DSCC 435 Optimization for Machine Learning Lecture 3

Gradient Methods

Lecturer: Jiaming Liang *September 4, 2025*

1 Convex smooth functions

1.1 Smoothness

Definition 1. *A differentiable function is called L-smooth on $\mathbb{R}^n$ if its gradient is L-Lipschitz continuous on $\mathbb{R}^n$, i.e.,*

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

Lemma 1. *Let f be a convex and differentiable function. Then, the following statements are equivalent:*

(i) f is L-smooth;

(ii) for all $x, y \in \mathbb{R}^n$,

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \le \frac{L}{2}\|x - y\|^2;$$

(iii) for all $x, y \in \mathbb{R}^n$,

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \le f(y);$$

(iv) for all $x, y \in \mathbb{R}^n$,

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \le \langle \nabla f(x) - \nabla f(y), x - y\rangle;$$

(v) for all $x, y \in \mathbb{R}^n$,

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \le L\|x - y\|^2.$$

Proof. **(i) $\implies$ (ii)**

Consider $g(t) = f(x + t(y - x))$ for $t \in [0, 1]$. Observe that $g(0) = f(x)$, $g(1) = f(y)$, and

$$g'(t) = \langle \nabla f(x + t(y - x)), y - x\rangle.$$

Then, we have

$$g(1) = g(0) + \int_0^1 g'(\tau)d\tau,$$

equivalently,

$$f(y)=f(x)+\int_0^1\langle\nabla f(x+\tau(y-x)),y-x\rangle d\tau. \tag{1}$$

It follows from the Cauchy-Schwarz inequality and the assumption that f is L-smooth that

$$\begin{aligned}f(y)-f(x)-\langle\nabla f(x),y-x\rangle&=\int_0^1\langle\nabla f(x+\tau(y-x))-\nabla f(x),y-x\rangle d\tau\\&\leq\int_0^1 L\tau\|y-x\|^2d\tau=\frac{L}{2}\|y-x\|^2.\end{aligned}$$

(ii) $\implies$ (iii)

Fix $x_0\in\mathbb{R}^n$ and consider a convex function $\phi(y)=f(y)-\langle\nabla f(x_0),y\rangle$. Note that the optimal solution is $y_*=x_0$. Using (ii), we have

$$\begin{aligned}\phi(y_*)=\min_{x\in\mathbb{R}^n}\phi(x)&\leq\min_{x\in\mathbb{R}^n}\left\{\phi(y)+\langle\nabla\phi(y),x-y\rangle+\frac{L}{2}\|x-y\|^2\right\}\\&=\min_{r\geq 0}\left\{\phi(y)-r\|\nabla\phi(y)\|+\frac{L}{2}r^2\right\}=\phi(y)-\frac{1}{2L}\|\nabla\phi(y)\|^2.\end{aligned}$$

Hence, (iii) holds in view of $\nabla\phi(y)=\nabla f(y)-\nabla f(x_0)$.

(iii) $\implies$ (iv)

We obtain (iv) by adding two copies of (iii) with x and y interchanged.

(iv) $\implies$ (i)

This is simply by (iv) and Cauchy-Schwarz inequality.

(ii) $\implies$ (v)

We obtain (v) by adding two copies of (ii) with x and y interchanged.

(v) $\implies$ (ii)

Using the identity (1) and (v), we have

$$\begin{aligned}f(y)-f(x)-\langle\nabla f(x),y-x\rangle&=\int_0^1\langle\nabla f(x+\tau(y-x))-\nabla f(x),y-x\rangle d\tau\\&\leq\int_0^1 L\tau\|y-x\|^2d\tau=\frac{L}{2}\|y-x\|^2.\end{aligned}$$

□

Remark 1. *Note that the convexity of f is explicitly used in the proof of (ii) $\implies$ (iii), and implicitly used in the proof of (iii) $\implies$ (iv). The convexity is necessary in Lemma 1, since even if smoothness disappears, i.e., $L=\infty$, the resulting inequalities of both (iii) and (iv) still imply that f is convex.*

1.2 Gradient method

Algorithm 1 Gradient method

Input: Initial point $x_0 \in \mathbb{R}^n$
for $k \leftarrow 0, \cdots, K-1$ **do**
  Step 1. Choose $h_k > 0$.
  Step 2. Compute $x_{k+1} = x_k - h_k \nabla f(x_k)$.
end for
Output: x_K

Theorem 1. *Assume f is convex and L-smooth and choose $h_k = h \in (0, 2/L)$ for every $k \geq 0$. Then, the Gradient Method generates a sequence of points $\{x_k\}$ satisfying*

$$f(x_k) - f_* \leq \frac{2[f(x_0) - f_*]\|x_0 - x_*\|^2}{2\|x_0 - x_*\|^2 + kh(2 - Lh)[f(x_0) - f_*]}, \quad \forall k \geq 0.$$

Proof. Let $r_k := \|x_k - x_*\|$. Then, we get

$$\begin{aligned} r_{k+1}^2 &= \|x_k - x_* - h\nabla f(x_k)\|^2 \\ &= r_k^2 - 2h \langle \nabla f(x_k), x_k - x_* \rangle + h^2 \|\nabla f(x_k)\|^2 \\ &= r_k^2 - 2h \langle \nabla f(x_k) - \nabla f(x_*), x_k - x_* \rangle + h^2 \|\nabla f(x_k)\|^2 . \end{aligned}$$

Using Lemma 1(iv), we have

$$r_{k+1}^2 \leq r_k^2 - h\left(\frac{2}{L} - h\right)\|\nabla f(x_k)\|^2 .$$

Therefore, $r_k \leq r_0$. Using Lemma 1(i), we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \alpha \|\nabla f(x_k)\|^2 \end{aligned}$$

where $\alpha = h(1 - Lh/2)$. This inequality gives the descent property of the function value. Define $\Delta_k = f(x_k) - f_*$. Then,

$$\Delta_k \leq \langle \nabla f(x_k), x_k - x_* \rangle \leq r_k \|\nabla f(x_k)\| \leq r_0 \|\nabla f(x_k)\| .$$

Thus,

$$\Delta_{k+1} \leq \Delta_k - \frac{\alpha}{r_0^2}\Delta_k^2.$$

Dividing the above inequality by $\Delta_{k+1}\Delta_k$, we have

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\alpha}{r_0^2}\frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\alpha}{r_0^2}.$$

Summing up these inequalities, we obtain

$$\frac{1}{\Delta_k} \geq \frac{1}{\Delta_0} + \frac{\alpha k}{r_0^2}.$$

The conclusion follows by inverting the above inequalities. $\square$

Choosing $h = 1/L$ maximizes $h(2-Lh)$ and hence the denominator, so it is the optimal stepsize. We have the following convergence rate of the Gradient Method:

$$f(x_k) - f_* \leq \frac{2L[f(x_0) - f_*]\|x_0 - x_*\|^2}{2L\|x_0 - x_*\|^2 + k[f(x_0) - f_*]}, \quad \forall k \geq 0.$$

Again, using the smoothness of f, we have

$$f(x_0) \leq f_* + \langle \nabla f(x_*), x_0 - x_* \rangle + \frac{L}{2}\|x_0 - x_*\|^2 = f_* + \frac{L}{2}\|x_0 - x_*\|^2.$$

Thus, we have the following result.

Corollary 1. *Assume f is convex and L-smooth and choose $h_k = h = 1/L$ for every $k \geq 0$. Then,*

$$f(x_k) - f_* \leq \frac{2L\|x_0 - x_*\|^2}{k+4}, \quad \forall k \geq 0.$$

Theorem 2. *If f is diffrentialble and convex on $\mathbb{R}^n$ and $\nabla f(x_*) = 0$, then x_* is the global minimum of f on $\mathbb{R}^n$.*

Proof. It follows from the convexity of f that for every $x \in \mathbb{R}^n$,

$$f(x) \geq f(x_*) + \langle \nabla f(x_*), x - x_* \rangle = f(x_*).$$

$\square$

Hence, it is also meaningful in finding a point with a small norm of the gradient: $\|\nabla f(x)\| \leq \varepsilon$.

2 Strongly convex and smooth functions

Definition 2. *A proper extended real-valued function f is μ-strongly convex if*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \lambda(1-\lambda)\frac{\mu}{2}\|x-y\|^2 \quad \textit{for all} \quad x, y \in \mathbb{R}^n, \lambda \in [0,1].$$

Lemma 2. *A function f is μ-strongly convex if and only if $f - \frac{\mu}{2}\|\cdot\|_2^2$ is convex.*

Theorem 3. *A continuously differentiable function f is μ-strongly convex on $\mathbb{R}^n$ if and only if for any $x, y \in \mathbb{R}^n$ we have*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x-y\|^2.$$

Proof. This theorem follows from Lemma 2 and Theorem 2 of Lecture 2. $\square$

Theorem 4. *A twice continuously differentiable function f is μ-strongly convex on $\mathbb{R}^n$ if and only if for any $x \in \mathbb{R}^n$ we have*

$$\nabla^2 f(x) \succeq \mu I.$$

Proof. This theorem follows from Lemma 2 and Theorem 3 of Lecture 2. $\square$

Lemma 3. *If a continuously differentiable function f is μ-strongly convex on $\mathbb{R}^n$, then we have*

(i) for all $x, y \in \mathbb{R}^n$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\mu} \|\nabla f(x) - \nabla f(y)\|^2;$$

(ii) for all $x, y \in \mathbb{R}^n$,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \frac{1}{\mu} \|\nabla f(x) - \nabla f(y)\|^2;$$

(iii) for all $x, y \in \mathbb{R}^n$,

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

Proof. The proof is left as a HW problem. $\square$

Lemma 4. *Assume f is μ-strongly convex and L-smooth. For every $x, y \in \mathbb{R}^n$, we have*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Proof. The proof is left as a HW problem. $\square$

We are now ready to estimate the performance of the Gradient Method on the class of strongly convex functions.

Theorem 5. *Assume f is μ-strongly convex and L-smooth and choose $0 \leq h \leq 2/(\mu + L)$. Then, the Gradient Method generates a sequence $\{x_k\}$ such that*

$$\|x_k - x_*\|^2 \leq \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|x_0 - x_*\|^2.$$

If $h = 2/(\mu + L)$, then

$$\|x_k - x_*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x_*\|,$$

and

$$f(x_k) - f_* \leq \frac{L}{2} \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x_0 - x_*\|^2, \quad \forall k \geq 0$$

where $\kappa = L/\mu$.

Proof. Let $r_k := \|x_k - x_*\|$. Then, we get

$$\begin{aligned} r_{k+1}^2 &= \|x_k - x_* - h\nabla f(x_k)\|^2 \\ &= r_k^2 - 2h\langle \nabla f(x_k), x_k - x_* \rangle + h^2 \|\nabla f(x_k)\|^2 \\ &= r_k^2 - 2h\langle \nabla f(x_k) - \nabla f(x_*), x_k - x_* \rangle + h^2 \|\nabla f(x_k)\|^2 . \end{aligned}$$

Using Lemma 4, we have

$$r_{k+1}^2 \le \left(1 - \frac{2h\mu L}{\mu + L}\right) r_k^2 + h\left(h - \frac{2}{\mu + L}\right) \|\nabla f(x_k)\|^2.$$

It follows from the assumption that $0 \le h \le 2/(\mu + L)$ that

$$r_{k+1}^2 \le \left(1 - \frac{2h\mu L}{\mu + L}\right) r_k^2.$$

So the first conclusion holds by applying the above inequality recursively. The second conclusion holds by plugging in $h = 2/(\mu + L)$. The last conclusion follows from Lemma 1(ii) and the second conclusion. □

Note that the fastest rage of convergence is achieved for $h = 2/(\mu + L)$. In this case, we have

$$\|x_k - x_*\| \le \left(\frac{L - \mu}{L + \mu}\right)^k \|x_0 - x_*\|.$$

3 Optimization with constraints

Let us consider now a smooth optimization problem with the set constraint:

$$\min_{x \in Q} f(x) \tag{2}$$

where Q is a closed convex set.

In the unconstrained case, the optimality condition is

$$\nabla f(x) = 0.$$

But this condition does not work with the set constraint. Consider the following univariate minimization problem:

$$\min_{x \ge 0} x.$$

Here $Q = \{x \in \mathbb{R} : x \ge 0\}$ and $f(x) = x$. Note that $x_* = 0$ but $f'(x_*) = 1 > 0$.

Theorem 6. *Let f be convex and differentiable and Q be closed and convex. A point x_* is a solution to* (2) *if and only if*

$$\langle \nabla f(x_*), x - x_* \rangle \geq 0 \tag{3}$$

for all $x \in Q$.

Proof. Indeed, if (3) is true, then

$$f(x) \geq f(x_*) + \langle \nabla f(x_*), x - x_* \rangle \geq f(x_*)$$

for all $x \in Q$. On the other hand, let x_* be a solution to (2). Assume that there exists some $x \in Q$ such that

$$\langle \nabla f(x_*), x - x_* \rangle < 0.$$

Consider the function

$$\phi(\alpha) = f(x_* + \alpha(x - x_*)), \quad \alpha \in [0, 1].$$

Note that

$$\phi(0) = f(x_*), \quad \phi'(0) = \langle \nabla f(x_*), x - x_* \rangle < 0.$$

Therefore, for α small enough we have

$$f(x_* + \alpha(x - x_*)) = \phi(\alpha) < \phi(0) = f(x_*).$$

This is a contradiction. □

The next statement is often addressed as the growth property of strongly convex functions.

Theorem 7. *If f is μ-strongly convex, then for any $x \in Q$, we have*

$$f(x) \geq f(x_*) + \frac{\mu}{2}\|x - x_*\|^2.$$

Proof. Indeed, by strong convexity and Theorem 6, we have

$$\begin{aligned} f(x) &\geq f(x_*) + \langle \nabla f(x_*), x - x_* \rangle + \frac{\mu}{2}\|x - x_*\|^2 \\ &\geq f(x_*) + \frac{\mu}{2}\|x - x_*\|^2. \end{aligned}$$

□

Theorem 8. *Let f be μ-strongly convex with $\mu > 0$ and the set Q is closed and convex. Then there exists a unique solution x_* to* (2).

Proof. The proof is left as a HW problem. □

3.1 Minimization over simple sets

Let us consider the following minimization problem over a set

$$\min_{x \in Q} f(x)$$

where f is μ-strongly convex and L-smooth, and Q is a closed convex set. We assume that Q is simple enough so that projection onto Q is easy to compute.

Definition 3. *Let Q be a closed set and $x_0 \in \mathbb{R}^n$. Define*

$$\operatorname{proj}_Q(x_0) = \arg\min_{x \in Q} \|x - x_0\|.$$

We call $\operatorname{proj}_Q(x_0)$ the Euclidean projection of the point x_0 onto the set Q.

Lemma 5. *For any two points x_1 and $x_2 \in \mathbb{R}^n$, we have*

$$\|\operatorname{proj}_Q(x_1) - \operatorname{proj}_Q(x_2)\| \leq \|x_1 - x_2\|.$$

Proof. The proof is left as a HW problem. □

Theorem 9. *Let x_* be an optimal solution to (2). Then, for any $h > 0$, we have*

$$\operatorname{proj}_Q(x_* - h\nabla f(x_*)) = x_*.$$

Proof. The proof is left as a HW problem. □

Examples

- nonnegative orthant $Q = \mathbb{R}^n_+$,
$$\operatorname{proj}_Q(x) = [x]_+;$$
- box $Q = \operatorname{Box}[\ell, u]$,
$$\operatorname{proj}_Q(x) = (\min\{\max\{x_i, \ell_i\}, u_i\})_{i=1}^n;$$
- affine set $Q = \{x \in \mathbb{R}^n : Ax = b\}$,
$$\operatorname{proj}_Q(x) = x - A^T \left(AA^T\right)^{-1}(Ax - b);$$
- l_2 ball $Q = B_{\|\cdot\|_2}[c, r]$,
$$\operatorname{proj}_Q(x) = c + \frac{r}{\max\{\|x - c\|_2, r\}}(x - c);$$

- half-space $Q = \{x : a^T x \le \alpha\}$,

$$\mathrm{proj}_Q(x) = x - \frac{[a^T x - \alpha]_+}{\|a\|^2} a.$$

Algorithm 2 Gradient method for simple set

Input: Initial point $x_0 \in Q$
for $k \leftarrow 0, \cdots, K-1$ **do**
  Step 1. Choose h_k.
  Step 2. Compute $x_{k+1} = \mathrm{proj}_Q\left(x_k - h_k \nabla f(x_k)\right)$.
end for
Output: x_K

Theorem 10. *Assume f is μ-strongly convex and L-smooth and choose $h_k = h \in (0, 2/(\mu+L)]$. Then, the Gradient Method generates a sequence $\{x_k\}$ such that*

$$\|x_k - x_*\| \le (1-\mu h)^k \|x_0 - x_*\|.$$

If $h = 2/(\mu+L)$, then

$$\|x_k - x_*\| \le \left(\frac{\kappa-1}{\kappa+1}\right)^k \|x_0 - x_*\|.$$

where $\kappa = L/\mu$.

Proof. Let $r_k := \|x_k - x_*\|$. Then, using Theorem 9 and step 2 of Algorithm 2, we get

$$\begin{aligned} r_{k+1}^2 &= \left\|\mathrm{proj}_Q(x_k - h\nabla f(x_k)) - \mathrm{proj}_Q(x_* - h\nabla f(x_*))\right\|^2 \\ &\le \|x_k - x_* - h[\nabla f(x_k) - \nabla f(x_*)]\|^2 \\ &= r_k^2 - 2h\left\langle \nabla f(x_k) - \nabla f(x_*), x_k - x_* \right\rangle + h^2 \|\nabla f(x_k) - \nabla f(x_*)\|^2 \end{aligned}$$

where the inequality is due to Lemma 5. Using Lemma 4, we have

$$r_{k+1}^2 \le \left(1 - \frac{2h\mu L}{\mu+L}\right) r_k^2 + h\left(h - \frac{2}{\mu+L}\right) \|\nabla f(x_k) - \nabla f(x_*)\|^2.$$

Using Lemma 3(iii), we have

$$\mu\|x-y\| \le \|\nabla f(x) - \nabla f(y)\|$$

and the assumption that $0 \le h \le 2/(\mu+L)$, we further have

$$r_{k+1}^2 \le \left(1 - \frac{2h\mu L}{\mu+L} + \mu^2 h\left(h - \frac{2}{\mu+L}\right)\right) r_k^2 = (1-\mu h)^2 r_k^2.$$

So the first conclusion holds by applying the above inequality recursively. The second conclusion holds by plugging in $h = 2/(\mu+L)$. $\square$

Note that the convergence rate here is the same as in the unconstrained one.