ESTR2520: Optimization Methods **2023–24 Second Term**

Handout 7: Analysis of Gradient Descent Method

Instructor: Hoi-To Wai Last updated: March 23, 2024

This note analyzes the gradient descent (GD) method introduced in the FTEC lecture. In particular, we will show that the GD method finds an optimal solution to an unconstrained convex optimization problem and also characterize the **rate of convergence**.

1 Recap: Gradient (Descent) Method and Assumptions

Our quest is the solve the following **unconstrained** problem:

$$\min_{x\in\mathbb{R}^n} \; f(x) \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a convex and differentiable function. Let us first recall the GD method:

Gradient Descent Method

Input: x^0 - initialization, set the iteration counter as $t=0$.

Repeat

Select a step size $\gamma>0$

$x^{t+1}=x^t-\gamma\nabla f(x^t)$

Until convergence / stopping criterion.

Our analysis will be focused on the case when $f(x)$ is **convex** and **smooth**. That is,

Assumption 1 *The function $f(x)$ is convex and lower bounded, i.e., $\min_{x\in\mathbb{R}^n} f(x)>-\infty$.*

The above condition can be satisfied if the Weierstrass theorem's condition holds, i.e., when there exists $\gamma>0$ such that the level set $\{x\in\mathbb{R}^n: f(x)\le\gamma\}$ is non-empty and compact.

Assumption 2 *The gradient of $f(x)$ is L Lipschitz continuous, i.e., there exists $L\in\mathbb{R}_+$ such that*

$$\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|,\quad x,y\in\mathbb{R}^n.$$

Denote the maximum eigenvalue of a matrix as $\lambda_{\max}(X)$, we observe the following fact:

Fact 1 *If the Hessian matrix of f satisfies*

$$L\cdot I_n-\nabla^2 f(x) \text{ is PSD} \iff \lambda_{\max}(\nabla f^2(x))\le L.$$

Then Assumption 2 holds with the constant L.

Example 1.1 *The function $f(x)=\frac{1}{2}x^\top x$ satisfies Assumption 2 with the constant $L=1$.*

Example 1.2 *The logistic regression objective function:*

$$f(w) = \sum_{t=1}^{T} \log(1 + \exp(-y_t x_t^\top w)) + \frac{1}{2} w^\top w$$

has the following expression for its Hessian:

$$\nabla^2 f(w) = \sum_{t=1}^{T} \frac{\exp(-y_t x_t^\top w)}{(1 + \exp(-y_t x_t^\top w))^2} x_t x_t^\top + \frac{1}{2} \cdot I_n$$

Using the fact that $\frac{\exp(-y_t x_t^\top w)}{(1+\exp(-y_t x_t^\top w))^2} \leq \frac{1}{2}$ *and* $\lambda_{\max}(A+B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$*, we observe*

$$\lambda_{\max}(\nabla^2 f(w)) \leq \frac{1}{2} \sum_{t=1}^{T} \lambda_{\max}(x_t x_t^\top) + \frac{1}{2} = \frac{1}{2} \left(1 + \sum_{t=1}^{T} \|x_t\|^2 \right) =: L < \infty$$

Our proof shall rely on the following **descent lemma** which is a consequence of Assumption 2.

Lemma 1 *If* $f(x)$ *satisfies Assumption 2, then*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|^2, \ \forall \ x, y \in \mathbb{R}^n.$$

The lemma can actually be derived from the Taylor's expansion of the function $f(y)$ around x.

2 Analysis for the Convex Case

Theorem 1: Convergence Rate for Convex Problems

Under Assumption 1, 2. Consider the iterates $\{x^t\}_{t \geq 0}$ generated by the GD method with a **constant step size** $\gamma = \frac{1}{L}$. Then

$$f(x^{T+1}) - \min_{x \in \mathbb{R}^n} f(x) \leq \frac{L \|x^0 - x^\star\|^2}{2T}, \ \forall \ T \geq 1.$$

where $x^\star \in \arg\min_{x \in \mathbb{R}^n} f(x)$.

Proof: We first establish the following auxilliary result that is instrumental to our proof.

Lemma 2 *Under Assumption 1, 2. Consider the iterates* $\{x^t\}_{t \geq 0}$ *generated by the GD method with* $\gamma = \frac{1}{L}$*. It holds*

$$f(x^{t+1}) \leq f(x^t) - \frac{1}{2L} \|\nabla f(x^t)\|^2, \ \forall \ t \geq 0.$$

Proof The lemma is a direct consequence of Lemma 1. In particular, we substitute $y = x^{t+1}$, $x = x^t$ in the said lemma, we get

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \nabla f(x^t)^\top (x^{t+1} - x^t) + \frac{L}{2} \|x^{t+1} - x^t\|^2 \\ &= f(x^t) + \nabla f(x^t)^\top (-\gamma \nabla f(x^t)) + \frac{L}{2} \|\gamma \nabla f(x^t)\|^2 \\ &= f(x^t) - \gamma \left(1 - \gamma \frac{L}{2} \right) \|\nabla f(x^t)\|^2 \end{aligned}$$

Substituting $\gamma = \frac{1}{L}$ completes the proof. $\square$

An important consequence of the lemma is that it shows $f(x^{t+1}) \leq f(x^t)$, i.e., the objective value is non-increasing. To prove the theorem, we begin by noticing that as f is convex,

$$f(x^\star) \geq f(x^t) + \nabla f(x^t)^\top (x^\star - x^t) \tag{2}$$

Now,

$$f(x^{t+1}) \overset{\text{Lemma 2}}{\leq} f(x^t) - \frac{1}{2L}\|\nabla f(x^t)\|^2 \overset{(2)}{\leq} f(x^\star) + \nabla f(x^t)^\top (x^t - x^\star) - \frac{1}{2L}\|\nabla f(x^t)\|^2.$$

Moreover,

$$\begin{aligned}
&\nabla f(x^t)^\top (x^t - x^\star) - \frac{1}{2L}\|\nabla f(x^t)\|^2 \\
&= \frac{L}{2}\left(\|x^t - x^\star\|^2 - \|x^t - x^\star\|^2 + \frac{2}{L}\nabla f(x^t)^\top (x^t - x^\star) - \frac{1}{L^2}\|\nabla f(x^t)\|^2 + \right) \\
&= \frac{L}{2}\left(\|x^t - x^\star\|^2 - \left\|x^t - \frac{1}{L}\nabla f(x^t) - x^\star\right\|^2\right) \\
&= \frac{L}{2}\left(\|x^t - x^\star\|^2 - \|x^{t+1} - x^\star\|^2\right)
\end{aligned}$$

This implies for any $t \geq 0$,

$$f(x^{t+1}) - f(x^\star) \leq \frac{L}{2}\left(\|x^t - x^\star\|^2 - \|x^{t+1} - x^\star\|^2\right).$$

Summing up the inequality from $t = 0$ to $t = T - 1$, we have

$$\sum_{t=1}^{T} f(x^t) - f(x^\star) \leq \frac{L}{2}\sum_{t=0}^{T-1}\left(\|x^t - x^\star\|^2 - \|x^{t+1} - x^\star\|^2\right) \leq \frac{L}{2}\|x^0 - x^\star\|^2.$$

Noting that the left hand side is lower bounded by $T \cdot (f(x^T) - f(x^\star))$ concludes the proof. $\square$

3 Analysis for the Strongly Convex Case

The convergence rate of the GD method can be further accelerated under the strong convexity condition, i.e.,

Assumption 3 *The function $f(x)$ is μ strongly convex (with $\mu > 0$) such that*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2, \ \forall\ x, y \in \mathbb{R}^n.$$

As a matter of fact, a function f is μ-strongly convex if $\nabla^2 f(x) - \mu \cdot I$ is PSD for any $x \in \mathbb{R}^n$.

Theorem 2: Convergence Rate for Strongly Convex Problems

Under Assumptions 1, 2, 3. Consider the iterates $\{x^t\}_{t\geq 0}$ generated by the GD method with a **constant step size** $\gamma = \frac{1}{L}$. Then

$$\|x^{t+1} - x^\star\|^2 \leq (1 - \mu/L)\|x^t - x^\star\|^2, \ \forall\ t \geq 0.$$

Proof: The proof is actually even more straightforward. Observe that

$$\begin{aligned}\|x^{t+1} - x^\star\|^2 &= \|x^t - (1/L)\nabla f(x^t) - x^\star\|^2 \\ &= \|x^t - x^\star\|^2 - (2/L)\nabla f(x^t)^\top (x^t - x^\star) + (1/L^2)\|\nabla f(x^t)\|^2\end{aligned}$$

Since $f(x)$ is strongly convex, we have

$$\nabla f(x^t)^\top (x^\star - x^t) \leq f(x^\star) - f(x^t) - \frac{\mu}{2}\|x^t - x^\star\|^2$$

We have

$$\begin{aligned}\|x^{t+1} - x^\star\|^2 &\leq (1 - \mu/L)\|x^t - x^\star\|^2 + (2/L)(f(x^\star) - f(x^t)) + (1/L^2)\|\nabla f(x^t)\|^2 \\ &= (1 - \mu/L)\|x^t - x^\star\|^2 + \frac{2}{L}\left(f(x^\star) - f(x^t) + \frac{1}{2L}\|\nabla f(x^k)\|^2\right)\end{aligned}$$

We note that by Lemma 2,

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

This concludes that $\|x^{t+1} - x^\star\|^2 \leq (1 - \mu/L)\|x^t - x^\star\|^2$. □

4 Perspectives

Note that when f is strongly convex, we have $f(x^t) - f(x^\star) = \mathcal{O}(\|x^t - x^\star\|^2)$. Now, from the above analysis, we see that the convergence rates for GD are:

$$\text{(Convex)} \quad f(x^t) - f(x^\star) = \mathcal{O}(1/t), \quad \text{(Str.-convex)} \quad f(x^t) - f(x^\star) = \mathcal{O}((1 - \mu/L)^t)$$

In other words, the convergence is **sublinear** for the convex case, and is **linear** for the strongly convex case.

The GD method is one of the simplest algorithm for (numerically) solving optimization problems, especially for those problems that are convex and differentiable. The algorithm design also yields the backbone for more advanced algorithms such as SGD, ADAM, etc.

For your curiosity, it is known that the **optimal** gradient-based algorithms will have the rates

$$\text{(Convex)} \quad f(x^t) - f(x^\star) = \mathcal{O}(1/t^2), \quad \text{(Str.-convex)} \quad f(x^t) - f(x^\star) = \mathcal{O}((1 - \sqrt{\mu/L})^t)$$

They are called the **Nesterov's accelerated gradient method** which are simply more complicated version GD, e.g., involving momentum terms. You may check out the references [1, 2] for more details.

References

[1] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[2] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.