Optimization Theory and Algorithm **Lecture 7 - 05/18/2021**

Lecture 7

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 Gradient Descent for Beta Smooth Function

Up to now, we have learned the gradient descent algorithm with linear search. The advantage of the algorithm is the simple interpretation. However, the linear search step involved in gradient descent algorithms makes more computational effort to find a proper step size. This also leads to difficulties in theoretical analysis (See Page 222).

Q: Whether exists a method to provide a proper step size s which can guarantee the convergence of the gradient descent algorithm without line search.

The answer is **Yes!** for the specific objective function.

Definition 1 *$f:\mathbb{R}^n\to\mathbb{R}$ is a β-smooth function if*

- *∇f exists which is continuous.*
- *For any $\mathbf{x}_1,\mathbf{x}_2\in dom(f)$,*
$$\|\nabla f(\mathbf{x}_1)-\nabla f(\mathbf{x}_2)\|_2\leq\beta\|\mathbf{x}_1-\mathbf{x}_2\|_2. \tag{1}$$
This means ∇f is a β-Lipshitz continuous function.

Let us show some examples:

- $f(\mathbf{x})=\langle\mathbf{b},A\mathbf{x}\rangle$ is a 0-smooth function.
- $f(\mathbf{x})=\frac{1}{2}\|\mathbf{b}-A\mathbf{x}\|^2$ is a $\lambda_{\max}(A^\top A)$-smooth function.

Lemma 1 *Let f be a β-smooth function, then for any $\mathbf{x}$ and $\mathbf{y}\in dom(f)$, then*
$$f(\mathbf{y})\leq f(\mathbf{x})+\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle+\frac{\beta}{2}\|\mathbf{y}-\mathbf{x}\|^2. \tag{2}$$

Proof 1 *Denote $g(t)=f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$, then $g(0)=f(\mathbf{x})$ and $g(1)=f(\mathbf{y})$. Then we have*
$$g(1)-g(0)=\int_0^1 g'(t)dt=\int_0^1\langle\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})),\mathbf{y}-\mathbf{x}\rangle dt. \tag{3}$$

Thus,

$$\begin{aligned}f(\mathbf{y})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle&=\int_0^1\langle\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle dt\\&\leq\int_0^1\|\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\nabla f(\mathbf{x})\|\|\mathbf{y}-\mathbf{x}\|dt\\&\leq\int_0^1 t\beta\|\mathbf{y}-\mathbf{x}\|^2dt=\frac{\beta}{2}\|\mathbf{y}-\mathbf{x}\|^2,\end{aligned}$$

where the first inequality due to Cahuchy inequality and the second inequality based on the definition of β-smooth function.

This Lemma tells us that if a function f is β-smooth, we could build a quadratic upper model at every point in the domain of f, namely, for any $\mathbf{x} \in dom(f)$ we can construct a function

$$m_{\mathbf{x}}(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}\|^2 \tag{4}$$

such that

- $m_{\mathbf{x}}(\mathbf{x}) = f(\mathbf{x})$,
- $f(\mathbf{y}) \leq m_{\mathbf{x}}(\mathbf{y})$ for any $\mathbf{y} \in dom(f)$.

Lemma 2 *f is β-smooth and $\nabla^2 f$ exists if and only if $\|\nabla^2 f\|_2 \leq \beta$.*

Hint: Consider the Taylor expansion of $f(\mathbf{y})$ at $\mathbf{x}$ is

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla^2 f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\top \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}),$$

where $0 \leq t \leq 1$.

Next, let us illustrate a gradient descent for β-smooth function in Algorithm 1. In the algorithm, it is notable that the step size is $s_t = \frac{1}{\beta}$ for all t.

Algorithm 1 Gradient Descent for β-smooth Function

1: **Input:** Given a initial starting point $\mathbf{x}^0 \in dom(f)$, a tolerance ϵ and $t = 0$
2: **while** $\|\nabla f(\mathbf{x}^t)\| \geq \epsilon$ **do**
3: $\quad \mathbf{x}^{t+1} = \mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t)$ and $t := t + 1$.
4: **end while**
5: **Output:** $\mathbf{x}^T$, where T is the last step index.

Example 1 *Let us consider the least squares problem again. That is*

$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 := f(\mathbf{x}).$$

We can compute that

- $\nabla f(\mathbf{x}) = A^\top(A\mathbf{x} - \mathbf{b})$,
- $\beta = \lambda_{\max}(A^\top A)$,
- *iterative step:*

$$\begin{aligned}
\mathbf{x}^{t+1} &= \mathbf{x}^t - \frac{1}{\lambda_{\max}(A^\top A)}\nabla f(\mathbf{x}^t) && (5)\\
&= \mathbf{x}^t - \frac{1}{\lambda_{\max}(A^\top A)}A^\top(A\mathbf{x}^t - \mathbf{b}) && (6)\\
&= \left(I - \frac{A^\top A}{\lambda_{\max}(A^\top A)}\right)\mathbf{x}^t + \frac{A^\top \mathbf{b}}{\lambda_{\max}(A^\top A)}. && (7)
\end{aligned}$$

This is an iterative algorithm and easy for implementation.

Q: *How to prove the algorithm converges???*

Theorem 1 *Suppose that* $\{\mathbf{x}^t\}_{t=0}^{\infty}$ *is generated by Algorithm 1 and the given tolerance* $\epsilon > 0$*, if* $T \geq \frac{2\beta(f(\mathbf{x}_0)-f^*)}{\epsilon^2}$*, then*

$$\min_{t=0,1,\ldots,T-1} \|\nabla f(\mathbf{x}^t)\| \leq \epsilon. \tag{8}$$

Proof 2 *Recall that* $m_t(\mathbf{x}) = f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t \rangle + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^t\|^2 \geq f(x)$ *is a quadratic function for any* $x \in dom(f)$*. And its minimizer is the solution of* $\nabla m_t(\mathbf{x}) = \nabla f(\mathbf{x}^t) + \beta(\mathbf{x}^t - \mathbf{x}) = 0$*. Thus,* $\mathbf{x}^* = \mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t) = \mathbf{x}^{t+1}$*, namely* $\mathbf{x}^{t+1}$ *is the global minimum of* $m_t(\mathbf{x})$*. Then*

$$f(\mathbf{x}^{t+1}) \leq m_t(\mathbf{x}^{t+1}) = f(\mathbf{x}^t) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}^t)\|^2 \leq m_t(\mathbf{x}^t) = f(\mathbf{x}^t). \tag{9}$$

So, we have that $f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) \leq -\frac{1}{2\beta}\|\nabla f(\mathbf{x}^t)\|^2$ *for all* t*. Inaddtion,*

$$f(\mathbf{x}^T) - f(\mathbf{x}^0) = \sum_{t=0}^{T-1}(f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)) \leq -\frac{1}{2\beta}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}^t)\|^2. \tag{10}$$

Therefore,

$$\frac{T}{2\beta}\min_{t=0,\ldots,T-1}\|\nabla f(\mathbf{x}^t)\|^2 \leq \frac{1}{2\beta}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}^t)\|^2 \leq f(\mathbf{x}^0) - f(\mathbf{x}^T) \leq f(\mathbf{x}_0) - f^*. \tag{11}$$

Based on this fact, we have

$$\min_{t=0,\ldots,T-1}\|\nabla f(\mathbf{x}^t)\|^2 \leq \sqrt{\frac{2\beta(f(\mathbf{x}^0) - f^*)}{T}}. \tag{12}$$

If $T \geq \frac{2\beta(f(\mathbf{x}_0)-f^*)}{\epsilon^2}$*, then*

$$\min_{t=0,1,\ldots,T-1} \|\nabla f(\mathbf{x}^t)\| \leq \epsilon. \tag{13}$$

Remark 1 *Tow facts should be discussed.*

- *Let us discuss the convergence property of* $\{\mathbf{x}^t\}_{t=1}^{\infty}$ *that generated by Algorithm 1. Assume that* f^* *exists, then based on the proof of Theorem 1, we have*

$$\frac{1}{2\beta}\sum_{t=0}^{\infty}\|\nabla f(\mathbf{x}^t)\|^2 \leq f(\mathbf{x}_0) - f^*. \tag{14}$$

  This implies $\{\|\nabla f(\mathbf{x}^t)\|\}_{t=0}^{\infty}$ *is a convergence sequence, and* $\lim_{t\to\infty}\|\nabla f(\mathbf{x}^t)\| = 0$*. So, if* $\mathbf{x}^t \to \mathbf{x}^*$*, then* $\mathbf{x}^*$ *is the stationary point of* f*.*

- *Local Minimum? Global Minimum???*

- *Let us discuss the convergence speed. Suppose that, take* $\epsilon = 10^{-2}$*, then according to Theorem 1, it should be* $T \geq 10^4$*. If we take* $\epsilon = 10^{-3}$*, then* $T \geq 10^6$*.*

- *Tooooo̊ slow!!!! Maybe* $T = O(1/\epsilon)$ *or* $T = O(1/\sqrt{\epsilon})$ *is better. Next subsection will show that the convex and* β*-smooth objective function can achieve the hopeful speed.*

References