ECE586 MDPs and Reinforcement Learning **University of Illinois at Urbana-Champaign**
Spring 2019 **Acknowledgment: R. Srikant's Notes and Discussions**

Lecture 4: Steepest and Gradient Descent-Part II

Instructor: Dimitrios Katselis *Scribe: William Wei, Aditya Deshmukh*

4.1 Introduction

In this lecture, we discuss the convergence properties of steepest descent with constant stepsize, under various assumptions on the loss function f to be optimized. We previously considered the scenario where $\nabla f(x)$ satisfied a Lipschitz continuity condition and we were able to show convergence of the steepest descent to a stationary point of f. We now consider the cases where f not only has a Lipschitz gradient, but is also convex or strongly convex, resulting in stronger convergence results and bounds on the rate of convergence.

4.2 Constant Stepsize Steepest Descent: L-smooth Convex Loss Function

In this section, we analyze the case in which the loss function f is convex and has a Lipschitz gradient. We first begin with the global underestimator property of convex functions, which is useful in proving convergence results.

Theorem 4.1 (Global underestimator property)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Then,* equivalently $\forall x, y \in \mathbb{R}^n$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

Proof. We first show the implication "f convex" $\Rightarrow$ "$f(y) \geq f(x) + \nabla f(x)^T (y - x), \forall x, y \in \mathbb{R}^n$". Let $x, y \in \mathbb{R}^n$. By the definition of convexity:

$$f(\lambda y + (1 - \lambda)x) \leq \lambda f(y) + (1 - \lambda) f(x), \ \ \forall \lambda \in [0, 1].$$

After rewriting, we get that

$$f(x + \lambda(y - x)) \leq f(x) + \lambda(f(y) - f(x)), \ \ \forall \lambda \in [0, 1].$$

Consequently, $\forall \lambda \in (0, 1]$,

$$f(y) - f(x) \geq \frac{f(x + \lambda(y - x)) - f(x)}{\lambda}.$$

Letting $\lambda \to 0$, we obtain:

$$f(y) - f(x) \geq \nabla f(x)^T (y - x),$$

where the right hand side is due to the definition of the directional derivative for a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$.

For the *reverse implication*: Let $x, y \in \mathbb{R}^n$ and consider a point z on the line segment joining x, y:

$$z = \lambda x + (1 - \lambda) y, \ \ \lambda \in [0, 1].$$

We then have:

$$f(x) \geq f(z) + \nabla f(z)^T(x-z),$$
$$f(y) \geq f(z) + \nabla f(z)^T(y-z).$$

Multiplying with λ and $1-\lambda$, respectively, and adding we obtain:

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) &\geq f(z) + \nabla f(z)^T(\lambda x + (1-\lambda)y - z) \\ &= f(z) \\ &= f(\lambda x + (1-\lambda)y).\end{aligned}$$

Therefore, f is convex.

□

To justify the title of this section, we give the following definition:

Definition 4.2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth if*

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

Equivalently, f is L-smooth if its gradient is L-Lipschitz:

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

We now proceed to show that if the function f is convex, has a Lipschitz gradient and the set of minimizers of f is nonempty, then for sufficiently small constant stepsize, steepest descent converges at rate $1/k$. Here, k represents the iteration index.

Theorem 4.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, continuously differentiable and L-smooth. Suppose that $\exists x^* \in \mathbb{R}^n$ such that $f(x^*) = \min_{x \in \mathbb{R}^n} f(x) > -\infty$. Then, the steepest descent iterates $\{x_k\}$, given by $x_{k+1} = x_k - \varepsilon \nabla f(x_k)$ with $0 < \varepsilon < \frac{2}{L}$, satisfy*

$$f(x_k) \xrightarrow[k\to\infty]{} f(x^*).$$

Moreover,

$$f(x_k) - f(x^*) = O\left(\frac{1}{k}\right).$$

Proof. We begin by upper bounding the distance between the iterates and the optimum point x^*. For any $k \in \mathbb{N}$ ($\mathbb{N}$ contains 0) ,

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - \varepsilon\nabla f(x_k) - x^*\|^2 \\ &= \|x_k - x^*\|^2 + \varepsilon^2\|\nabla f(x_k)\|^2 - 2\varepsilon\nabla f(x_k)^T(x_k - x^*) \\ &\leq \|x_k - x^*\|^2 + \varepsilon^2\|\nabla f(x_k)\|^2 - 2\varepsilon(f(x_k) - f(x^*)). \qquad (4.1)\end{aligned}$$

The last inequality (4.1) follows from Theorem 4.1. Rearranging (4.1), we get:

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

Summing from $k = 0$ to $k = N$, we obtain:

$$2\varepsilon\sum_{k=0}^{N}(f(x_k) - f(x^*)) \leq \|x_0 - x^*\|^2 - \|x_{N+1} - x^*\|^2 + \varepsilon^2\sum_{k=0}^{N}\|\nabla f(x_k)\|^2. \qquad (4.2)$$

From Theorem 3.2 in Lecture 3 notes, we have that

$$\sum_{k=0}^{N} \|\nabla f(x_k)\|^2 \leq \frac{f(x_0) - f(x^*)}{\varepsilon\left(1 - \frac{1}{2}\varepsilon L\right)}. \tag{4.3}$$

Plugging (4.3) in (4.2), we get:

$$\begin{aligned} 2\varepsilon \sum_{k=0}^{N} (f(x_k) - f(x^*)) &\leq \|x_0 - x^*\|^2 - \|x_{N+1} - x^*\|^2 + \frac{\varepsilon(f(x_0) - f(x^*))}{1 - \frac{1}{2}\varepsilon L} \\ &\leq \|x_0 - x^*\|^2 + \frac{\varepsilon(f(x_0) - f(x^*))}{1 - \frac{1}{2}\varepsilon L}. \end{aligned} \tag{4.4}$$

Recall from the proof of Theorem 3.2 in Lecture 3 notes that $\forall k \in \mathbb{N}$,

$$f(x_{k+1}) - f(x_k) \leq -\varepsilon\left(1 - \frac{1}{2}\varepsilon L\right)\|\nabla f(x_k)\|^2 \leq 0.$$

Hence, $\forall k \in \mathbb{N}$ we have:

$$f(x_{k+1}) \leq f(x_k).$$

Consequently, this implies that

$$\sum_{k=0}^{N} (f(x_k) - f(x^*)) \geq N(f(x_N) - f(x^*)). \tag{4.5}$$

Using (4.5) in (4.4), we get

$$2\varepsilon N(f(x_N) - f(x^*)) \leq \|x_0 - x^*\|^2 + \frac{\varepsilon(f(x_0) - f(x^*))}{1 - \frac{1}{2}\varepsilon L}.$$

Thus, $\forall N \in \mathbb{N}$,

$$0 \leq f(x_N) - f(x^*) \leq \frac{1}{2\varepsilon N}\left[\|x_0 - x^*\|^2 + \frac{\varepsilon(f(x_0) - f(x^*))}{1 - \frac{1}{2}\varepsilon L}\right] = O\left(\frac{1}{N}\right).$$

We conclude that $f(x_k) \xrightarrow[k\to\infty]{} f(x^*)$ at rate $1/k$.

□

Remark 4.4. *A deficiency of the above theorem is that we need to know either L or an upper bound on L to be able to choose $\varepsilon \in \left(0, \frac{2}{L}\right)$.*

4.3 Strongly convex functions

In this section, we give some background material on strongly convex functions.

Definition 4.5 (Strong convexity)**.** *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be strongly convex with parameter $m > 0$ if*

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq m\|x - y\|^2, \ \ \forall x, y \in \mathbb{R}^n. \tag{4.6}$$

Moreover, a function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with parameter $m > 0$ if and only if $f(x) - \frac{m}{2}\|x\|^2$ is a convex function.

Note: Roughly speaking, strong convexity means that f is "as least as convex" as a quadratic function.

Remarks:

1. Strong convexity $\Rightarrow$ strict convexity $\Rightarrow$ convexity.
2. Using the Cauchy-Schwarz inequality to upper bound the inner product in (4.6), we obtain that strong convexity with parameter $m > 0$ implies that

$$\|\nabla f(x) - \nabla f(y)\| \geq m\|x - y\|, \ \ \forall x, y \in \mathbb{R}^n.$$

If in addition ∇f is L-Lipschitz, then last inequality implies that $L \geq m$.

The following characterization is particularly useful when a strongly convex function is twice differentiable.

Proposition 4.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. Then, the following are equivalent:*

1. *f is strongly convex with parameter $m > 0$.*
2. *$\forall x \in \mathbb{R}^n$, $\nabla^2 f(x) \succeq mI$.*
3. *$\forall x, y \in \mathbb{R}^n$, $f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|^2$.*

Here, $\nabla^2 f(x)$ represents the Hessian matrix of f at x, and I represents the $n \times n$ identity matrix. $A \succeq B$ denotes that $A - B$ is positive semi-definite.

Proof. $(1 \implies 2)$ For any $x, v \in \mathbb{R}^n$, we have

$$\lim_{h \to 0} \frac{\nabla f(x + hv) - \nabla f(x)}{h} = \nabla^2 f(x)v.$$

Using the definition (4.6) of strong convexity , we get

$$\begin{aligned} v^T \nabla^2 f(x) v &= \lim_{h \to 0} \frac{(\nabla f(x + hv) - \nabla f(x))^T (hv)}{h^2} \\ &\geq \lim_{h \to 0} \frac{m\|hv\|^2}{h^2} \\ &= m\|v\|^2. \end{aligned}$$

Thus, $\forall x, v \in \mathbb{R}^n$,

$$v^T(\nabla^2 f(x) - mI)v \geq 0.$$

This implies $\forall x \in \mathbb{R}^n$, $\nabla^2 f(x) - mI$ is positive semi-definite.

$(2 \implies 3)$ Let $x, y \in \mathbb{R}^n$. Using second-order Taylor expansion around x, we have that

$$f(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x), \tag{4.7}$$

where $z = x + \lambda(y - x)$ and $\lambda \in [0, 1]$. By assumption, $\nabla^2 f(z) \succeq mI$. Hence,

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T mI(y - x) \\ &= f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|^2. \end{aligned}$$

($3 \implies 1$) For any $x, y \in \mathbb{R}^n$, we have that

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|y-x\|^2, \tag{4.8}$$

and similarly

$$f(x) \geq f(y) + \nabla f(y)^T (x-y) + \frac{m}{2}\|y-x\|^2. \tag{4.9}$$

Adding (4.8) and (4.9), we get

$$(\nabla f(x) - \nabla f(y))^T (x-y) \geq m\|x-y\|^2.$$

□

Sometimes, functions that satisfy the above inequalities are called *elliptic*.

Univariate f: If $f : \mathbb{R} \to \mathbb{R}$ and f is twice continuously differentiable, then:

- f is convex $\Leftrightarrow f''(x) \geq 0,\ \forall x \in \mathbb{R}$.
- f is strictly convex *if* $f''(x) > 0,\ \forall x \in \mathbb{R}$. (Note that this condition is sufficient for strict convexity but not necessary.)
- f is strongly convex $\Leftrightarrow f''(x) \geq m > 0,\ \forall x \in \mathbb{R}$.

Note: In general, it is easier to work with strongly convex functions than convex or strictly convex functions. Moreover, strongly convex functions have unique minima on compact sets, a common property with strictly convex functions.

4.4 Constant Stepsize Steepest Descent: L-smooth Strongly Convex Loss function

We now proceed to analyze the case where f is strongly convex and has a Lipschitz gradient. We show that we can get stronger convergence results for the constant stepsize steepest descent.

Theorem 4.7. *If f is strongly convex with parameter m and L-smooth, the constant stepsize steepest descent iterate x_k converges to x^* at a* linear *or a* geometric *rate*[1]*, where ε is the corresponding stepsize.*

Proof.

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - \varepsilon \nabla f(x_k) - x^*\|^2 \\
&= \|x_k - x^* - \varepsilon(\nabla f(x_k) - \nabla f(x^*))\|^2 && (\nabla f(x^*) = 0) \\
&= \|x_k - x^*\|^2 + \varepsilon^2 \|\nabla f(x_k) - \nabla f(x^*)\|^2 - 2\varepsilon \nabla f(x_k)^T (x_k - x^*) \\
&\leq \|x_k - x^*\|^2 + \varepsilon^2 L^2 \|x_k - x^*\|^2 + 2\varepsilon \nabla f(x_k)^T (x^* - x_k). && (4.10)
\end{aligned}$$

Recall that f is strongly convex, i.e., f is necessarily convex. Therefore, using the global underestimator property we have that

$$f(x^*) \geq f(x_k) + \nabla f(x_k)^T (x^* - x_k) \Leftrightarrow \nabla f(x_k)^T (x^* - x_k) \leq f(x^*) - f(x_k). \tag{4.11}$$

[1] I.e., the distance $\|x_k - x^*\|$ is decreased at least as fast as the geometric progression $\{\rho^k \|x_0 - x^*\|\}$ for some $\rho \in (0,1)$. This is called **linear** or **geometric** convergence.

By (4.10) and (4.11), we get,

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\|^2 + \varepsilon^2 L^2 \|x_k - x^*\|^2 + 2\varepsilon(f(x^*) - f(x_k)). \tag{4.12}$$

Moreover, since f is strongly convex,

$$f(x_k) \geq f(x^*) + \nabla f(x^*)^T(x_k - x^*) + \frac{m}{2}\|x_k - x^*\|^2 \Leftrightarrow -\frac{m}{2}\|x_k - x^*\|^2 \geq f(x^*) - f(x_k). \tag{4.13}$$

By (4.12) and (4.13),

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 + \varepsilon^2 L^2 \|x_k - x^*\|^2 - \varepsilon m \|x_k - x^*\|^2 \\ &= (1 + \varepsilon^2 L^2 - \varepsilon m)\|x_k - x^*\|^2. \end{aligned} \tag{4.14}$$

Iterating (4.14), we have,

$$\|x_k - x^*\|^2 \leq (1 + \varepsilon^2 L^2 - \varepsilon m)^k \|x_0 - x^*\|^2.$$

Therefore, $x_k \xrightarrow[k\to\infty]{} x^*$ geometrically fast.

Note that,

$$1 + \varepsilon^2 L^2 - \varepsilon m = \left(L\varepsilon - \frac{m}{2L}\right)^2 + 1 - \left(\frac{m}{2L}\right)^2 \geq 1 - \left(\frac{m}{2L}\right)^2.$$

Equality is achieved if

$$L\varepsilon - \frac{m}{2L} = 0 \Leftrightarrow \varepsilon = \frac{m}{2L^2}.$$

Therefore, the best convergence rate of x_k to x^* is given by the progression:

$$\|x_k - x^*\|^2 \leq \left[1 - \left(\frac{m}{2L}\right)^2\right]^k \|x_0 - x^*\|^2.$$

□

This may be bad if $\frac{m}{L} \ll 1$ or equivalently $\frac{L}{m} \gg 1$, i.e., if the condition number of the Hessian $\nabla^2 f(x)$ is very large.

Condition Number: If $mI \preceq \nabla^2 f(x) \preceq LI$, the ratio L/m is an upper bound on the condition number of the matrix $\nabla^2 f(x)$, i.e., the ratio of its largest eigenvalue to its smallest eigenvalue. This upper bound translates to an upper bound on the condition number of the sublevel sets of f $\{x \in \mathbb{R}^n | f(x) \leq \beta\}$ for $f(x^*) < \beta \leq f(x_0)$. The condition number of a convex set (the sublevel sets of f are convex sets, since f is strongly convex[2]) gives a measure of its *anisotropy* or *eccentricity*. If the condition number is small, the set is nearly spherical. If the condition number is large, the set is anisotropic, i.e., wider in some directions than in others. From the above theorem, we see that the condition number of the sublevel sets of f (which is bounded by L/m) has a strong effect on the efficiency of the constant stepsize steepest descent algorithm.

[2] Note that “f convex” implies convex sublevel sets. Nevertheless, a function whose sublevel sets are convex may fail to be convex. A function whose sublevel sets are convex is called a **quasiconvex** function.

Remark 4.8. *Consider a twice differentiable function f that is strongly convex with parameter m, i.e.,*

$$mI \preceq \nabla^2 f(x),$$

and L-smooth:

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

Therefore,

$$mI \preceq \nabla^2 f(x) \preceq LI. \tag{4.15}$$

Rearranging (4.7), we have that

$$\frac{1}{2}(y-x)^T \nabla^2 f(z)(x-y) = f(y) - f(x) - \nabla f(x)^T (y-x).$$

Applying the bounds in (4.15), we get:

$$\frac{1}{2}m\|y-x\|^2 \leq f(y) - f(x) - \nabla f(x)^T(y-x) \leq \frac{1}{2}L\|y-x\|^2.$$

Thus,

$$\frac{1}{2}m\|x_k - x^*\|^2 \leq f(x_k) - f(x^*) \leq \frac{1}{2}L\|x_k - x^*\|^2, \tag{4.16}$$

since $\nabla f(x^*) = 0$.

In Theorem 4.7 we proved that $x_k \xrightarrow[k\to+\infty]{} x^*$ at a linear or geometric rate under the assumptions of strong convexity and L-smoothness. From (4.16), we can see that bounds on $\|x_k - x^*\|$ and $f(x_k) - f(x^*)$ can be related to each other.