6FMAI19 Nonlinear Optimization	Spring, 2022
Lecture #4 — Feb. 9th, 2022	
Lecturer: Yura Malitsky	*Scribe: Jianan Bai*

1 Gradient Descent with Convex Objective Function

During last lecture, we introduced the gradient descent algorithm

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k), \quad \text{for } k = 0, 1, \cdots, K-1, \tag{1}$$

and analyzed the convergence with fixed step size when the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is *L-smooth* and *bounded below* by f^*. Now, we provide the convergence analysis of gradient descend when the $f(\cdot)$ is also convex.

1.1 Convergence Analysis

Assumption 1. *(L-Smoothness) The objective function is differentiable and L-smooth, such that* $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ *for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Assumption 2. *(Convexity) The objective function is convex, such that* $\theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) \geq f(\theta \mathbf{x} + (1-\theta)\mathbf{y})$ *for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *and* $\theta \in [0, 1]$.

Theorem 3. *Under Assumptions 1 and 2, the gradient descent algorithm, with fixed step size* $\alpha_k = \alpha \in (0, 1/L]$*, gives*

$$f(\mathbf{x}_K) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2\alpha K}, \tag{2}$$

where $\mathbf{x}^*$ *is the optimal solution, i.e.,* $\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$.

Proof. First consider

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k) - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\alpha \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle + \alpha^2 \|\nabla f(\mathbf{x}_k)\|^2. \end{aligned} \tag{3}$$

Using the first-order condition of convex functions, we have

$$\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle \geq f(\mathbf{x}_k) - f(\mathbf{x}^*). \tag{4}$$

Meanwhile, by utilizing L-smoothness

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= f(\mathbf{x}_k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(\mathbf{x}_k)\|^2 \\ &\overset{(a)}{\leq} f(\mathbf{x}_k) - \frac{\alpha}{2} \|\nabla f(\mathbf{x}_k)\|^2, \end{aligned} \tag{5}$$

where (a) follows by the condition $\alpha \in (0, 1/L]$. The above equation can be rearranged as

$$\|\nabla f(\mathbf{x}_k)\|^2 \leq -\frac{2}{\alpha} \left(f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \right). \tag{6}$$

Substitute (4) and (6) into (3), we obtain

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2 &\leq \|\mathbf{x}_k-\mathbf{x}^*\|^2 - 2\alpha\left(f(\mathbf{x}_k)-f(\mathbf{x}^*)\right) - 2\alpha\left(f(\mathbf{x}_{k+1})-f(\mathbf{x}_k)\right)\\ &= \|\mathbf{x}_k-\mathbf{x}^*\|^2 - 2\alpha\left(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)\right).\end{aligned} \tag{7}$$

Since (7) holds for all $k=0,1,\cdots,K-1$, we can sum the LHS and RHS over k, i.e.,

$$\sum_{k=0}^{K-1}\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2 \leq \sum_{k=0}^{K-1}\|\mathbf{x}_k-\mathbf{x}^*\|^2 - 2\alpha\sum_{k=0}^{K-1}\left(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)\right), \tag{8}$$

which can be rearranged as

$$\begin{aligned}2\alpha\sum_{k=0}^{K-1}\left(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)\right) \leq& \|\mathbf{x}_0-\mathbf{x}^*\|^2 - \|\mathbf{x}_K-\mathbf{x}^*\|^2\\ \leq& \|\mathbf{x}_0-\mathbf{x}^*\|^2.\end{aligned} \tag{9}$$

Notice that, as indicated by (5), $\{f(\mathbf{x}_k)\}$ is non-increasing sequence. Therefore, the LHS of (9) can be lower bounded by

$$2\alpha\sum_{k=0}^{K-1}\left(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)\right) \geq 2\alpha K\left(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)\right). \tag{10}$$

After substituting the above inequality into (9), we obtain

$$f(\mathbf{x}_K)-f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_0-\mathbf{x}^*\|^2}{2\alpha K}, \tag{11}$$

which completes the proof.

□

Remark 4. *Theorem 3 implies that the convergence rate of gradient descent with convex objective function is $O(1/k)$. Equivalently, to achieve an accuracy $\epsilon > 0$, such that $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$, we need to run gradient decent for $O(1/\epsilon)$ iterations.*

1.2 Nesterov Acceleration

Instead of using the update equation in (1), an alternative method to improve the convergence is to use *Nesterov's accelerated gradient* (NAG), with the update in the k-th iteration:

$$\begin{aligned}\mathbf{y}_k &= \mathbf{x}_k + \frac{2}{k+2}(\mathbf{x}_k-\mathbf{x}_{k-1}),\\ \mathbf{x}_{k+1} &= \mathbf{x}_k - \alpha\nabla f(\mathbf{y}_k).\end{aligned} \tag{12}$$

This achieves

$$f(\mathbf{x}_k)-f(\mathbf{x}^*) \leq \frac{c}{k^2}, \tag{13}$$

where $c > 0$ is a constant. Obviously, NAG achieves a convergence rate $O(1/k^2)$. Equivalently, to achieve an ϵ-solution, such that $f(\mathbf{x}_k)-f(\mathbf{x}^*) \leq \epsilon$, we need to run NAG for $O(1/\sqrt{\epsilon})$ iterations.

1.3 Drawbacks of Gradient Descent

There are two major drawbacks of gradient descent:[1]

1. The convergence is slow. As we can see from Theorem 3, the convergence rate of gradient descent for convex function is $O(1/k)$, rather than $O(1/k^2)$.

2. It is usually difficult to know the Lipschitz constant L and choose the fixed step size α.

There are some remedies to the choice of step size. Here, we introduce two of them:

1.3.1 Backtracking Line Search

Algorithm 1 Gradient Descent with Backtracking Line Search

Require: $\mathbf{x}_0$, $\beta \in (0, 1)$
Ensure: $\mathbf{x}_K$
  for $k = 0, \cdots, K-1$ **do**
    $\alpha := 1$
    while $f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|^2$ **do**
      $\alpha := \beta\alpha$
    end while
    $\mathbf{x}_{k+1} := \mathbf{x}_k - \alpha\nabla f(\mathbf{x}_k)$
  end for

1.3.2 Adaptive Learning Rate

Another way to select the learning rate is to use adaptive learning rate. One example is

$$\alpha_k = \min\left\{\frac{\|\mathbf{x}_k - \mathbf{x}_{k-1}\|}{2\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})\|}, \sqrt{1 + \frac{\alpha_{k-1}}{\alpha_{k-2}}}\alpha_{k-1}\right\}. \tag{14}$$

2 Gradient Descent with Constraints

Now, we consider the optimization problem with constraints:

$$\underset{\mathbf{x}\in\mathcal{C}}{\text{minimize}} \quad f(\mathbf{x}), \tag{15}$$

where $\mathcal{C}$ is a *convex* and *closed* set. Notice that interesting problems usually have the optimal solution on the boundary of $\mathcal{C}$. Otherwise, we can simply solve an unconstrained problem.

2.1 Optimality Condition

Theorem 5. *For the problem in (15), we have the following optimality conditions:*

- *For differentiable $f(\cdot)$, if $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$, then $\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0$ for all $\mathbf{x} \in \mathcal{C}$.*

- *If $f(\cdot)$ is differentiable and convex, then $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$ if and only if $\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0$ for all $\mathbf{x} \in \mathcal{C}$.*

[1]We restrict our discussion to the case in Theorem 3.

Proof. Let us start by proving the first part. For any $\mathbf{x} \in \mathcal{C}$, since $\mathcal{C}$ is convex, we have

$$\mathbf{x}^* + \theta(\mathbf{x} - x^*) \in \mathcal{C}, \tag{16}$$

for any $\theta \in [0, 1]$. Since $\mathbf{x}^*$ is the optimal solution, we have

$$f(\mathbf{x}^* + \theta(\mathbf{x} - \mathbf{x}^*)) \geq f(\mathbf{x}). \tag{17}$$

Then,

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle = \lim_{\theta \to 0} \frac{f(\mathbf{x}^* + \theta(\mathbf{x} - \mathbf{x}^*)) - f(\mathbf{x})}{\theta} \geq 0, \tag{18}$$

which proves the first part.

For the second part, the sufficiency is obvious from the the first part. Now, we prove the necessity. Assume $\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0$ holds for all $\mathbf{x} \in \mathcal{C}$. Since $f(\cdot)$ is convex, by using the first-order condition

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq f(\mathbf{x}^*), \ \forall \mathbf{x} \in \mathcal{C}, \tag{19}$$

which completes the proof. □