Lecture 11

Unconstrained Optimization

Differentiable Case

October 9, 2008

Outline

- Terminology and Assumptions
- Gradient Methods

Unconstrained Minimization

minimize $f(x)$

Assumption 1:

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

When f has a domain $\mathrm{dom} f \neq \mathbb{R}^n$, we need continuously differentiable over $\mathrm{dom}\, f$ and the additional assumption that

- The level sets of f are closed.

The level set closedness condition:

- Not always easy to verify
- Satisfied when **all level sets are closed**; this is guaranteed when:
  - The epigraph $\mathrm{epi} f$ of f is closed [f closed]
  - f increases to $+\infty$ as the boundary of the domain is approached: $f(x) \to \infty$ as $x \to \bar{x}$ with $\bar{x} \in \mathrm{bd}\,(\mathrm{dom} f)$

An example of a differentiable convex function with closed level sets:

$$f(x) = -\sum_{i=1}^{m} \ln(b_i - a_i^T x) \qquad \mathrm{dom} f = \{x \mid Ax < b\}$$

Minimization Methods

- Iterative methods of the form

$$x_{k+1} = x_k + \alpha_k d_k \qquad \text{starting with some } x_0 \in \mathbb{R}^n$$

  - $\alpha_k > 0$ is a stepsize, d_k is a direction

- Goal: Generate a sequence of points $\{x_k\}$ such that

$$f(x_k) \to f^*$$

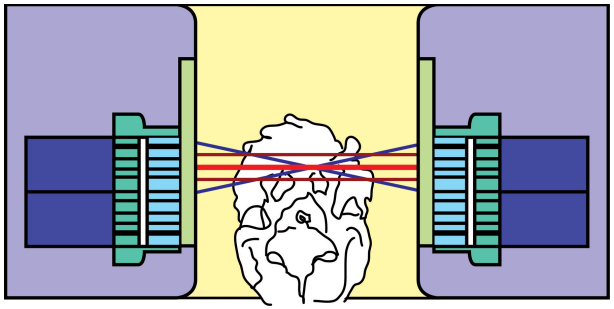
- Can be interpreted as iterative methods for solving the system of equations corresponding to the optimality condition

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

- Renewed interest in gradient methods due to their simplicity (in large distributed optimization)

Motivating Example

Image Reconstruction in PET-scan [Ben-Tal, 2005]



- Maximum Likelihood Model results in convex optimization

$$\min_{x \geq 0,\ e'x \leq 1} \left\{ -\sum_{i=1}^{m} y_i \ln \left(\sum_{j=1}^{n} p_{ij} x_j \right) \right\}$$

  - x is a decision vector
  - y models measured data (by PET detectors)
  - p_{ij} probabilities modeling detections of emitted positrons
- Constrained differentiable problem

Intuition/Motivation

- Suppose we are at some x_k. Consider the first-order Taylor expansion of the function f at x_k along a fixed nonzero direction $d \in \mathbb{R}^n$, as a function of the step $\alpha > 0$,

$$\begin{aligned} f(x_k + \alpha d) &= f(x_k) + \alpha \nabla f(x_k)^T d + o(\alpha) \\ &= f(x) + \alpha \left(\nabla f(x_k)^T d + \frac{o(\alpha)}{\alpha} \right) \end{aligned}$$

- Suppose that α is small so that $\frac{o(\alpha)}{\alpha}$ is negligible.
If $\nabla f(x_k)^T d < 0$, we have for a range of sufficiently small stepsizes α

$$f(x_k + \alpha d) \approx f(x) + \alpha \nabla f(x_k)^T d < f(x_k)$$

Gradient Descent Method

- Uses $d_k = -\nabla f(x_k)$
- Known as Gradient Descent (our focus)

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

- Stepsize Choices
  - Constant
  - Diminishing $\alpha_k \to 0$ with $\sum_k \alpha_k = +\infty$
  - Line search types
    - Exact line search: $\alpha_k = \operatorname{argmin}_{\alpha>0} f(x_k + \alpha d_k)$
    - Backtracking line search:

Additional Assumptions on f for Convergence

- Convergence analysis
  - Bounded Gradients: for some $L > 0$

$$\|\nabla f(x)\| \leq L \qquad \text{for all } x, y \in \mathbb{R}^n,$$

  - Lipschitz Gradients: for some scalar M

$$\|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^n,$$

- Convergence rate estimates
  - Strong convexity of f
  - f with a sharp set of minima

Basic Iterate Relation

- **Basic Lemma** Let Assumption 1 hold. Let $y \in \mathbb{R}^n$ be arbitrary but fixed. Then, for the gradient method with any stepsize rule, we have for all k,

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k(f(x_k) - f(y)) + \alpha_k^2 \|\nabla f(x_k)\|^2.$$

- *Proof:* By the definition of the method, it follows that for any k,

$$\begin{aligned} \|x_{k+1} - y\|^2 &= \|x_k - \alpha_k \nabla f(x_k) - y\|^2 \\ &= \|x_k - y\|^2 - 2\alpha_k \nabla f(x_k)^T (x_k - y) + \alpha_k^2 \|\nabla f(x_k)\|^2. \end{aligned}$$

By the convexity of f, we have for any k and any $y \in \mathbb{R}^n$,

$$f(x_k) - f(y) \leq \nabla f(x_k)^T (x_k - y).$$

The desired relation follows by combining the preceding two inequalities.

Bounded Gradients

- **Theorem** Let Assumption 1 hold, and suppose that the gradients are bounded. Let the stepsize be constant, $\alpha_k = \alpha$ for some $\alpha > 0$ and all k. Then, the gradient method generates the sequence $\{x_k\}$ such that

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

- *Proof:* To arrive at a contradiction, assume that the given relation does not hold, i.e., assume that

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

Then, for some sufficiently small $\epsilon > 0$, we have

$$f(x_k) \geq f^* + \frac{\alpha L^2}{2} + 2\epsilon \qquad \text{for all } k.$$

The function f is continuous over X, so that there exists $\widehat{y} \in \mathbb{R}^n$ such that $f(\widehat{y}) = f^* + \epsilon$, implying that

$$f(x_k) - f(\widehat{y}) \geq \frac{\alpha L^2}{2} + \epsilon \qquad \text{for all } k.$$

Using the relation of Lemma 1 with $\alpha_k = \alpha$ and $y = \widehat{y}$, we obtain for all k,

$$\begin{aligned} \|x_{k+1} - \widehat{y}\|^2 &\leq \|x_k - \widehat{y}\|^2 - 2\alpha(f(x_k) - f(\widehat{y})) + \alpha^2 \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - \widehat{y}\|^2 - 2\alpha \left(\frac{\alpha L^2}{2} + \epsilon \right) + \alpha^2 L^2, \end{aligned}$$

where we use the uniform boundedness of the gradients. Hence

$$\|x_{k+1} - \widehat{y}\|^2 \leq \|x_k - \widehat{y}\|^2 - 2\alpha\epsilon \qquad \text{for all } k,$$

and by summing the preceding inequalities over k, we obtain

$$\|x_k - \widehat{y}\|^2 \leq \|x_0 - \widehat{y}\|^2 - 2k\alpha\epsilon.$$

However, the preceding relation fails to hold for sufficiently large k. Therefore, we must have

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

Lipschitz Gradients

- **Q-approximation Lemma**
  For continuously differentiable function with Lipschitz gradients, we have

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

- **Theorem** Let Assumption 1 hold, and assume that the gradients of f are Lipschitz. Then, for a constant stepsize α with $\alpha < \frac{2}{M}$, we have

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

  If in addition, an optimal solution exists [i.e., the $min_x f(x)$ is attained at some x^*], then every accumulation point of the sequence $\{x_k\}$ is optimal.

- *Proof:* Using Q-approximation Lemma with $y = x_{k+1}$ and $x = x_k$, we have

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

By summing these relations and using $2 - \alpha M > 0$, we can see that

$$\sum_k \|\nabla f(x_k)\|^2 < \infty,$$

implying $\lim_{k \to \infty} \|f(x_k)\| = 0$.

Suppose a solution exists, and the sequence $\{x_k\}$ has an accumulation point $\bar{x}$. Then, by continuity of the gradient we have $\nabla f(x_k) \to \nabla f(\bar{x})$ along an appropriate subsequence.
Since $\nabla f(x_k) \to 0$, it follows $\nabla f(\bar{x}) = 0$, implying that $\bar{x}$ is a solution.

Strong Convexity Assumption and Implications

- **Strong convexity assumption**: f is twice continuously differentiable and there exists an $m > 0$ such that

$$\nabla^2 f(x) \succeq mI \qquad \text{for all } x \in \mathbb{R}^n$$

- Can be weakened by assuming the strong convexity only over the level set $L_f(\gamma)$ for $\gamma = f(x_0)$ when we have the guarantee that the iterates x_k stay in that level set.

- **Implications**:
  - **Lower Bound on f over $\mathbb{R}^n$**:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|x - y\|_2^2 \qquad \text{for all } x, y \in \mathbb{R}^n$$

    - 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 stopping criterion (if you know m)
  - Relation (1) with $x = x_0$ and $f(y) \leq f(x_0)$ implies that the level set $L_f(f(x_0))$ **is bounded**
  - Relation (1) also yields for an optimal x^* and any $x \in \mathbb{R}^n$,

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

Upper Bound on Hessian and f over the Level Set

For a *strongly convex* f:

- The level set $L_0 = \{x \mid f(x) \leq f(x_0)\}$ is bounded (just shown)
- 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 strongly convex f: $\quad mI \preceq \nabla^2 f(x) \preceq MI \quad$ 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 $mI \preceq \nabla^2 f(x) \preceq MI$ 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: $B_{inner} \subseteq L_0 \subseteq B_{outer}$, 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**