Lecture 13

Gradient Methods for Constrained Optimization

October 16, 2008

Outline

- Gradient Projection Algorithm
- Convergence Rate

Constrained Minimization

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject} \quad x \in X \end{aligned}$$

- Assumption 1:
  - The function f is convex and continuously differentiable over $\mathbb{R}^n$
  - The set X is closed and convex
  - The optimal value $f^* = \inf_{x \in \mathbb{R}^n} f(x)$ is finite

- Gradient projection algorithm

$$x_{k+1} = P_X[x_k - \alpha_k \nabla f(x_k)]$$

starting with $x_0 \in X$.

Bounded Gradients

Theorem 1 *Let Assumption 1 hold, and suppose that the gradients are uniformly bounded over the set X. Then, the projection gradient method generates the sequence $\{x_k\} \subset X$ such that*

- *When the constant stepsize $\alpha_k \equiv \alpha$ is used, we have*

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \frac{\alpha L^2}{2}$$

- *When diminishing stepsize is used with $\sum_k \alpha_k = +\infty$, we have*

$$\liminf_{k\to\infty} f(x_k) = f^*.$$

Proof: We use projection properties and the line of analysis similar to that of unconstrained method. HWK 6

Lipschitz Gradients

- **Lipschitz Gradient Lemma** For a differentiable convex function f with Lipschitz gradients, we have for all $x, y \in \mathbb{R}^n$,

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

where L is a Lipschitz constant.

- **Theorem 2** *Let Assumption 1 hold, and assume that the gradients of f are Lipschitz continuous over X. Suppose that the optimal solution set X^* is not empty. Then, for a constant stepsize $\alpha_k \equiv \alpha$ with $0 < \alpha < \frac{2}{L}$ converges to an optimal point, i.e.,*

$$\lim_{k\to\infty} \|x_k - x^*\| = 0 \qquad \text{for some } x^* \in X^*.$$

Proof:

Fact 1: If $z = P_X[z - v]$ for some $v \in \Re^n$, then $z = P_X[z - \tau v]$ for any $\tau > 0$.

Fact 2: $z \in X^*$ if and only if $z = P_X[z - \nabla f(z)]$.

These facts imply that $z \in X^*$ if and only if $z = P_X[z - \tau \nabla f(z)]$ for any $\tau > 0$.

By using the definition of the method and the preceding relation with $\tau = \alpha$, we obtain for any $z \in X^*$,

$$\|x_{k+1} - z\|^2 = \|P_X[x_k - \alpha \nabla f(x_k)] - P_X[z - \alpha \nabla f(z)\|^2.$$

By non-expansiveness of the projection, it follows

$$\begin{aligned} \|x_{k+1} - z\|^2 &= \|x_k - z - \alpha(\nabla f(x_k) - \nabla f(z))\|^2 \\ &= \|x_k - z\|^2 - 2\alpha(x_k - z)^T(\nabla f(x_k) - \nabla f(z)) \\ &\quad + \alpha^2 \|\nabla f(x_k) - \nabla f(z)\|^2 \end{aligned}$$

Using Lipschitz Gradient Lemma, we obtain for any $z \in X^*$,

$$\|x_{k+1} - z\|^2 \leq \|x_k - z\|^2 - \frac{\alpha}{L}(2 - \alpha L)\|\nabla f(x_k) - \nabla f(z)\|^2. \qquad (1)$$

Hence, for all k,

$$\frac{\alpha}{L}(2 - \alpha L)\|\nabla f(x_k) - \nabla f(z)\|^2 \leq \|x_k - z\|^2 - \|x_{k+1} - z\|^2.$$

By summing the preceding relations from arbitrary K to N, with $K < N$, we obtain

$$\frac{\alpha}{L}(2-\alpha L) \sum_{k=K}^{N} \|\nabla f(x_k)-\nabla f(z)\|^2 \leq \|x_K-z\|^2-\|x_{N+1}-z\|^2 \leq \|x_K-z\|^2.$$

In particular, setting $K = 0$ and letting $N \to \infty$, we see that

$$\frac{\alpha}{L}(2 - \alpha L) \sum_{k=0}^{\infty} \|\nabla f(x_k) - \nabla f(z)\|^2 \leq \|x_0 - z\|^2 < \infty. \tag{2}$$

As a consequence, we also have

$$\lim_{k \to \infty} \nabla f(x_k) = \nabla f(z). \tag{3}$$

By discarding the non-positive term in the right hand side of Eq. (1), we have for any $z \in X^*$ and all k,

$$\|x_{k+1} - z\|^2 \leq \|x_k - z\|^2 + (2 - \alpha L)\|\nabla f(x_k) - \nabla f(z)\|^2.$$

By summing these relations over $k = K, \ldots, N$ for arbitrary K and N with $K < N$, we obtain

$$\|x_{N+1} - z\|^2 \leq \|x_K - z\|^2 + (2 - \alpha L) \sum_{k=K}^{N} \|\nabla f(x_k) - \nabla f(z)\|^2.$$

Taking limsup as $N \to \infty$, we obtain

$$\limsup_{N\to\infty} \|x_{N+1} - z\|^2 \leq \|x_K - z\|^2 + (2 - \alpha L) \sum_{k=K}^{\infty} \|\nabla f(x_k) - \nabla f(z)\|^2.$$

Now, taking liminf as $K \to \infty$ yields

$$\begin{aligned}\limsup_{N\to\infty} \|x_{N+1} - z\|^2 \leq & \liminf_{K\to\infty} \|x_K - z\|^2 \\ & + (2 - \alpha L) \lim_{K\to\infty} \| \sum_{k=K}^{\infty} \nabla f(x_k) - \nabla f(z)\|^2 \\ = & \liminf_{K\to\infty} \|x_K - z\|^2,\end{aligned}$$

where the equality follows in view of the relation in (2). Thus, we have that the sequence $\{\|x_k - z\|\}$ is convergent for every $z \in X^*$.
By the inequality in Eq. (1), we have that

$$\|x_k - z\| \leq \|x_0 - z\| \qquad \text{for all } k.$$

Hence, the sequence $\{x_k\}$ is bounded, and it has an accumulation point. Since the scalar sequence $\{\|x_k - z\|\}$ is convergent for every $z \in X^*$, the sequence $\{x_k\}$ must be convergent.
Suppose now that $x_k \to \bar{x}$. By considering the definition of the iterate x_{k+1}, we have

$$x_{k+1} = P_X[x_k - \alpha \nabla f(x_k)].$$

Letting $k \to \infty$ and using $x_k \to \bar{x}$, and continuity of the gradient $\nabla f(x)$, we obtain

$$\bar{x} = P_X[\bar{x} - \alpha \nabla f(\bar{x})].$$

In view of facts 1 and 2, the preceding relation is equivalent to $\bar{x} \in X^*$. $\square$

Modes of Convexity: Strict and Strong

- **Def.** f is strictly convex if for all $x \neq y$ and $\alpha \in (0,1)$ we have

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha)f(y)$$

- **Def.** f is strongly convex if there exists a scalar $\nu > 0$ such that

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\nu}{2}\alpha(1-\alpha)\|x-y\|^2$$

for all $x, y \in \Re^n$ and any $\alpha \in [0,1]$.
The scalar ν is referred to as *strongly convex constant*.
The function is said to be *strongly convex with constant* ν.

Modes of Convexity: Differentiable Function

- Let $f : \Re^n \to \mathbb{R}$ be continuously differentiable.
- Modes of convexity can be equivalently characterized in terms of the **linearization properties** of the function $\nabla f : \Re^n \to \Re^n$.
- We have
  - f is convex if and only if

$$f(x) + \nabla f(x)^T (y - x) \leq f(y) \qquad \text{for all } x, y \in \Re^n$$

  - f is strictly convex if and only if

$$f(x) + \nabla f(x)^T (y - x) < f(y) \qquad \text{for all } x \neq y$$

  - f is strongly convex with constant ν if and only if

$$f(x) + \nabla f(x)^T (y - x) + \frac{\nu}{2} \|y - x\|^2 \leq f(y) \qquad \text{for all } x, y \in \Re^n$$

Modes of Convexity: Gradient Mapping

- Let $f : \Re^n \to \mathbb{R}$ be continuously differentiable.
- Modes of convexity can be equivalently characterized in terms of the **monotonicity properties** of the gradient mapping $\nabla f : \Re^n \to \Re^n$.
- We have
  - f is convex if and only if

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0 \qquad \text{for all } x, y \in \Re^n$$

  - f is strictly convex if and only if

$$(\nabla f(x) - \nabla f(y))^T (x - y) > 0 \qquad \text{for all } x \neq y$$

  - f is strongly convex with constant ν if and only if

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \nu \|x - y\|^2 \qquad \text{for all } x, y \in \Re^n$$

Modes of Convexity: Twice Differentiable Function

- Let $f : \Re^n \to \mathbb{R}$ be twice continuously differentiable.
- Modes of convexity can be equivalently characterized in terms of the **definiteness of the Hessians** $\nabla^2 f(x)$ for $x \in \Re^n$.
- We have
  - f is convex if and only if

$$\nabla^2 f(x) \geq 0 \qquad \text{for all } x \in \Re^n$$

  - f is strictly convex **if**

$$\nabla^2 f(x) > 0 \qquad \text{for all } x \in \Re^n$$

  - f is strongly convex with constant ν if and only if

$$\nabla^2 f(x) \geq \nu\, I \qquad \text{for all } x \in \Re^n$$

Strong Convexity: Implications

Let f be continuously differentiable and strongly convex* over $\mathbb{R}^n$ with constant m

- **Implications**:
  - **Lower Bound on f over $\mathbb{R}^n$**: for all $x, y \in \mathbb{R}^n$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|x - y\|_2^2 \tag{4}$$

    - minimize w/r to y in the right-hand side:

$$f(y) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|^2$$

    - minimum over $y \in \mathbb{R}^n$:

$$f(x) - f^* \leq \frac{1}{2m} \|\nabla f(x)\|^2$$

  - Useful as a stopping criterion (if you know m)

*Strong convexity over $\mathbb{R}^n$ can be replaced by a strong convexity over a set X. Then, all the relations stay valid over the set

- Relation (4) with $x = x_0$ and $f(y) \leq f(x_0)$ implies that the level set $L_f(f(x_0))$ **is bounded**
- Relation (4) also yields for an optimal x^* and any $x \in \mathbb{R}^n$,

$$\frac{m}{2}\|x - x^*\|^2 \leq f(x) - f(x^*)$$

- Last two bullets HWK6 assignment.

Convergence Rate: Once Differentiable

Theorem 3 *Let Assumption 1 hold, and assume that the gradients of f are Lipschitz continuous over X with constant $L > 0$. Suppose that the function is strongly convex with constant $m > 0$. Then:*

- *A solution x^* exists and it is unique.*

- *The iterates generated by the gradient projection method with $\alpha_k \equiv \alpha$ and $\alpha < \frac{2}{L}$ converge to x^* with geometric rate, i.e.,*

$$\|x_{k+1} - x^*\|^2 \leq q^k \, \|x_k - x^*\|^2 \qquad \text{for all } k$$

with $q \in (0, 1)$ depending on m and L.

Proof: HWK 6.

Convergence Rate: Twice Differentiable

Theorem 4 *Let Assumption 1 hold. Assume that the function is twice continuously differentiable and strongly convex with constant* $m > 0$. *Assume also that* $\nabla f^2(x) \leq L$ *for all* $x \in X$. *Then:*

- *A solution* x^* *exists and it is unique.*

- *The iterates generated by the gradient projection method with* $\alpha_k \equiv \alpha$ *and* $\alpha < \frac{2}{L}$ *converge to* x^* *with geometric rate, i.e.,*

$$\|x_{k+1} - x^*\| \leq q^k \ \|x_k - x^*\| \qquad \text{for all } k$$

with $q = \max\{|1 - \alpha m|, |1 - \alpha L\}$.

Proof: The q here is different from the one in the preceding theorem. Since $\nabla f^2(x) \leq L$ for all $x \in X$, it follows that the gradients are Lipschitz continuous over X with constant L. By the definition of the method and the non-expansive property of the projection, we have for $z = x^*$ and any k,

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|P_X[x_k - \alpha \nabla f(x_k)] - P_X[x^* - \nabla f(x^*)]\|^2 \\ &\leq \|x_k - x^* - \alpha(\nabla f(x_k) - \nabla f(x^*))\|^2. \end{aligned} \tag{5}$$

Mean Value Theorem for vector functions When $g : \mathbb{R}^n \to \mathbb{R}$ is differentiable on $[x, y]$, we have

$$g(y) = g(x) + \int_0^1 \nabla g(x + \tau(y - x))\, d\tau$$

Applying this Theorem with $g = \nabla f$, $y = x_k$ and $x = x^*$, we obtain

$$\nabla f(x_k) = \nabla f(x^*) + \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))\, d\tau$$

Hence,

$$\nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau. \tag{6}$$

By introducing $A_k(x - x^*) = \nabla f(x_k) - \nabla f(x^*)$ and using this in relation (5), we obtain

$$\|x_{k+1} - x^*\| \leq \|(I - \alpha A_k)(x_k - x^*)\| \leq \|I - \alpha A_k\|\, \|x_k - x^*\|$$

The matrix A_k is symmetric, and hence $\|I - A_k\|$ is equal to the max absolute eigenvalue of $I - A_k$, i.e.,

$$\|I - \alpha A_k\| = \max\{|1 - \alpha\lambda_{\max}(A_k)|, |1 - \alpha\lambda_{\min}(A_k)|\}.$$

In view of Eq. (6), we have $A_k = \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))d\tau$. By the strong convexity of f, we have $\nabla^2 f(x) \geq mI$ for all x, while by the given condition, we have $\nabla^2 f(x) \leq L\, I$. Therefore,

$$\lambda_{\max}(A_k) \leq L, \qquad \lambda_{\min}(A_k) \geq m,$$

implying that

$$\|I - \alpha A_k\| \leq \max\{|1 - \alpha m|, |1 - \alpha M|\}.$$

□

- The parameter q is minimized when $\alpha^* = \frac{2}{m+L}$ in which case

$$q^* = \frac{L-m}{L+m} \quad \Longleftrightarrow \quad q^* = \frac{cond(f)-1}{cond(f)+1},$$

with $cond(f) = \frac{L}{m}$.

Upper Bound on Hessian and f over the Level Set

For a twice differentiable *strongly convex* f:

- The level set $L_0 = \{x \mid f(x) \leq f(x_0)\}$ is bounded
- The maximum eigenvalue of the Hessian $\nabla^2 f(x)$ is a continuous function of x over L_0
- Hence, the maximum eigenvalue of the Hessian is bounded over L_0:

  there is a constant M such that $\nabla^2 f(x) \preceq MI$ for all $x \in L_0$

- **Upper Bound on f over L_0**:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2}\|y - x\|^2 \quad \text{for all } x, y \in L_0$$

  - minimize over $y \in L_0$ in both sides:

$$f^* \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|^2 \qquad \text{for all } x \in L_0$$

Condition Number of a Matrix

For a a twice differentiable strongly convex f: $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in L_0$

- The condition number $cond(A)$ of a positive definite matrix A:

$$cond(A) = \frac{\text{largest eigenvalue of } A}{\text{smallest eigenvalue of } A}$$

- The ratio $\frac{M}{m}$ is an upper bound on the condition number $\nabla^2 f(x)$ for every $x \in L_0$

Strong Convexity and Condition Number of Level Sets

Assume a minimizer x^* of f over $\mathbb{R}^n$ exists and f is strongly convex.
Consider the level set $L_0 = \{x \mid f(x) \leq f(x_0)\}$

- We have seen that $\quad mI \preceq \nabla^2 f(x) \preceq MI \quad$ for all $x \in L_0$
- Also, we have

$$f^* + \frac{m}{2}\|x - x^*\|^2 \leq f(x) \leq f^* + \frac{M}{2}\|x - x^*\|^2$$

- Hence: $\quad B_{inner} \subseteq L_0 \subseteq B_{outer}$, where

$$B_{inner} = \left\{x \mid \|x - x^*\| \leq \sqrt{\left(2\left(f(x_0) - f^*\right)/M\right)}\right\}$$

$$B_{outer} = \left\{x \mid \|x - x^*\| \leq \sqrt{\left(2\left(f(x_0) - f^*\right)/m\right)}\right\}$$

- Therefore, we have a bound on $cond(L_0)$

$$cond(L_0) \leq \frac{M}{m}$$

- The condition number of level sets **affects the efficiency of the algorithms**