CS292A Convex Optimization: Gradient Methods and Online Learning **Spring 2019**

Lecture 3: April 16

Lecturer: Yu-Xiang Wang *Scribes: Yijun Xiao*

3.1 Gradient descent method and interpretation

Consider the unconstrained smooth convex optimization:

$$\min_x f(x) \tag{3.1}$$

where f is convex and differentiable with $\mathrm{dom}(f) = \mathbb{R}^n$.

Gradient descent method is an iterative method that takes a step along the negative gradient direction at each iteration. It produces a sequence of points $x^{(k)}$ with $k = 1, 2, 3, \cdots$. The sequence follows the following update rule:

$$x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f(x^{(k)}) \tag{3.2}$$

where $t^{(k)} > 0$ is the step size chosen for the k-th iteration.

This formulation is equivalent to minimizing a quadratic approximation of f at $x^{(k)}$. To see this, consider the following quadratic approximation of $f(x)$:

$$f(x) \approx f(x^{(k)}) + \nabla f(x^{(k)})^\top (x - x^{(k)}) + \frac{1}{2t^{(k)}} \|x - x^{(k)}\|_2^2 \tag{3.3}$$

To minimize the right-hand side term in Equation 3.3, we take the first order derivative w.r.t x and set it to zero:

$$\nabla f(x^{(k)}) + \frac{1}{t^{(k)}} (x - x^{(k)}) = 0 \tag{3.4}$$

Clearly, this is equivalent to Equation 3.2.

3.2 Line search

The choice of t at each iteration is a key step in gradient descent. A large t could lead to divergence of the objective while a small one could make the algorithm very slow. Different approaches to choose t are referred to as line search methods. We introduce two basic line search methods: exact line search and backtracking line search in this section.

3.2.1 Exact line search

Exact line search aims to find the best possible step size at each iteration. In other words, it solves the following minimization problem at each iteration:

$$t = \text{argmin}_{s \geq 0} f(x - s\nabla f(x)) \tag{3.5}$$

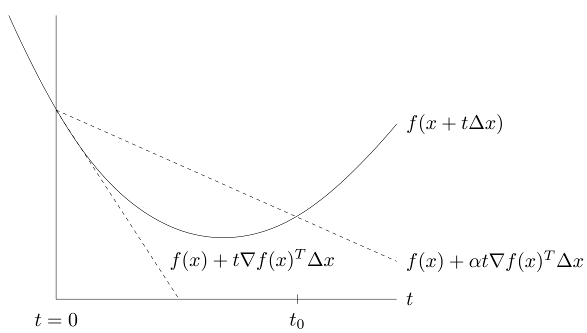
This method is used when the single variable minimization problem can be computed analytically or efficiently.

3.2.2 Backtracking line search

Backtracking line search chooses step size adaptively. It depends on two constants α, β with $0 < \alpha < 0.5, 0 < \beta < 1$. At each iteration:

Algorithm 1: Backtracking line search
Input: $\alpha \in (0, 0.5), \beta \in (0, 1)$, function f and gradient $\nabla f(x)$.
$t := t_{\text{init}}$
while $f(x - t\nabla f(x)) > f(x) - \alpha t \|\nabla f(x)\|_2^2$ **do**
| $t := \beta t$
end

Intuitively, backtracking line search iteratively decrease step size t by a factor of β until a "sufficient" descent is achieved. The sufficient descent is a fraction of decrease predicted by the linear extrapolation at current point.



For us $\Delta x = -\nabla f(x)$

Figure 3.1: An illustration of backtracking line search. The lower dashed line shows the linear extrapolation of f, and the upper dashed line has a slope a factor of α smaller. The backtracking condition is that f lies below the upper dashed line.

3.3 Convergence analysis

In this section, we give formal analysis of the convergence guarantee gradient descent methods provides in different conditions. First, we introduce the descent lemma:

Lemma 3.1 ***(Descent lemma)*** *Assume that f is differentiable with $dom(f) = \mathbb{R}^n$ and ∇f is Lipschitz continuous with constant $L > 0$. For $\forall t \le \frac{1}{L}$, the following inequality holds:*

$$f(x - t\nabla f(x)) \le f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2$$

Proof: Denote $x^+ = x - t\nabla f(x)$, by gradient Lipschitz condition, we have

$$\begin{aligned} f(x^+) &\le f(x) + \nabla f(x)^\top (x^+ - x) + \frac{L}{2}\|x^+ - x\|_2^2 \\ &= f(x) - \nabla f(x)^\top t\nabla f(x) + \frac{L}{2}t^2\|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{2t - Lt^2}{2}\|\nabla f(x)\|_2^2 \\ &\le f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \end{aligned} \tag{3.6}$$

■

Next we analyze the convergence rate mainly for two cases: f is convex and f is strongly convex. Results for first-order methods and gradient descent for non-convex functions are simply presented.

3.3.1 Convex function

Denote x^* as an optimal solution for f, and f^* as the optimal value. We have the following convergence guarantee for convex L-smooth function f:

Theorem 3.2 *Gradient descent with fixed step size $t \le 1/L$ satisfies*

$$f(x^{(k)}) - f^* \le \frac{\|x^{(0)} - x^*\|_2^2}{2tk}$$

Proof: By the convexity of f, we have:

$$f(x) \le f(x^*) + \nabla f(x)^\top (x - x^*) \tag{3.7}$$

Combined with Lemma 3.1, we have:

$$\begin{aligned} f(x^+) &\le f(x^*) + \nabla f(x)^\top (x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &= f(x^*) - \frac{1}{2t}\left(\|t\nabla f(x)\|_2^2 - 2t\nabla f(x)^\top (x - x^*) + \|x - x^*\|_2^2 - \|x - x^*\|_2^2\right) \\ &= f(x^*) - \frac{1}{2t}\left(\|t\nabla f(x) - (x - x^*)\|_2^2 - \|x - x^*\|_2^2\right) \\ &= f(x^*) - \frac{1}{2t}\left(\|x^+ - x^*\|_2^2 - \|x - x^*\|_2^2\right) \end{aligned} \tag{3.8}$$

Telescoping from 1 to k, we can conclude the proof:

$$\begin{aligned}\sum_{i=1}^{k} f(x^{(i)}) - kf^* &\le -\frac{1}{2t}\left(\|x^{(k)} - x^*\|_2^2 - \|x^{(0)} - x^*\|_2^2\right)\\ \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f^* &\le \frac{\|x^{(0)} - x^*\|_2^2}{2tk}\\ f(x^{(k)}) - f^* &\le \frac{\|x^{(0)} - x^*\|_2^2}{2tk} \qquad (3.9)\end{aligned}$$

The last step is due to the fact that $f(x^{(k)}) = \min_{i\in\{1,\cdots,k\}} f(x^{(i)}) \le \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)})$. ■

In other words, to find a ϵ-suboptimal solution, gradient descent takes $O(1/\epsilon)$ iterations for convex objectives.

The above analysis is for a fixed step size. When using backtracking line search, by descent lemma we know that the search will terminate with $\forall\alpha \in (0, 0.5)$ and the terminating step size $t > \frac{\beta}{L}$. Therefore, we have a same sub-linear convergence rate with backtracking line search by replacing t with $\frac{\beta}{L}$.

3.3.2 Strongly convex function

With the same notation, now assume f is m-strongly convex and L-smooth. Recall that m-strongly convexity of f indicates that $f(x) - \frac{m}{2}\|x\|_2^2$ is convex for some $m > 0$. We have the following convergence guarantee:

Theorem 3.3 *Gradient descent with fixed step size $t \le 2/(m + L)$ or with backtracking line search satisfies*

$$f(x^{(k)}) - f^* \le c^k \frac{L}{2}\|x^{(0)} - x^*\|_2^2$$

where $0 < c < 1$

Proof: By smoothness of f, we have

$$\begin{aligned}f(x^+) - f^* &\le \frac{L}{2}\|x^+ - x^*\|_2^2\\ &= \frac{L}{2}\|x - x^* - t\nabla f(x)\|_2^2\\ &= \frac{L}{2}\left(\|x - x^*\|_2^2 - 2t\nabla f(x)^\top(x - x^*) + t^2\|\nabla f(x)\|_2^2\right)\\ &\le \frac{L}{2}\left(\|x - x^*\|_2^2 - \frac{2tLm}{L+m}\|x - x^*\|_2^2 - \frac{2t}{L+m}\|\nabla f(x)\|_2^2 + t^2\|\nabla f(x)\|_2^2\right)\\ &= \frac{L}{2}\left[\left(1 - \frac{2tLm}{L+m}\right)\|x - x^*\|_2^2 + \left(t^2 - \frac{2t}{L+m}\right)\|\nabla f(x)\|_2^2\right]\\ &\le \frac{L}{2}\left(1 - \frac{2tLm}{L+m}\right)\|x - x^*\|_2^2 \qquad (3.10)\end{aligned}$$

Iteratively apply the above equation, we conclude our proof:

$$f(x^k) - f^* \le \left(1 - \frac{2tLm}{L+m}\right)^k \frac{L}{2}\|x^{(0)} - x^*\|_2^2 \qquad (3.11)$$

We could easily check that when $0 < t \leq 2/(m+L)$, $c = 1 - \frac{2tLm}{L+m} \in (0,1)$ ■

An alternative proof can be found in Section 3.4.2 when proving Theorem 3.7 as PL condition is a weaker condition than strong convexity.

We can see that the convergence rate is linear. We find ϵ-suboptimal point in $O(\log(1/\epsilon))$ iterations. A key note on the convergence rate is that the contraction factor c depends adversely on the condition number L/m. To see this, take $t = \frac{2}{m+L}$, then $c = 1 - \frac{4Lm}{(L+m)^2} = (\frac{L}{m} - 1)^2/(\frac{L}{m} + 1)^2$. With a higher condition number, the convergence gets slower.

3.3.3 First-order method

First-order methods are iterative methods which update $x^{(k)}$ in

$$x^{(0)} + \text{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \cdots, \nabla f(x^{(k-1)})\} \tag{3.12}$$

The span of the gradients is called a Krylov space.

For first-order methods, we have the following lower bound on convergence:

Theorem 3.4 ***(Nesterov)*** *For any $k \leq (n-1)/2$ and $x^{(0)}$, there is a function f in the problem class such that any first-order method satisfies*

$$f(x^{(k)}) - f^* \geq \frac{3L\|x^{(0)} - x^*\|_2^2}{32(k+1)^2}$$

3.3.4 Nonconvex function

Assume f is differentiable and L-smooth but nonconvex. Then we have the following upper bound on the gradient norm:

Theorem 3.5 *Gradient descent with fixed step size $t \leq 1/L$ satisfies*

$$\min_{i=0,\cdots,k} \|\nabla f(x^{(i)})\|_2 \leq \sqrt{\frac{2(f(x^{(0)}) - f^*)}{t(k+1)}}$$

This rate cannot be improved by any deterministic algorithm as shown in [CDHS17].

3.4 Generalization of strong convexity

Under certain conditions that are weaker than strong convexity, linear convergence can also be obtained. We introduce two such conditions in the following sections: restricted strong convexity (RSC) and Polyak-Łojasiewicz (PL) condition. We briefly touch on other conditions in Section 3.4.3.

3.4.1 Restricted strong convexity

Restricted strong convexity [NYWR09,ZC15] defines functions with the following conditions:

1. f is convex;
2. f obeys the restricted secant inequality (RSI):

$$(\nabla f(x) - \nabla f(x_{\text{prj}}))^\top (x - x_{\text{prj}}) \geq m\|x - x_{\text{prj}}\|_2^2, \forall x \tag{3.13}$$

   where x_{prj} is the projection of x to the solution set $\mathcal{X}^*$

It has been shown that RSC is sufficient and necessary for global linear convergence [ZC15].

Theorem 3.6 *If function f is L-smooth and m-RSC, then gradient descent method with fixed step size $t \leq 1/L$ converges linearly with*

$$\|x^{(k+1)} - x_{\text{prj}}^{(k+1)}\| \leq (1 - m/L)^{1/2}\|x^{(k)} - x_{\text{prj}}^{(k)}\|.$$

Conversely, assuming f has a unique solution x^* and gradient descent algorithm achieves a linear convergence rate from any $x^{(0)}$ with contraction ratio between 0 and 1, then f is RSC with $m > 0$.

3.4.2 Polyak-Łojasiewicz condition

Polyak-Łojasiewicz (PL) condition states that

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq m(f(x) - f^*), \forall x \tag{3.14}$$

We can show that an L-smooth function satisfying PL condition has linear convergence with gradient descent.

Theorem 3.7 *If f is L-smooth and satisfies PL condition with constant $m > 0$, the gradient descent method with fixed step size $t \leq 1/L$ converges linearly with*

$$f(x^{(k)}) - f^* \leq \frac{L}{2}(1 - mt)^k \|x^{(0)} - x^*\|_2^2$$

Proof: By descent lemma, we have:

$$f(x^{(k)}) \leq f(x^{(k-1)}) - \frac{t}{2}\|\nabla f(x^{(k-1)})\|_2^2 \tag{3.15}$$

Next apply PL condition:

$$\begin{aligned} f(x^{(k)}) &\leq f(x^{(k-1)}) - mt\left(f(x^{(k-1)}) - f^*\right) \\ f(x^{(k)}) - f^* &\leq (1 - mt)\left(f(x^{(k-1)}) - f^*\right) \\ &\leq (1 - mt)^k\left(f(x^{(0)}) - f^*\right) \\ &\leq \frac{L}{2}(1 - mt)^k\|x^{(0)} - x^*\|_2^2 \end{aligned} \tag{3.16}$$

■

Clearly, the contraction factor $1 - \frac{m}{L} \leq 1 - mt < 1$. Therefore this is linear convergence.

Note that PL condition is weaker than strongly convexity condition. If function f is m-strongly convex, then f satisfies PL condition with parameter m. The proof is as follows:

Proof: By strong convexity, we have for $\forall x$:

$$
\begin{aligned}
f^* &\geq f(x) + \nabla f(x)^\top (x^* - x) + \frac{m}{2}\|x^* - x\|_2^2 \\
&= f(x) + \frac{1}{2m}(m^2\|x^* - x\|_2^2 + 2m\nabla f(x)^\top (x^* - x) + \|\nabla f(x)\|_2^2 - \|\nabla f(x)\|_2^2) \\
&= f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2 + \frac{1}{2m}\|m(x^* - x) + \nabla f(x)\|_2^2 \\
&\leq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2 \qquad (3.17)
\end{aligned}
$$

which is equivalent to Equation 3.14. ■

3.4.3 Other conditions for linear convergence

There are some other conditions with linear convergence for gradient descent:

1. Quadratic growth (QG) condition

$$f(x) - f^* \geq \frac{m}{2}\|x_{\text{prj}} - x\|^2, \forall x$$

2. Error bounds (EB) condition

$$\|\nabla f(x)\| \geq m\|x_{\text{prj}} - x\|, \forall x$$

It has been shown [KNS16] that for smooth functions (RSI)→(EB)≡(PL)→(QG). If f is also convex, then (RSC)≡(PL)≡(QG)≡(EB).

References

[CDHS17] Y. Carmon, J.C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points i", *arXiv preprint arXiv:1710.11606*, 2017.

[NYWR09] S. Negahban, B. Yu, M.J. Wainwright, and P.K. Ravikumar, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers", *In Advances in Neural Information Processing Systems*, 2009, pp. 1348–1356.

[ZC15] H. Zhang and L. Cheng, "Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization", *Optimization Letters 9.5*, 2015, pp. 961–979.

[KNS16] H. Karimi, J. Nutini and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak- lojasiewicz condition", *ECML-PKDD*, 2016.