function strong convexity

Definition 12 *A convex function f is μ-strong convex if*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{\mu}{2}\theta(1-\theta)\|x - y\|^2$$

holds for all $x, y \in dom f, \theta \in [0, 1]$.

Remark 13 (first order condition)

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2$$

Remark 14 (second order condition) *The strongly convex holds if and only if it holds for f restricted to arbitrary lines:*

$$f(x + t(y - x)) - \frac{\mu}{2}t^2\|x - y\|^2$$

If f is twice differentiable then

$$\begin{aligned} f \text{ is } \mu - convex &\iff \nabla^2 f(x) \succeq \mu I \\ &\iff f(x) - \frac{\mu}{2}\|x\|_2^2 \text{ is convex} \\ &\iff \langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2, \quad \text{for all } x, y \in \operatorname{dom} f. \end{aligned}$$

The last inequality is called strong monotonicity /coercivity of ∇f.

Remark 15 *Difference between strictly convex and strongly convex function.*

- *Strictly convex function: for all $x \neq y$ and $\lambda \in (0,1)$,*

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y).$$

  . (First order condition)

$$f(y) > f(x) + \nabla f(x)^T (y-x) \text{ for all } x, y \in \operatorname{dom} f, x \neq y$$

  . (Second-order condition)

$$\nabla^2 f(x) > 0$$

  .

- *Strongly convex functions: there exists an $\alpha > 0$ such that*

$$f(x) - \alpha \|x\|_2^2$$

  is convex.

- *strong convexity $\Rightarrow$ strict convexity $\Rightarrow$ convexity.*

2 Analysis of gradient descent (GD)

This section talks about the unconstrained optimization of minimizing a convex function

$$\min_x f(x)$$

Lemma 16 *For convex functions, local minima are global minima.*

Proof:.

2.1 GD for convex and smooth functions

Definition 17 *The Gradient descent is defined by the following updating rule:*

$$x_{k+1} = x_k - t_k \nabla f(x_k).$$

for fixed step size or back tracking line search.
Now we are going to analyze the problem where f is only convex and $L-$smooth. We can bound the function value difference by

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} ||x_{k+1} - x_k||^2 \quad \textit{(L-smoothness)} \\ &= -t_k ||\nabla f(x_k)||^2 + \frac{L}{2} t_k^2 ||\nabla f(x_k)||^2 \\ &= (-t_k + \frac{L}{2} t_k^2) ||\nabla f(x_k)||^2 \end{aligned}$$

The last line is minimized when $t_k = \frac{1}{L}$. We take t_k to be this value and obtain

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L} ||\nabla f(x_k)||^2.$$

Based on the above observation, we have the following theorem.

Theorem 18 *Suppose f is convex and L-smooth, then if we pick step size $t_k = \frac{1}{L}$ in gradient descent, we have*

$$f(x_{k+1}) - f(x^*) \leq \frac{L||x_0 - x^*||^2}{2k} = \mathcal{O}(\frac{1}{k}).$$

As in the previous section, we would also like to investigate the rate of convergence—the difference between the point we get in kth iteration and the minimum point. We have the following theorem.

Theorem 19 *Using the gradient descent defined as before with step size $t_k = \frac{1}{L}$, we have*

$$||x_{k+1} - x^*||^2 \leq ||x_k - x^*||^2 - \frac{||\nabla f(x_k)||^2}{L^2}.$$

Proof:.

Remark 20 *Note that x^* here may not be unique—this result applies to all optimizers attaining the same function value as $f(x^*)$.*

2.2 GD for strongly convex functions (linear convergence)

Theorem 21 *If f is μ-strong convex and L-smooth, then for $t_k = \frac{2}{\mu+L}$ and $\kappa = \frac{L}{\mu} \geq 1$, we have*

$$\|x_k - x^*\| \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|x_0 - x^*\|.$$

Proof:.

Remark 22 *In this lecture all the norms are the L_2 norm and the above convergence rate is called Linear Convergence in optimization.*

Remark 23 *Assuming L-smoothness, we have*

$$f\left(x_k\right) - f\left(x^*\right) \leq \frac{L}{2}\left\|x_k - x^*\right\|^2 \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k}\left\|x_0 - x^*\right\|^2.$$

3 Stochastic Gradient Descent

In machine learning, we solve the finite-sum problem. Given a finite dataset $D = \{(\xi_i, y_i)\}_{i=1,\ldots,n}$, we minimize

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} l\left(x; \xi_i, y_i\right)$$

. In practice, the number of data n can be very large in modern machine learning. It is difficult to do gradient descent in this case because the gradient. Stochastic gradient descent for the finite-sum case performs the following iterations:

$$x^{t+1} = x^t - \eta \nabla l\left(x^t; \xi_{\omega t}, y_{\omega t}\right)$$

. The datum $(\xi_{\omega t}, y_{\omega t})$ over which we compute the gradient before updating the weights is picked randomly from the dataset D.

Remark 24 (Mini-batch version of SGD)

$$x^{t+1} = x^t - \frac{\eta}{\vartheta}\sum_{k=1}^{\vartheta} \nabla l\left(x^t; \xi_{\omega t}^k, y_{\omega t}^k\right)$$

.

Lemma 25 (Descent lemma for stochastic updates) *The next update for SGD satisfies*

$$\mathbb{E}_{\omega_t}\left[f\left(x^{t+1}\right)\right]-f\left(x^t\right)\leq-\alpha\langle\nabla f\left(x^t\right),\mathbb{E}_{\omega_t}\left[\nabla f\left(x^t\right)\right]\rangle+\frac{L\alpha^2}{2}\mathbb{E}_{\omega_t}\left[\left\|\nabla f\left(x^t\right)\right\|^2\right].$$

Proof. Use the descent lemma, substitute the iterates of SGD and take an expectation on both sides over the index of the datum ωt.
[1]

References

References

[1] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[3] L. Vandenberghe. *lecture notes in UCLA ECE236C: Spring 2020*. 2020. URL: `http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf`.

[1] Check [1, 2] and lecture notes in UCLA ECE236C [3] for reference.