Compact course notes

Combinatorics and Optimization 367 Winter 2012

Non-linear optimization

Lecturer: N. Gillis
transcribed by: J. Lazovskis
University of Waterloo
April 10, 2012

Contents

0.1 Definitions . . . 2

1 Situations **2**
1.1 Uncontsrained problems . . . 2
1.2 Constrained problems . . . 3
1.3 Convexity . . . 3

2 Strategies **4**
2.1 Unconstrained nonlinear optimization . . . 4
2.2 Duality . . . 5
2.3 Perturbation and sensitivity analyisis . . . 6

0.1 Definitions

Definition 0.1.1. A matrix $A \in M_{n\times n}$ is termed positive semi-definite iff for all $x \in \mathbb{R}^n$, $x^T Ax \geqslant 0$. Moreover, A is positive definite if strict equality holds.

· Also note that such a matrix has nonnegative (or strictly positive) eigenvalues.

1 Situations

1.1 Uncontsrained problems

Theorem 1.1.1. Let $f \in C^2$ with x^* a local minimum of f. Then $\nabla f(x^*) = 0$.

Proof: Suppose that $\nabla f(x^*) \neq 0$.
Let $d = -\nabla f(x^*)$.
Then the directional derivative is $\nabla f(x^*)^T d = -\|\nabla f(x^*)\|^2 < 0$, so

$$\begin{aligned} f(x^* + td) &= f(x^*) + t\nabla f(x^*)^T d + O(t) \\ &< f(x^*) \end{aligned}$$

■

Theorem 1.1.2. Let $f \in C^2$ with x^* a local minimum of f. Then $\nabla^2 f(x^*) \geqslant 0$.

Proof: Since x^* is a local minimum, from above $\nabla f(x^*) = 0$.
Suppose that $\nabla^2 f(x^*) \ngeqslant 0$.
Then there exists a vector d such that $d^T \nabla^2 f(x^*)d < 0$, so

$$\begin{aligned} f(x^* + td) &= f(x^*) + t\nabla f(x^*)^T d + \tfrac{1}{2}t^2 d^T \nabla^2 f(x^*)d + O(t^2) \\ &= f(x^*) + \tfrac{1}{2}t^2 d^T \nabla^2 f(x^*)d \\ &< f(x^*) \end{aligned}$$

■

Theorem 1.1.3. Let $f \in C^2$ with x^* in the domain of f such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) > 0$. Then x^* is a local minimum.

Proof: Since $f \in C^2$, use Taylor to approximate f around x^*.
Then for any vector d that represents the change in x and small positive $t < 1$,

$$\begin{aligned} f(x^* + td) &= f(x^*) + t\nabla f(x^*)^T d + \tfrac{1}{2}t^2 d^T \nabla^2 f(x^*)d + O(t^2) \\ &= f(x^*) + \tfrac{1}{2}t^2 d^T \nabla^2 f(x^*)d + O(t^2) \\ &> f(x^*) \end{aligned}$$

■

Remark 1.1.4. If $\nabla^2 f(x^*)$ has one positive and one negative eigenvalue, then x^* is termed a saddle point. It is neither a maximum nor a minimum, but is still a stationary point of f.

Remark 1.1.5. A function can be bounded but still have no minimum. For example, $\arctan(x)$ is bounded on $x \in \mathbb{R}$ by $\pm\frac{\pi}{2}$, but it never reaches the infimum of $-\frac{\pi}{2}$.

1.2 Constrained problems

Theorem 1.2.1. [Extreme value theorem]
If f is continuous and bounded over a compact domain D, then f attains its infimum on D.

1.3 Convexity

Definition 1.3.1. An optimization problem is termed a convex optimization problem iff
· it is a minimization problem
· the objective function is convex
· the feasible domain is convex

Definition 1.3.2. A set X is termed convex iff for all $x, y \in X$ and $\lambda \in (0,1)$,

$$\lambda x + (1-\lambda)y \in X$$

Definition 1.3.3. A function $f : X \to Y$ is termed convex iff $\text{dom}(f)$ is convex and for all $x, y \in X$ and $\lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y)$$

Definition 1.3.4. A function $f : X \to Y$ is termed concave iff $\text{dom}(f)$ is convex and for all $x, y \in X$ and $\lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda)y) \geqslant \lambda f(x) + (1-\lambda)f(y)$$

Theorem 1.3.5. A function $f \in C^2$ over a domain D is convex iff D is convex and

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

Proof: ($\Rightarrow$) Suppose f is convex but D is not.

For $x, y \in D$ and $0 < t < 1$, we have $f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y)$.
Divide both sides by t to get

$$\begin{aligned} f(y) &\geqslant f(x) + \frac{f((1-t)x+ty) - f(x)}{t} \\ &= f(x) + \frac{f(x+t(y-x)) - f(x)}{t} \\ &= f(x) + \nabla f(x)(y-x) \quad \text{as } t \to 0 \end{aligned}$$

($\Leftarrow$) Now suppose that $f(y) \geqslant f(x) + \nabla f(x)(y-x)$.
Since D is convex, choose $x \neq y$ both in D and let $z = tx + (1-t)y$.
Applying the premise twice, we find that

$$\begin{aligned} f(x) &\geqslant f(z) + \nabla f(z)(x-z) \\ f(y) &\geqslant f(z) + \nabla f(z)(y-z) \end{aligned}$$

Multiply the first by t, the second by $(1-t)$, then add to get $tf(x) + (1-t)f(y) \geqslant f(z)$. ■

Theorem 1.3.6. A function $f \in C^2$ over a domain D is convex iff D is convex and $\nabla^2 f(x) \geqslant 0 \;\forall\; x \in D$.

Proof: ($\Rightarrow$) Using the Taylor expansion and the premise, we have that

$$f(y) = f(x) + \nabla f(x + t(y-x))^T(y-x) \geqslant f(x) + \nabla f(x)^T(y-x)$$

This implies that $\nabla f(x+t(y-x))(y-x) \geqslant \nabla f(x)(y-x)$.
If we assume that $y > x$, then since $0 < t < 1$,

$$\begin{aligned} \tfrac{\nabla(f(x+t(y-x))+\nabla f(x)}{(y-x)} &\geqslant 0 \\ \implies \lim_{x\to y}\left[\tfrac{\nabla(f(x+t(y-x))+\nabla f(x)}{t(y-x)}\right]t = \nabla^2 f(x)t &\geqslant 0 \\ \implies \nabla^2 f(x) &\geqslant 0 \end{aligned}$$

($\Leftarrow$) Since $\nabla^2 f(x) \geqslant 0$, every directional derivative $\nabla f_i(x)$ is a non-decreasing function.
Let $0 < t < 1$.
If $y \geqslant x$, then $\nabla f_i(x+t(y-x)) \geqslant \nabla f_i(x)$, and so $\nabla f_i(x+t(y-x))(y-x) \geqslant \nabla f_i(x)(y-x)$.
If $y \leqslant x$, then $\nabla f_i(x+t(y-x)) \leqslant \nabla f_i(x)$, and so $\nabla f_i(x+t(y-x))(y-x) \geqslant \nabla f_i(x)(y-x)$.
For all directions together, $\nabla f(x+t(y-x))(y-x) \geqslant \nabla f(x)(y-x)$ for all $x, y \in D$. ■

Theorem 1.3.7. [Karush, Kuhn, Tucker]

· In the convex case, if x^* satisfies the KKT conditions, then it is a global optimum.

Definition 1.3.8. A feasible point x satisfies LICQ, the linear independence constraint qualification, iff the gradients of the active constraints are linearly independent.

Theorem 1.3.9. A convex function is continuous on the interior of its domain.

2 Strategies

2.1 Unconstrained nonlinear optimization

Definition 2.1.1. The two main strategies for unconstrained nonlinear programming are line search methods and trust region methods.

In the *trust region* method, we approximate f by a quadratic in a ball around our current iterate.
· The second-order term in this approach is found with the SR1 or BFGS strategy

Proposition 2.1.2. Suppose that f is convex and continuously differentiable. Then $\nabla f(x)$ is the gradient of f at x iff for all $y \in D$,

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

Proof: Taylor's theorem. ■

Definition 2.1.3. For a function f over its domain D, define the subdifferetial of f at $x \in D$ to be

$$\partial f(x) := \{z \in \mathbb{R}^n \mid f(y) \geqslant f(x) + z^T(y-x)\ \forall\ y \in D\}$$

Proposition 2.1.4. If f is convex, then
1. $\partial f(x) \neq \emptyset$ for all $x \in D$
2. f differentiable at x implies $\partial f(x) = \{\nabla f(x)\}$
3. f not differentiable at x implies $\lambda(\partial f(x)) > 0$

Corollary 2.1.5. The function f has a minimum at x iff $0 \in \partial f(x)$.

Definition 2.1.6. For a function f with domain D, define the Fenchel conjugate of f as the convex function

$$f^*(y) := \sup_{x \in D} \{y^T x - f(x)\}$$

Theorem 2.1.7. [Subgradient method]
Let f be Lipschitz continuous with Lipschitz constant L and $x^* \in \arg \inf_{x \in \mathbb{R}^n} \{f(x)\}$ with $\|x_0 - x^*\|_2 \leqslant R$. If α_k is the implemented step length at iteration k,

$$\min_k \Big\{ |f(x_k) - f(x^*)| \Big\} \leqslant \frac{R^2 + L^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i}$$

2.2 Duality

Definition 2.2.1. Given a function $f(x)$ for optimization and its associated Lagrangian $\mathcal{L}(x, \lambda)$, define the dual function to be the extended-real valued concave function

$$g(\lambda) = \inf_x \{\mathcal{L}(x, \lambda)\} = \inf_x \left\{ f(x) - \sum_{i \in \mathcal{E}, \mathcal{I}} \lambda_i c_i(x) \right\}$$

The values of λ for which $g(\lambda) \in \mathbb{R}$ are termed dual feasible.

This is a general analysis of functions and problems associated with duality.

Primal

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & x \in \left\{ x \in \mathbb{R} \;\middle|\; \begin{smallmatrix} c_i(x) = 0, \; i \in \mathcal{E} \\ c_j \geqslant 0, \; j \in \mathcal{I} \end{smallmatrix} \right\} = D \end{aligned}$$

Lagrangian fit

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E}, \mathcal{I}} \lambda_i c_i(x), \;\; \lambda_i \geqslant 0 \;\text{ for }\; i \in \mathcal{I}$$

Lagrangial dual fit

$$\begin{aligned} & g(\lambda) = \inf_x \{\mathcal{L}(x, \lambda)\} \leqslant \inf_{x \in \mathcal{D}} \{f(x)\} \\ & \text{dom}(g) = \{\lambda \in \mathbb{R}^{|\mathcal{E}| + |\mathcal{I}|} \mid \lambda_i \geqslant 0 \;\text{ for }\; i \in \mathcal{I} \;\text{ and }\; g(\lambda) > -\infty\} \end{aligned}$$

Lagrangian dual

$$\begin{aligned} \max_\lambda \quad & g(\lambda) \\ \text{s.t.} \quad & \lambda_i \geqslant 0 \;\text{ for }\; i \in \mathcal{I} \end{aligned}$$

Proposition 2.2.2. The Lagrangian dual function g is concave.

Proof: This is just computation, from the definition of concave. ■

Proposition 2.2.3. For any Lagrange multiplier λ with $\lambda_i \geqslant 0$ for $i \in \mathcal{I}$, the Lagrange dual function is a lower bound on the optimal value of the primal. That is,

$$g(\lambda) = \inf_x \{\mathcal{L}(x, \lambda)\} \leqslant \inf_x \{f(x)\} = f^*$$

Proof: Note that $\lambda_i \geqslant 0$ and $c_i(x) \geqslant 0$ for all i. ■

2.3 Perturbation and sensitivity analyisis

Definition 2.3.1. Consider the perturbed problem

$$\begin{aligned} f^0(\varepsilon) \quad &= \quad \min_{x \in D^\varepsilon} \quad f(x) \\ &\quad\quad \text{s.t.} \quad D^\varepsilon = \{x \in \mathbb{R}^n \mid c_i(x) = \varepsilon_i \text{ for } i \in \mathcal{E} \text{ and } c_i \geqslant \varepsilon_i \text{ for } i \in \mathcal{I}\} \end{aligned}$$

If $\varepsilon_i < 0$ for $i \in \mathcal{I}$, the problem is relaxed.
If $\varepsilon_i > 0$ for $i \in \mathcal{I}$, the problem is tightened.

Proposition 2.3.2. Suppose that an optimization problem has strong duality. Then for an optimal dual solution λ^*,

$$f^0(\varepsilon) \geqslant f^0(0) + \lambda^{*T}\varepsilon$$

Proof: Suppose that λ^* is an optimal dual solution to the original problem $f(x) = f^0(0)$.
So then for any $x \in D^\varepsilon$,

$$\begin{aligned} f^0(0) &= g(\lambda^*) \\ &\leqslant f(x) - \sum_i \lambda_i^* c_i(x) \\ &\leqslant f(x) - \sum_i \lambda_i^* \varepsilon_i \end{aligned}$$

Rearranging and restating gives the desired result. ∎

Proposition 2.3.3. Suppose that an optimization problem has strong duality and f^0 is differentiable at 0. Then for an optimal dual solution λ^*,

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

Or equivalently,

$$f^0(\varepsilon) = f^0(0) + \lambda^{*T}\varepsilon + \mathcal{O}(\varepsilon^2)$$