Final Review

IE417

In the Beginning...

- In the beginning, *Weierstrass's theorem* said that a continuous function achieves a minimum on a compact set.
- Using this, we showed that for a convex set S and y not in the set, there is a unique point in S with minimum distance from y.
- This allowed us to show that we can separate a convex set S from any point not in the set.
- Finally, we arrived at *Farkas' Theorem* which is at the heart of all optimization theory.

Convex Functions

- Recall that if f:S$\rightarrow$ $\mathbf{R}^n$ is twice-differentiable, then f is convex if and only if the Hessian of f is positive semidefinite at each point in S.
- If f is convex and S is a convex set, the point $x^* \in S$ is an optimal solution to the problem $\min_{x \in S} f(x)$ if and only if f has a subgradient ξ such that $\xi^T(x - x^*) \geq 0 \;\; \forall x \in S$.
- Note that this is equivalent to there begin no improving, feasible directions.
- Hence, if S is open, then x^* is an optimal solution if and only if there is a zero subgradient of f at x^*.

Characterizing Improving Directions

Unconstrained Optimization

- Consider the unconstrained optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in X \end{array}$$

where X is an open set (typically $\mathbf{R}^n$).

- If f is differentiable at x^* and there exists a vector d such that $\nabla f(x^*)^T d < 0$, then d is an improving direction.
- If $\nabla f(x^*)^T d > 0 \ \forall d \in \mathbf{R}^n$, then there are no improving directions.

Optimality Conditions

Unconstrained Optimization

- If x^* is a local minimum and f is differentiable at x^*, then $\nabla f(x^*) = 0$ and $H(x^*)$ is positive semi-definite.
- If f is differentiable at x^*, $\nabla f(x^*) = 0$, and $H(x^*)$ is positive definite, then x^* is a local minimum.
- If f is convex and x^* is a local minimum, then x^* is a global minimum.
- If f is strictly convex and x^* is a local minimum, then x^* is the unique global minimum.
- If f is convex and differentiable on the open set X, then $x^* \in X$ is a global minimum if and only if $\nabla f(x^*) = 0$.

Constrained Optimization

- Now consider the constrained optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \ \forall\ i \in [1, m] \\ & h_i(x) = 0 \ \forall\ i \in [1, l] \\ & x \in \mathrm{X} \end{aligned}$$

where X is again an open set (typically $\mathbf{R}^n$).

Feasible and Improving Directions

Constrained Optimization

- Definition: Let S be a nonempty set in $\mathbf{R}^n$ and let $x^* \in \text{cl } S$. The *cone of feasible directions* of S at x^* is given by

$$D = \{d: d \neq 0 \text{ and } x^* + \lambda d \in S, \forall \lambda \in (0, \delta), \exists \delta > 0\}$$

- *Definition*: Let S be a nonempty set in $\mathbf{R}^n$ and let $x^* \in \text{cl } S$. Given a function $f{:}\mathbf{R}^n \to \mathbf{R}$, the *cone of improving directions* of f at x^* is given by

$$F = \{d: f(x^* + \lambda d) < f(x^*), \forall \lambda \in (0, \delta), \exists \delta > 0\}$$

Necessary Conditions

Constrained Optimization

- If x^* is a local minimum, then $F \cap D = \varnothing$.
- The converse is not true.
- Given a feasible $x^* \in X$, set $I = \{i: g_i(x^*) = 0\}$.
- Define $F_0 = \{d: \nabla f(x^*)^T d < 0\}$ and $F_0' = \{d: d \neq 0, \nabla f(x^*)^T d \leq 0\}$. Then $F_0 \subseteq F \subseteq F_0'$.
- Define $G_0 = \{d: \nabla g_i(x^*)^T d < 0 \ \forall i \in I\}$ and $G_0' = \{d: d \neq 0, \nabla g_i(x^*)^T d \leq 0 \ \forall i \in I\}$. Then $G_0 \subseteq D \subseteq G_0'$.

Fritz-John Conditions

Constrained Optimization

- If x^* is a local minimum, then $F_0 \cap G_0 = \emptyset$.
- $F_0 \cap G_0 = \emptyset$ if and only if there exists $\mu, \nu \in \mathbf{R}^m$ such that

$$\mu_0 \nabla f(x^*) + \sum \mu_i \nabla g_i(x^*) + \sum \nu_i \nabla h_i(x^*) = 0$$

$$\mu_i g_i(x^*) = 0 \quad \forall\, i \in [1, m]$$

$$\mu \geq 0$$

$$(\mu, \nu) \neq 0$$

- These are the FJ conditions.

KKT Conditions

Constrained Optimization

- Assuming that $\nabla g_i(x^*)$ and $\nabla h_i(x^*)$ are linearly independent, then $\mu_0 > 0$ and we obtain the KKT conditions:

$$\nabla f(x^*) + \sum \mu_i \nabla g_i(x^*) + \sum \nu_i \nabla h_i(x^*) = 0$$

$$\mu_i g_i(x^*) = 0 \quad \forall\, i \in [1, m]$$

$$\mu \geq 0$$

- x^* is a KKT point is and only if $F_0 \cap G_0' = \varnothing$.

The Restricted Lagrangian

- Recall the restricted Lagrangian at x^* with respect to dual multipliers $u^* \geq 0$ and v^*:

$$L(x) = f(x) + \Sigma_{i \in I}\, u_i^* g_i(x) + \Sigma v_i^* h_i(x)$$

- The KKT conditions are equivalent to $\nabla L(x^*) = 0$.
- Notice that this is an attempt to include the requirement for feasibility into the objective function.
- This converts constrained optimization into unconstrained.

Second-order Conditions

- Suppose x^* is a KKT point with restricted Lagrangean function L.
  - If $\nabla^2 L$ is positive semi-definite $\forall x \in S$, Then x^* is a global minimum.
  - If $\nabla^2 L$ is positive semi-definite in a neighborhood of x^*, then x^* is a local minimum.
  - If $\nabla^2 L(x^*)$ is positive definite, then x^* is a strict local minimum.
- From this, we can derive second-order necessary and sufficient conditions.

Convex Programs

- The KKT conditions are sufficient for *convex programs*:
  - f is convex
  - $g_1, \ldots, g_m$ is convex
  - $h_1, \ldots, h_l$ is linear
- The KKT conditions are necessary and sufficient for convex programs with all linear constraints.
- Recall that convex functions are exactly those for which the set of improving directions can be characterized.

A Word about Necessary and Sufficient Conditions

- If the KKT conditions are *sufficient* (as when we have convexity), then any KKT point will be optimal.
- However, just because there does not exist a KKT point does not mean there is no optimal solution.
- On the other hand, if the KKT conditions are *necessary* and there is a KKT point, this *does not* mean that the problem has an optimal solution. The problem can still be unbounded.
- Only in cases where the KKT conditions are *necessary and sufficient* can we simply enumerate all KKT points and draw a definite conclusion about optimality.

Other Constraint Qualifications

- There are other (less restrictive) conditions which imply the necessity of the KKT conditions (Chapter 5).
- For convex programs, the *Slater condition* implies the necessity of the KKT conditions.
  - $\nabla h_i(x^*)$ are linearly independent.
  - there exists $x' \in S$ such that $g_i(x') < 0 \; \forall \; i \in I$.

The Lagrangian Dual

- Let $\Phi(x, u, v) = f(x) + u^T g(x) + v^T h(x)$.
- We now formulate the following *dual problem* D:

$$\begin{aligned} \text{max} \quad & \Theta(u, v) \\ \text{s.t.} \quad & u \geq 0 \end{aligned}$$

where $\Theta(u, v) = \inf\{\Phi(x, u, v) : x \in X\}$.

- It is straightforward to show weak duality.

Interpreting Lagrangean duality

- Assume we have a well-behaved problem (no duality gap).
- Suppose we know the optimal dual multipliers. Then the optimal primal solution is given by

$$\min \mathrm{L}(x) = \Phi(x, u^*, v^*), \ \mathrm{x} \in \mathrm{X}$$

- Alternatively, if we know the optimal primal solution, then the optimal dual multipliers are given by

$$\max \Theta(u, v) = \Phi(x^*, u, v), \ u \geq 0$$

Lagrangian Saddle Points

- Intuitively, a saddle point of $\Phi(x, u, v)$ is a triple (x^*, u^*, v^*) that simultaneously satisfies solves $\max \Theta(u, v)$ and $\min L(x)$.
- Hence, a (feasible) saddle point solution will automatically be optimal for the primal and the dual.
- The following are equivalent:
  - the existence of a feasible saddle point solution (x^*, u^*, v^*),
  - the absence of a duality gap,
  - the primal-dual optimality of (x^*, u^*, v^*).

Properties of the Dual Function

- The dual function $\Theta(w)$ is concave.
- If it is differentiable at w^*, then $\nabla\Theta(w^*) = (\mathbf{g}(x^*), \mathbf{h}(x^*))$.
- Otherwise, the direction of steepest ascent is given by the subgradient of Θ at w^* with the smallest norm.
- To maximize the dual, we generally use subgradient optimization.

Algorithms

- An algorithm is defined by its *algorithmic map*.
- Given our current location, where do we go next?
- This is determined by a mapping A: X $\rightarrow 2^{|X|}$ which maps each point in the *domain* X to a set of possible "next iterates".
- In other words, if the current iterate is x_k, then
  $x_{k+1} \in A(x_k)$.
- After terminating the algorithm, the final iterate x^* will be called a *solution*.

Closed Maps

- An algorithmic map A is said to be *closed* at $x \in X$ if

  - $x_k \in X$ and $\{x_k\} \to x$
  - $y_k \in A(x_k)$ and $\{y_k\} \to y$

  implies that $y \in A(x)$.
- The map A is said to be closed on $Z \subseteq X$ if it is closed at each point in Z.
- Under mild conditions, algorithms with closed maps will converge.

Line Search Methods

- Line search is fundamental to all optimization algorithms.
- Analytic Methods
  - Solve the line search problem analytically.
  - Take the derivative with respect to λ and set it to zero.
- Iterative Methods
  - Methods using function evaluations (Golden Section)
  - Methods using derivatives (Newton's method)
  - Generally guaranteed to converge for pseudoconvex functions.
- We also distinguish between exact and inexact methods.

Algorithms for Unconstrained Optimization

- These algorithms are composed of two components
  - Choosing a search direction
  - Performing a line search
- Under mild conditions, the exact line search map is closed.
- There are two basic classes
  - Methods using function evaluations,
  - Methods using derivatives.

Derivative-free methods

- The basic idea is to search in a sequence of orthogonal directions, thereby insuring convergence.
- An acceleration step can be inserted after each sequence (Hooke and Jeeves).
- The directions can also be recomputed after each sequence (Rosenbrock).
- These methods are generally inferior to those using derivative information, but are easy to implement and do not require much memory.

Methods Using Derivative Info

- Recall that $-\nabla f(x^*)$ is the direction of steepest descent.
- All of these methods are based on moving in a modified steepest descent direction.
- The method of steepest descent has difficulty with problems for which the Hessian is ill-conditioned.
- Newton's method deflects the steepest descent direction to $-H(x^*)^{-1}\nabla f(x^*)$ correct for the ill-conditioning, but is not globally convergent.
- The problem occurs when the Hessian is not positive definite.

Levenberg-Marquardt and Trust-Region Methods

- L-M methods
  - Perturb the Hessian until it is positive definite.
  - Perform a line search in the resulting direction.
  - Dynamically adjust the amount of perturbation.
- Trust region methods
  - Use a quadratic approximation to the function within a defined trust region.
  - Solve the approximated problem.
  - Adjust the trust region.

Methods of Conjugate Directions

- If $H \in \mathbf{R}^{n\times n}$ is symmetric, the linearly independent vectors $d_1, \ldots, d_n$ are called *H-conjugate* if $d_i^T H d_j = 0$ for $i \neq j$.
- With conjugate directions, we can minimize a quadratic function by performing line searches.
- Quasi-Newton methods
  - Idea 1: Use a search direction $d_j = -D_j \nabla f(x)$ where D_j is symmetric positive definite and approximates H^{-1}.
  - Idea 2: Update D_j at each step so that d_{j+1} is a conjugate direction.

Conjugate Gradient Methods

- A simpler version of quasi-Newton requiring less computation and less memory.
- Idea: Let the next search direction depend on the last one, i.e., $d_{j+1} = -\nabla f(y_{j+1}) + \alpha_j d_j$
- However, we maintain the requirement that the directions be conjugate.
- This turns out to be similar to a "memoryless" quasi-Newton method.
- These methods are more appropriate for large problems.

Subgradient Methods

- Suppose f is convex/concave, but not differentiable.
- Instead of using the direction $-\nabla f(x)$, find a subgradient ξ and use $-\xi$ as the search direction.
- The direction is not necessarily a descent direction, but if the step size is chosen as follows, these methods do converge.
  - $\{\lambda_k\} \to 0$
  - $\Sigma \lambda_k = \infty$
- Most often used for solving the Lagrangian dual.

Methods for Constrained Optimization

- Two classes
  - Methods that implicitly enforce the constraints by converting to an equivalent unconstrained problem.
    - Interior Methods (barrier)
    - Exterior Methods (penalty)
  - Methods that explicitly enforce the constraints by only searching in feasible directions.

Penalty and Barrier Functions

- A *penalty function* α is $\alpha(x) = \Sigma\, \phi(g_i(x)) + \Sigma\, \psi(h_i(x))$, where
  - $\phi(y) = 0$ if $y \leq 0$, $\phi(y) > 0$ if $y > 0$
  - $\psi(y) = 0$ if $y = 0$, $\psi(y) > 0$ if $y \neq 0$
- A *barrier function* is $B(x) = \Sigma\, \phi(g_i(x))$
  - $\phi(y) \geq 0$ if $y < 0$
  - $\lim_{y \to 0+} \phi(y) = \infty$

Implementing Penalty-Barrier

- Initialization: Choose termination scalar $\varepsilon > 0$, an initial point x_1, an initial penalty parameter $\mu_1 > 0$, and a scalar $\beta \in (0,1)$. Set $k = 1$.
- Loop
  - Starting at x_k, minimize $f(x) + \alpha(x)/\mu_k + \mu_k B(x)$ subject to $x \in X$ to obtain x_{k+1}.
  - If $\alpha(x_{k+1})/\mu_k + \mu_k B(x_{k+1}) < \varepsilon$, then STOP. Otherwise, let $\mu_{k+1} = \beta\mu_k$, replace k by $k + 1$ and iterate.

Performance of Penalty-Barrirer

- For penalty methods, under some mild conditions, if there *exists* a solution x_μ the penalty problem for each μ and $\{x_\mu\}$ is contained in a compact set, then $\{x_\mu\} \rightarrow x^*$, and $\inf\{f(x): x \in X, g(x) \geq 0, h(x) = 0\} = \lim \{\Theta(\mu)\}$.
- For barrier methods, under some mild conditions on the feasible set and the location of the optimal solution, then $\{x_\mu\} \rightarrow x^*$, and

$$\inf\{f(x): x \in X, g(x) \geq 0\} = \lim \{\Theta(\mu)\}$$

Comments on Penalty-Barrier

- Note that these methods depend on the ability to solve the penalty-barrier problem.
- These methods are subject to computational difficulties with extremely small/large multipliers.
- This is the reason for the incremental algorithms that are presented in the text.
- Ill-conditioning can cause further problems.
- In well-behaved examples, we can recover the optimal KKT multipliers.

Augmented Lagrangian Methods

- Consider the penalty function $\psi(h_i(x)) = [h_i(x) - \theta_i]^2$.
- Assuming only equality constraints, the penalized objective function can then be written as

$$F(x, v) = f(x) + \Sigma\, v_i h_i(x) + \mu \Sigma\, [h_i(x)]^2$$

- Any KKT point satisfying second-order sufficiency conditions for being a local min will be a local min of this function for sufficiently large μ.

Methods of Feasible Direction

- General Method
  - Generate an improving, feasible direction by solving a direction-finding program or using projection.
  - Perform a line search in that direction.
- These methods are most closely tied to the KKT conditions.
- The direction-finding program is the "alternative" to the existence of a KKT point.

Summary

- Factors to consider when faced with solving an NLP
  - Unconstrained
    - Is the function to be minimized convex?
    - Is the Hessian ill-conditioned?
    - What is the dimension of the problem?
  - Constrained
    - Is the feasible region convex?
    - Are the Hessians of the constraints ill-conditioned?
    - Is there a relaxation that is "easy"?