Course: Learning and Optimization in Multiagent Decision-Making Systems

Notes 1: Single Agent Static Optimization

Instructor: Rasoul Etesami

May 29, 2025

1 Preliminaries: Necessary and Sufficient Optimality Conditions

We consider optimization problems of the form

$$\min_{x\in X} f(x),$$

where $x=(x_1,\dots,x_n)$ is a decision variable, $X\subseteq\mathbb{R}^n$ is a constrained set of variables (also called a feasible set), and $f:\mathbb{R}^n\to\mathbb{R}$ is an objective cost. We often assume that f is twice continuously differentiable. If $X=\mathbb{R}^n$, the problem is called an unconstrained problem. Otherwise, it is called a constrained problem.

- Linear program is a special case when $f(x)=c'x$ is a linear function and X is characterized by a set of linear constraints.
- Nonlinear program refers to the case where $f(x)$ is a nonlinear function and $X\subseteq\mathbb{R}^n$ is a continuous set of decision variables.

Definition 1. *x^* is called a local minimum if $\exists\epsilon>0$ such that $f(x^*)\le f(x),\forall x\in X$ with $\|x-x^*\|<\epsilon$. If this inequality holds strictly, x^* is called a strict local minimum.*

Definition 2. *x^* is called a global minimum if $f(x^*)\le f(x),\forall x\in X$. If this inequality holds strictly, x^* is called a strict global minimum.*

Remark 1. *Local and global maximum are defined similarly by replacing f with $-f$.*

Proposition 3 (Necessary Optimality Conditions). *Let x^* be a local minimum for the unconstrained problem $\min_{x\in\mathbb{R}^n} f(x)$, where f is continuously differentiable in an open set containing x^*. Then,*

$$\nabla f(x^*)=0, \qquad \textit{(first order necessary condition)}$$
$$\nabla^2 f(x^*)\ge 0. \qquad \textit{(second order necessary condition)}$$

Proof: Fix a direction $d \in \mathbb{R}^n$, and consider the scalar function $g(\alpha) = f(x^* + \alpha d)$. Then,

$$0 \leq \lim_{\alpha \to 0^+} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} = \frac{d}{d\alpha} g(0) = d' \nabla f(x^*),$$

where the first inequality holds because x^* is a local minimum. By repeating the same argument when d is replaced by $-d$, we get $d' \nabla f(x^*) \leq 0$. Therefore, $d' \nabla f(x^*) = 0, \forall d \in \mathbb{R}^n \;\Rightarrow\; \nabla f(x^*) = 0$.

Next assume f is twice continuously differentiable, and let d be any vector in $\mathbb{R}^n$. For all $\alpha \in \mathbb{R}$, the second order expansion yields

$$f(x^* + \alpha d) - f(x^*) = \alpha \nabla f(x^*)' d + \frac{\alpha^2}{2} d' \nabla^2 f(x^*) d + o(\alpha^2).$$

For sufficiently small $\alpha \to 0$, we have $\frac{o(\alpha^2)}{\alpha^2} \to 0$. Since $\nabla f(x^*) = 0$, we have

$$0 \leq \lim_{\alpha \to 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha^2} = \frac{1}{2} d' \nabla^2 f(x^*) d, \; \forall d \in \mathbb{R}^n \;\Rightarrow\; \nabla^2 f(x^*) \geq 0.$$

□

Proposition 4 (Sufficient Condition). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable over an open set S. If $x^* \in S$ satisfies $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) > 0$, then, x^* is a strict local minima of f. In particular, $\exists\, \epsilon, \gamma > 0 : f(x) > f(x^*) + \frac{\gamma}{2}\|x - x^*\|^2, \forall x : \|x - x^*\| < \epsilon$.*

Proof: Since $\nabla^2 f(x)$ is positive definite, its smallest eigenvalue denoted by λ must be positive, and in particular for any $d \in \mathbb{R}^n$, we have,

$$d' \nabla^2 f(x^*) d \geq d'(\lambda I) d = \lambda \|d\|^2.$$

Using this relation into a second order approximation and because $\nabla f(x^*) = 0$, we have $f(x^* + d) - f(x^*) = \nabla f(x^*) d + \frac{1}{2} d' \nabla^2 f(x^*) d + o(\|d\|^2) \geq \frac{\lambda}{2}\|d\|^2 + o(\|d\|^2) = (\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2})\|d\|^2$. Thus, if $\|d\|$ is sufficiently small (e.g. $\|d\| < \epsilon$), the term $(\frac{\lambda}{2} + \frac{O(\|d\|^2)}{\|d\|^2})$ is strictly bounded above zero (e.g., greater than $\frac{\gamma}{2}$), and we get:

$$f(x^* + d) \geq f(x^*) + \frac{\gamma}{2}\|d\|^2, \forall d : \|d\| < \epsilon.$$

□

Definition 5. *A subset $C \subseteq \mathbb{R}^n$ is called convex if $\alpha x + (1-\alpha) y \in C, \forall x, y \in C, \forall \alpha \in [0,1]$. A function $f : C \to \mathbb{R}$ is called convex if $f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y), \forall x, y \in C, \forall \alpha \in [0,1]$. If in addition f is continuously differentiable, one can show that f is convex if and only if*

$$f(y) \geq f(x) + \nabla f(x)'(y - x), \forall x, y \in C.$$

We say f is strictly convex if the above inequalities hold strictly. A continuously differentiable function f is called σ-strongly convex (for some $\sigma > 0$) if

$$f(y) \geq f(x) + \nabla f(x)'(y - x) + \frac{\sigma}{2}\|y - x\|^2, \forall x, y \in C.$$

Finally, f is (stictly/strongly) concave if $-f$ is (stictly/strongly) convex.

Proposition 6. *Let X be a convex set and $f : X \to \mathbb{R}$ be a convex function.*

(a) If f is continuously differentiable, then $\nabla f(x^*)'(x - x^*) \geq 0$, $\forall x \in X$ is necessary and sufficient for a vector $x^* \in X$ to be a global minimum of f over X.

(b) If X is an open *set and f is continuously differentiable over X, then $\nabla f(x^*) = 0$ is a necessary and sufficient condition for a vector $x^* \in X$ to be a global minimum of f over X.*

Proof: (a) By convexity $f(x) \geq f(x^*) + \nabla f(x^*)'(x - x^*), \forall x \in X \implies f(x) \geq f(x^*)$, $\forall x \in X$. Conversely, if x^* minimizes f over x, for any $x \in X$ we have,

$$f(x^*)'(x - x^*) = \lim_{\alpha \to 0^+} \frac{f(x^* + \alpha(x - x^*)) - f(x^*)}{\alpha} \geq 0 \tag{1}$$

(b) If $\nabla f(x^*) = 0$, the optimality of x^* follows as a special case of part (a). Conversely, if x^* minimizes f over X, and X is an open set. Using the necessary condition in the previous proposition, we have $\nabla f(x^*) = 0$.

□

Definition 7. *Let $X \subseteq \mathbb{R}^n$ be a closed convex set and $\|\cdot\|$ be the Euclidean norm. The projection of an arbitrary vector $x \in \mathbb{R}^n$ on X is defined by $[x]^+ = \arg\min_{y \in X} \|y - x\|$.*

Lemma 8. *The function $f : \mathbb{R}^n \to X$ defined by $f(x) = [x]^+$ is continuous and non-expansive, i.e., $\|[x]^+ - [y]^+\| \leq \|x - y\|, \forall x, y \in \mathbb{R}^n$.*

2 Gradient Methods

The goal is to develop an iterative scheme to solve $\min_{x \in \mathbb{R}^n} f(x)$, or obtain a stationary point, i.e., a point x^* at which $\nabla f(x^*) = 0$. Gradient method refers to a class of algorithms to achieve this task and are generally in the form of:

$$x^{k+1} = x^k + \alpha^k d^k, \quad k = 0, 1, 2, \ldots,$$

where $\alpha^k \geq 0$ is the stepsize, and d^k is the update direction and often is taken such that results in a descent in the objective value, i.e., $f(x^{k+1}) < f(x^k)$. If d^k is a descent direction, the gradient method is called a gradient descent method. One way to ensure d^k is a descent direction is as follows:

$$f(x^{k+1}) = f(x^k + \alpha^k d^k) = f(x^k) + \alpha^k \nabla f(x^k)' d^k + o(\alpha^k)$$

If d^k is chosen such that $\nabla f(x^k)' d^k < 0$ and α^k is sufficiently small, then $f(x^{k+1}) < f(x^k)$. There are a large variety of choices for d^k and α^k in gradient methods. Following are just a few:

2.1 Descent Direction

$d^k = -D^k \nabla f(x^k)$, where D^k is a positive definite matrix. Then,

$$\nabla f(x^k)' d^k = \nabla f(x^k)' D^k \nabla f(x^k) < 0,$$

which satisfies the descent direction criterion. Some special choices for D^k include:

- **Steepest Descent**: $D^k = I,\ \forall k - 0, 1, 2, ...$ Among all directions d^k that have unit norm $\|d^k\| = 1$, the choice $d^k = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$ attains the minimum value for $\nabla f(x^k)'d^k$. However, the steepest descent often suffers from slow convergence, especially if the gradient at an iterate is orthogonal to the direction that leads to a minimum.

- **Newton's Method**: $D^k = (\nabla^2 f(x^k))^{-1},\quad k = 0, 1, 2, ...$, provided $\nabla^2 f(x^k)$ is positive definite. Newton's method often converges fast in a neighborhood of a local minimum. However, it requires more computation per iteration.

2.2 Stepsize Rules

There are many rules for choosing the stepsize in gradient methods. Some popular ones include:

- **Minimization Rule**: Given direction d^k, choose α^k that the cost is minimized along d^k, i.e.,

$$\alpha^k = \arg\min_{\alpha \geq 0} f(x^k + \alpha d^k)$$

  This requires solving a scalar minimization problem that is often done by a line search, to approximate α^k. Limited minimization rule is an alternative way that limits the feasible range for α in the above minimization to some interval, say $[0, S]$, i.e.,

$$\alpha^k = \arg\min_{\alpha \in [0,S]} f(x^k + \alpha d^k)$$

  .

- **Constant Stepsize:** Here a fixed stepsize $s > 0$ is selected and $\alpha^k = s,\ \forall k = 1, 2, \ldots$ However, if s is chosen too large, divergence will occur, while if s is too small, the rate of convergence may be very slow.

- **Diminishing Stepsize:** Choose α^k such that $\lim_{k\to\infty} \alpha^k = 0$ and $\sum_{k=1}^{\infty} \alpha^k = \infty$. The difference between this stepsize rule and previous ones is that it does not guarantee descent at each iteration (although it becomes more likely as $\alpha^k \to 0$). The second condition ensures that the stepsize does not diminish so fast that iterates stay away from a stationary point.

3 Convergence of Gradient Methods

In order for the iterates $\{x^k\}$ to converge to a stationary point, we need to avoid the situation that the direction d^k become asymptotically orthogonal to the gradient direction, i.e., as x approaches a nonstationary point, we should not have $\nabla f(x^k)'d^k \to 0$. Otherwise, there is a chance that $\{x^k\}$ will get stuck near a non-stationary point. That motivates the following definition:

Definition 9. *Let $\{x^k, d^k\}$ be a sequence generated by a given gradient method. We say the directional sequence $\{d^k\}$ is gradient related to $\{x^k\}$ if for any subsequence $\{x^k\}_{k\in K}$ that converges to a nonstationary point, the corresponding subsequence $\{d^k\}_{k\in K}$ is bounded and satisfies* $\limsup_{k\to\infty, k\in K} \nabla f(x^k)'d^k < 0$.

Theorem 10. *Let $\{x_k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, and assume that $\{d^k\}$ is gradient related and α^k is chosen by the minimization rule or the limited minimization rule. Then every limit point of $\{x^k\}$ is a stationary point for* $\min_{x\in\mathbb{R}^n} f(x)$.

3.1 Convergence Results for L-Smooth Functions

Definition 11. *A function $f(x)$ is called L-smooth (or has L-Lipschitz gradient) if*

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

Lemma 12. *[Descent Lemma] Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and L-smooth. Then*

$$f(y) \le f(x) + \nabla f(x)'(y-x) + \frac{L}{2}\|y-x\|^2 \quad \forall x, y.$$

Proof: Let $t \in [0,1]$ be a scalar, and define $g(t) = f(x+tz)$. Then, using the chain rule we have

$$\frac{d}{dt}g(t) = z'\nabla f(x+tz).$$

We can write

$$\begin{aligned} f(x+z) - f(x) = g(1) - g(0) = \int_0^1 \frac{d}{dt}g(t)dt &= \int_0^1 z'\nabla f(x+tz)dt \\ &\le |\int_0^1 z'\left(\nabla f(x+tz) - \nabla f(x)\right)dt| + \int_0^1 z'\nabla f(x)dt \\ &\le \int_0^1 z'\nabla f(x)dt + \|z\|\int_0^1 \|\nabla f(x+tz) - \nabla f(x)\|dt \\ &\le z'\nabla f(x) + \|z\|\int_0^1 Lt\|z\|dt \\ &= z'\nabla f(x) + \frac{L}{2}\|z\|^2. \end{aligned}$$

The result follows by choosing $z = y - x$. □

Theorem 13. *Let $\{x_k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related. Assume $f(x)$ is L-smooth, $d^k \neq 0, \forall k$, and $\epsilon \le \alpha^k \le (2-\epsilon)\bar{\alpha}^k$, where $\bar{\alpha}^k = \frac{|\nabla f(x^k)'d^k|}{L\|d^k\|^2}$ and $\epsilon \in (0,1)$ is a fixed scalar. Then every limit point of $\{x^k\}$ is a stationary point of* $\min_{x\in\mathbb{R}^n} f(x)$.

Proof: Using the descent lemma 12, we have:

$$\begin{aligned} f(x^k) - f(x^k + \alpha^k d^k) &\ge -\alpha^k f(x^k)'d^k - \frac{L}{2}(\alpha^k)^2\|d^k\|^2 \\ &= \alpha^k(|\nabla f(x^k)'d^k| - \frac{L}{2}\alpha^k\|d^K\|^2) \\ &\ge \alpha^k\frac{\epsilon}{2}|\nabla f(x^k)'d^k| \\ &\ge \frac{\epsilon^2}{2}|\nabla f(x^k)'d^k| \end{aligned}$$

Now, if a subsequence of $\{x_k\}_{k\in K}$ converges to a nonstationary point $\bar{x}$, we have $f(x^k) - f(x^{k+1}) \to 0$, which implies $|\nabla f(x^k)'d^k| \to 0$. This contradicts the assumption that $\{d_k\}$ is gradient related. Hence every limit point of $\{x_k\}$ is stationary. □

Remark 2. *In the case of steepest descent $d_k = -\nabla f(x^k)$, the condition on the stepsize becomes $\epsilon \le \alpha^k \le \frac{2-\epsilon}{L}$. Thus, a constant stepsize in the middle of the interval $(0, \frac{2}{L})$ guarantees convergence.*

4 Optimization Over a Convex Set

In this part, we consider constrained optimization problems of the form

$$\min_{x\in X} f(x),$$

where $X \subseteq \mathbb{R}^n$ is a closed convex set and f is continuously differentiable over an open set containing X.

Proposition 14 (Necessary Optimality Condition). *If x^* is a local minimum of f over X, then*

$$\nabla f(x^*)'(x-x^*) \geq 0, \quad \forall x \in X.$$

If f is convex over X, then the above condition is also sufficient for x^* to minimize f over X.

Proof: Fix $x \in X$ and define $g(\alpha) = f(x^* + \alpha(x-x^*))$ over $\alpha \in [0,1]$. Note that by convexity of X, $x^* + \alpha(x-x^*) \in X$, so $g(\cdot)$ is well-defined. By using the chain rule, we have

$$0 \leq \lim_{\alpha\to 0^+} \frac{f(x^*+\alpha(x-x^*))-f(x^*)}{\alpha} = \frac{dg(0)}{d\alpha} = \nabla f(x^*)'(x-x^*).$$

If f is convex, then $f(x) \geq f(x^*) + \nabla f(x^*)'(x-x^*)\, \forall x \in X \Rightarrow f(x) \geq f(x^*), \forall x \in X$. □

Definition 15. *x^* is called a stationary point for* $\min_{x\in X} f(x)$ *if* $\nabla f(x^*)'(x-x^*) \geq 0$, $\forall x \in X$.

4.1 Feasible Direction Methods

Given a feasible vector $x \in X$, a feasible direction at x is a vector d such that $x + \alpha d \in X$ for all sufficiently small $\alpha > 0$. A feasible direction method starts with a feasible vector x^0 and generates a sequence of feasible vectors $\{x^k\}$ according to $x^{k+1} = x^k + \alpha^k d^k$, where if x^k is not a stationary point for $\min_{x\in X} f(x)$, d^k is a feasible direction at x^k, which is also a descent direction, i.e., $\nabla f(x^k)'d^k < 0$, and $\alpha^k > 0$ is chosen such that $x^{k+1} = x^k + \alpha^k d^k \in X$. If x^k is a stationary point, the method stops, i.e., $x^{k+1} = x^k$.

Remark 3. *When X is a convex set, every feasible direction at x^k is of the form $d^k = \bar{x}^k - x^k$, where $\bar{x}^k$ is some feasible vector in X. Thus, for a nonstationary point x^k, the feasible direction method can be written as $x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k) \in X$, where $\alpha^k \in (0,1]$, $\bar{x}^k \in X$, and $\nabla f(x^k)'(\bar{x}^k - x^k) < 0$.*

As before, there could be many ways for choosing the stepsize, such as:

- Limited minimization rule: α^k is chosen such that $f(x^k + \alpha^k d^k) = \min_{\alpha\in[0,1]} f(x^k + \alpha d^k)$.
- Constant Stepsize: We set $\alpha^k = \alpha, \forall k$.

Proposition 16. *Consider* $\min_{x\in X} f(x)$*, where X is a closed convex set. Let $\{x^k\}$ be a sequence generated by the feasible direction method $x^{k+1} = x^k + \alpha^k d^k$. Assume that $\{d^k\}$ is gradient related*[1] *and that α^k is chosen by the minimization rule or the limited minimization rule. Then every limit point of $\{x^k\}$ is a stationary point.*

Remark 4. *To apply a feasible direction method, it is necessary to have an initial feasible point. Finding such a point could be difficult if X is characterized by a set of nonlinear inequalities. However, if set X is a polyhedron, one can find a feasible point by solving a linear program.*

[1] Recall that $\{d^k\}$ is gradient related to $\{x^k\}$ if for any subsequence $\{x^k\}_{k\in K}$ that converges to a nonstationary point, the corresponding subsequence $\{d^k\}_{k\in K}$ is bounded and satisfies $\limsup_{k\in K, k\to\infty} \nabla f(x^k)'d^k < 0$.

4.2 The Frank-Wolfe Method (a.k.a. Conditional Gradient Method)

A natural way to find a feasible direction $d^k = \bar{x}^k - x^k$ that satisfies the descent condition $\nabla f(x^k)'(\bar{x}^k - x^k) < 0$ is to solve the optimization problem

$$\min\{\nabla f(x^k)'(x - x^k) : x \in X\},$$

and use its optimal solution as $\bar{x}^k$ in the feasible direction method. The corresponding feasible direction method is called *Frank-Wolfe* method, also know as *conditional gradient method*.

Proposition 17. *Every limit point of the conditional gradient method with the minimization rule or limited minimization rule over the convex compact set X is stationary.*

Proof: Using Proposition 16, we only need to show that the direction sequence of the conditional gradient method is gradient related. Indeed, suppose that $\{x^k\}_{k\in K}$ converges to a nonstationary point $\tilde{x}$. Since $\bar{x}^k,\ x^k \in X$, and X is compact, the direction $d^k = \bar{x}^k - x^k$ is bounded. Moreover, by the definition of $\bar{x}^k$, we can write $\nabla f(x^k)'(\bar{x}^k - x^k) \leq \nabla f(x^k)'(x - x^k)\ \forall x \in X$. Therefore,

$$\limsup_{k\to\infty,\ k\in K} f(x^k)'(\bar{x}^k - x^k) \leq \nabla f(\tilde{x})'(x - \tilde{x}) < 0,$$

where the last inequality is obtained by taking minimum over $x \in X$ and noting that $\tilde{x}$ is a nonstationary point. □

4.2.1 Application of the Frank-Wolfe Method for Multicommodity Flows

We are given a directed graph with a set of directed arcs A and a set W of ordered origin-destination pairs $w = (i, j)$. For each w, we are given a scalar input traffic r_w. The goal is to divide each r_w among the many paths from origin to destination in a way that the total arc flow minimizes a suitable cost function.

Let P_w be a given set of paths that start at the origin and end at the destination of w, and x_p be the portion of r_w assigned to path p, also called the path flow on p. Let f_a denote the total flow on arc $a \in A$, which is the sum of path flows traversing arc a, i.e., $f_a = \sum_{p:a\in p} x_p$ and $D_a(f_a)$ is a convex cost function of the flow on arc $a \in A$. The total flow cost is given by $\sum_{a\in A} D_a(f_a)$, where $D_a(\cdot)$ is a convex function associated to the arc a. Therefore, the optimal routing problem can be formulated as:

$$\min_{f\in F} G(f) := \sum_{a\in A} D_a(f_a),$$

where the constraint set F is given by

$$F = \Big\{ f_a = \sum_{p:a\in p} x_p, \quad \sum_{p\in p_w} x_p = r_w\ \forall w \in W, \quad x_p \geq 0\ \forall p \in P_w,\ \ w \in W \Big\}.$$

Note that the constraint set F is a polyhedron. Given the current flow vector f^k, the Frank-Wolfe method finds $\bar{f}^k$ such that $\bar{f}^k = \arg\min_{f\in F} \nabla G(f^k)'(f - f^k)$, and updates $f^{k+1} = f^k + \alpha^k(\bar{f}^k - f^k)$, where α^k is obtained by the line minimization as

$$\alpha^k \in \arg\min_{\alpha\in[0,1]} G\big(f^k + \alpha(\bar{f}^k - f^k)\big).$$

Note that solving for $\bar{f}^k$ can be done using a shortest path problem, while the length of each arc a is $\frac{\partial G(f^k)}{\partial f_a}$ (due to linearity of the objective cost $\nabla f(x^k)'(f - f^k)$, the optimal solution for $\bar{f}^k$ is obtained by solving multiple shortest path problems, one for each w, to get the shortest paths $\bar{x}_{p_w}$, and set $\bar{f}_a^k = \sum_{p_w:a\in p_w} \bar{x}_{p_w}$). Thus, the subproblems for finding a feasible direction $\bar{f}^k - f^k$ can be solved efficiently using $|W|$ shortest path problems.

4.3 Gradient Projection Methods:

The conditional gradient method uses a feasible direction obtained by solving a subproblem with linear cost. Gradient projection methods use instead a subproblem with quadratic cost. While this subproblem may be more difficult to solve, the resulting convergence rate is typically better.

$$\bar{x}^k = \left[x^k - s^k \nabla f(x^k)\right]^+ = \arg\min_{x\in X}\left\{\nabla f(x^{k\prime})(x - x^k) + \frac{1}{2s^k}\|x - x^k\|^2\right\}$$
$$x^{k+1} = x^k + \alpha^k\left(\bar{x}^k - x^k\right). \tag{2}$$

i) Take a step $-s^k \nabla f(x^k)$ along the negative gradient,

ii) Project the result $x^k - s^k \nabla f(x^k)$ on X to obtain a feasible vector $\bar{x}^k$,

iii) Take a step along the feasible direction $\bar{x}^k - x^k$ using the stepsize α^k.

Remark 5. *For $s > 0$, we have $x^* = [x^* + s\nabla f(x^*)]^+$ if and only if x^* is a stationary point of $\min_{x\in X} f(x)$. The reason is that using the projection property $(x^* - s\nabla f(x^*) - x^*)'(x - x^*) \le 0 \quad \forall x \in X$, if and only if x^* is the projection of $x^* - s\nabla f(x^*)$ on X. That means $\nabla f(x^*)'(x - x^*) \ge 0$, $\forall x \in X$, which implies that x^* is a stationary point. Thus, the gradient projection method stops if and only if it encounters a stationary point.*

Proposition 18. *Let $\{x^k\}$ be a sequence generated by the gradient projection method (2) with choices of α^k by the minimization rule or the limited minimization rule, and $s^k = s > 0\ \forall k$. Then, every limit point of $\{x^k\}$ is stationary.*

Proof: Using Proposition 16, we only need to show that the direction sequence $\{\bar{x}^k - x^k\}$ is gradient related. Suppose $\{x^k\}$ converges to a nonstationary point $\tilde{x}$. Then, by continuity of the projection, we have $\limsup_{k\to\infty, k\in K} \|\bar{x}^k - x^k\| = \|[\tilde{x} - s\nabla f(\tilde{x})]^+ - \tilde{x}\| < \infty$.

Next, by the characteristic property of the projection, we have

$$(x^k - s^k\nabla f(x^k) - \bar{x}^k)'(x - \bar{x}^k) \le 0, \quad \forall x \in X.$$

Choosing $x = x^k$ in the above relation, we get $\nabla f(x^k)'(\bar{x}^k - x^k) \le -\frac{1}{s}\|x^k - \bar{x}^k\|^2$. Finally, by taking limit from the last expression, and because $s > 0$, we get:

$$\limsup_{k\to\infty,\ k\in K} \nabla f(x^k)'(\bar{x}^k - x^k) \le -\frac{1}{s}\|[\tilde{x} - s\nabla f(\tilde{x})]^+ - \tilde{x}\|^2 < 0,$$

where the last inequality follows from Remark 5 and the fact that $\tilde{x}$ is a nonstationary point. □

4.4 Proximal Algorithms

The proximal algorithm for minimizing a convex function f over a closed convex set X is given by

$$x^{k+1} = \arg\min_{x\in X}\left\{f(x) + \frac{1}{2c^k}\|x - x^k\|^2\right\},$$

where x^0 is an arbitrary starting point and c^k is a positive scalar parameter. Note that compared to the gradient projection method that only a linear approximation $\nabla f(x^k)'(x - x^k)$ is used in the optimization, here the entire function $f(x)$ is used in the proximal iteration. This makes the algorithm applicable even to nondifferentiable convex cost functions f.

Proposition 19. *For the proximal algorithm, the following properties hold:*

i) $f(x^{k+1}) \le f(x^k) - \frac{1}{2c^k}\|x^{k+1} - x^k\|^2$.

ii) $f(y) \ge f(x^{k+1}) + \frac{1}{c^k}(x^k - x^{k+1})'(y - x^{k+1}),\ \forall y \in X$.

iii) $\|x^{k+1} - y\|^2 \le \|x^k - y\|^2 - 2c^k(f(x^{k+1}) - f(y)),\ \forall y \in X$.

Proof: (i) Using the proximal iteration $x^{k+1} = \arg\min_{x\in X}\left\{f(x) + \frac{1}{2c^k}\|x - x^k\|^2\right\}$, the cost is reduced at each iteration, because by selecting $x = x^k$ in the above optimization, we obtain

$$f(x^{k+1}) \le f(x^k) - \frac{1}{2c^k}\|x^{k+1} - x^k\|^2.$$

(ii) Consider the following two convex sets: $C_1 = \{(x,w):\ f(x) < w,\ x \in X\}$ and $C_2 = \left\{(x,w):\ w \le \gamma^k - \frac{1}{2C^k}\|x - x^k\|^2,\ x \in \mathbb{R}^n\right\}$, where $\gamma^k = f(x^{k+1}) + \frac{1}{2c^k}\|x^{k+1} - x^k\|$. From the definition of x^{k+1} via the proximal iteration, these two sets are convex and disjoint, and their closures touch each other at the unique single point $(x^{k+1}, f(x^{k+1}))$. So there exists a separating hyperplane H passing through $(x^{k+1}, f(x^{k+1}))$, where we have the normal vector of H is given by the gradient of $\gamma^k - \frac{1}{2c^k}\|x - x^k\|^2$ at $x = x^{k+1}$, which is $\left(\frac{x^k - x^{k+1}}{c^k}, 1\right)$. Since C_1 must lie on the opposite side of H, we have $(y, f(y))'\left(\frac{x^k - x^{k+1}}{c^k}, 1\right) \ge (x^{k+1}, f(x^{k+1}))'\left(\frac{x^k - x^{k+1}}{c^k}, 1\right)$, for every $y \in X$, which after simplification gives us the desired inequality in (ii).
(iii) This part can be shown by expanding

$$\begin{aligned}\|x^k - y\|^2 = \|x^k - x^{k+1} + x^{k+1} - y\|^2 &= \|x^k - x^{k+1}\|^2 + \|x^{k+1} - y\|^2 - 2(x^k - x^{k+1})'(y - x^{k+1}) \\ &\ge \|x^{k+1} - y\|^2 - 2(x^k - x^{k+1})'(y - x^{k+1}),\end{aligned}$$

and using the result of part (ii) to bound the last crossover term. □

Theorem 20. *Let $f^* = \inf_{x\in X} f(x)$ (which may be $-\infty$) and X^* be the set of minima of f (which may be empty). Moreover, let $\{x^k\}$ be a sequence generated by the proximal algorithm. Then, if $\sum_{k=0}^{\infty} c^k = \infty$, we have $f(x^k) \to f^*$, and if $X^* \ne \emptyset$, $\{x^k\}$ converges to some point in X^*.*

Proof: Using Proposition 19 (part i), $\{f(x^k)\}$ is monotonically nonincreasing. Thus $f(x^k) \to f^\infty \ge f^*$. Moreover, by Proposition 19 (part iii), we have

$$\|x^{k+1} - y\|^2 \le \|x^k - y\|^2 - 2c^k(f(x^{k+1}) - f(y)) \quad \forall k, \forall y \in X. \tag{3}$$

By summing this inequality over $k = 0, 1, \ldots, N$, we get

$$2\sum_{k=0}^{N} c^k(f(x^{k+1}) - f(y)) \le \|x^0 - y\|^2 - \|x^{N+1} - y\|^2,\ \forall y \in X,\ N \ge 0.$$

Taking the limit as $N \to \infty$, we have $\sum_{k=0}^{\infty} c^k(f(x^{k+1}) - f(y)) \le \frac{1}{2}\|x^0 - y\|^2\ \forall y \in X,\ N \ge 0$. To derive a contradiction, assume $f^\infty > f^*$ and let $\hat{y} \in X$ be such that $f^\infty > f(\hat{y}) > f^*$. Since $\{f(x^k)\}$ is monotonically nonincreasing, $f(x^{k+1}) - f(\hat{y}) \ge f^\infty - f(\hat{y}) > 0$, and we can write

$$\infty = \left(f^\infty - f(\hat{y})\right)\sum_{k=0}^{\infty} c^k \le \sum_{k=0}^{\infty} c^k(f(x^{k+1}) - f(y)) \le \frac{1}{2}\|x^0 - y\|^2.$$

This contradiction shows that $f^{\infty} = f^*$.

Consider now the case where $X^* \neq \emptyset$ and let x^* be any point in X^*. Using (3) with $y = x^*$,

$$\|x^{k+1} - x^*\| \leq \|x^k - x^*\|^2 - 2c^k(f(x^{k+1}) - f(x^*)),$$

which shows that $\{\|x^k - x^*\|^2\}$ is monotonically nonincreasing and so $\{x^k\}$ is bounded. If $\bar{x}$ is a limit point of $\{x^k\}$, we have $\bar{x} \in X$ (as X is closed) and $f(\bar{x}) = \lim_{k\to\infty,\ k\in K} f(x^k) = f^{\infty} = f^*$ for any subsequence $\{x^k_{k\in K}\} \to \bar{x}$. Hence $\bar{x}$ minimizes f over X and must belong to X^*. Finally, since the distance of $\{x^k\}$ to any $x^* \in X^*$ is monotonically nonincreasing, $\{x^k\}$ must converge to a unique point $\bar{x} \in X^*$. □

The role of convergence of the proximal method depends on the choice of c^k, (larger c^k results in wider parabola and typically faster convergence) as well as on the order of growth of f near the optimal solution set (sharper minima can be achieved faster).

Theorem 21. *Assume $X^* \neq \emptyset$ and for some scalars $\beta, \delta, \gamma > 0$, we have $f^* + \beta d^{\gamma}(x) \leq f(x)$ for all x such that $d(x) \leq \delta$, where $d(x) = inf_{x^* \in X^*} \|x - x^*\|$. Let $\sum_{k=0}^{\infty} c^k = \infty$, so the proximal sequence converges to some point in X^*. Then, for all k sufficiently large and $x^{k+1} \notin X^*$, we have $d(x^{k+1}) + \beta c^k d(x^{k+1})^{\gamma-1} \leq d(x^k)$.*

5 General Constrained Optimization

5.1 Necessary and Sufficient Optimality Conditions for Equality Constraints

In this section, we consider the general equality constrained optimization of the form

$$\min\{f(x) : h_i(x) = 0,\ i = 1, \ldots, m\},$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $h_i : \mathbb{R}^n \to \mathbb{R}$, and are continuously differentiable functions. All the results presented here related to a local minima also hold if f and h_i are continuously differentiable within just an open set containing the local minimum.

Definition 22. *A feasible vector x is called regular if the constraint gradients $\nabla h_1(x), \ldots, \nabla h_m(x)$ are linearly independent. For a feasible vector x, we let $V(x) = \{\mathbf{y} : \nabla h_i(x)'\mathbf{y} = 0,\ \forall i = 1, \ldots, m\}$ be the subspace of first order feasible variations.*

Theorem 23. *Let x^* be a local minimum of f subject to $h(x) := (h_1(x), \ldots, h_m(x)) = 0$ and assume that x^* is regular. Then there exists a unique Lagrange multiplier vector $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)$ such that*

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) = 0.$$

If in addition f and h are twice continuously differentiable, then we have the second order necessary condition $y'(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*))y \geq 0\ \forall y \in V(x^*)$.

Proof: The main idea is to approximate the original constrained problem by an unconstrained one that involves a penalty for violation of the constraints and then applying the unconstraint optimality conditions. More precisely, for $k = 1, 2, \ldots$, let

$$F_k(x) = f(x) + \frac{k}{2}\|h(x)\|^2 + \frac{\alpha}{2}\|x - x^*\|^2,$$

where x^* is the local minimum of the constrained problem and $\alpha > 0$. The term $\frac{k}{2}\|h(x)\|^2$ imposes a penalty for violation of the constraint $h_i(x) = 0$ and $\frac{\alpha}{2}\|x - x^*\|_2^2$ is introduced to ensure x^* is a strict local minima. Since x^* is a local minimum $\exists\epsilon > 0$ such that $f(x^*) \le f(x)$ for all $x \in \bar{B}(x^*) := \{x : \|x - x^*\| \le \epsilon\}$. Let x_k be an optimal solution of the problem $\min_{x\in\bar{B}(x^*)} F_k(x)$. Then $\lim_{k\to\infty} x^k = x^*$. The reason is that $F^k(x^k) = f(x^k) + \frac{k}{2}\|h(x^k)\|^2 + \frac{\alpha}{2}\|x^k - x^*\|^2 \le F^k(x^*) = f(x^*) \implies \lim_{k\to\infty}\|h(x^k)\| = 0$, which implies that every limit point $\bar{x}$ of $\{x^k\}$ satisfies $h(\bar{x}) = 0$. Moreover, using the above relation as $k \to \infty$, we have $f(\bar{x}) + \frac{\alpha}{2}\|\bar{x} - x^*\|^2 \le f(x^*) \le f(\bar{x})$, where the last inequality holds because $\bar{x} \in \bar{B}(x^*)$. Since $\alpha > 0 \implies \bar{x} = x^*$. This shows that $\lim_{k\to\infty} x^k = x^*$ and in particular, for k sufficiently large, x^k is an interior point of $\bar{B}(x^*)$. This means that for sufficiently large k, x^k is an unconstrained local minimum of $F^k(x)$.

For sufficiently large k, using the first order unconstrained necessary conditions, we have:

$$0 = \nabla F^k(x^k) = \nabla f(x^k) + k\nabla h(x^k)h(x^k) + \alpha(x^k - x^*). \tag{4}$$

Since by regularity assumption, $\nabla h(x^*)$ has rank m, the same is true for $\nabla h(x^k)$, for sufficiently large k because $x^k \to x^*$. Therefore, $\nabla h(x^k)'\nabla h(x^k)$ is invertible, and by multiplying (4) with $(\nabla h(x^k)'\nabla h(x^k))^{-1}\nabla h(x^k)'$, we get:

$$kh(x^k) = -(\nabla h(x^k)'\nabla h(x^k))^{-1}\nabla h(x^k)'(\nabla f(x^k) + \alpha(x^k - x^*)) \implies \lim_{k\to\infty} kh(x) = \lambda^*$$

where $\lambda^* = -(\nabla f(x^*)'\nabla h(x^*))^{-1}\nabla h(x^*)'\nabla f(x^*)$. This in view of (4) as $k \to \infty$ also implies $\nabla f(x^*) + h(x^*)\lambda^* = 0$, completing the first part of the theorem.

The proof of the second part follows using the second order necessary condition for the unconstrained optimization $\min\{F^k(x) : x \in \bar{B}(x^*)\}$ when k is sufficiently large, i.e., $\nabla^2 F(x^k) \ge 0$. For the sake of brievity we omit the technical details here. □

Remark 6. *The existence of a Lagrange multiplier vector is not always guaranteed (e.g., in the absence of regularity of x^*). However, one can obtain some conditions under which λ^* exists. For instance, for linear constraints, it is known that a Lagrange multiplier vector always exists.*

5.2 The Lagrangian Function

Sometimes, it is easier to write the necessary conditions in terms of the Lagrangian function:

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m}\lambda_i h_i(x).$$

Then, if x^* is a regular local minimum, the conditions of the previous theorem can be written as:

$$\nabla_x L(x^*,\lambda^*) = 0, \quad \nabla_\lambda L(x^*,\lambda^*) = 0, \quad y'\nabla^2_{xx}L(x^*,\lambda^*)y \ge 0 \quad \forall y \in V(x^*).$$

The first two conditions represent a system of $n + m$ equations with $n + m$ unknowns (x^*, λ^*). Every local minimum that is regular together with its associated λ^* will be a solution of this system. However, a solution of this system need not correspond to a local minimum.

Example: Consider the optimization problem

$$\min\Big\{\frac{1}{2}(x_1^2 + x_2^2 + x_3^2) : \ x_1 + x_2 + x_3 = 3\Big\}.$$

The first order necessary conditions yield $x_1^* + \lambda^* = 0$, $x_2^* + \lambda^* = 0$, $x_3^* + \lambda^* = 0$, $x_1^* + x_2^* + x_3^* = 3$, $\implies x_1^* = x_2^* = x_3^* = 1$, $\lambda^* = -1$. Thus, $x^* = (1,1,1)$ is the unique candidate for a local minimum. Note that as the constraint gradient is $(1,1,1)$, all feasible vectors (and in particular x^*) are regular. Since $\nabla^2_{xx}L(x^*,\lambda^*) = I \geq 0$. This implies the second order necessary condition is also satisfied. In fact, by using the convexity of the objective function and the constraint set, we can use sufficiency conditions to argue that $x^* = (1,1,1)$ is a global minimum.

5.3 Extensions to Inequality Constraints (ICP)

Here we consider a more general problem involving both equality and inequality constraints:

$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t. } h_i(\mathbf{x}) = 0, \quad i = 1,\ldots,m, \\ &\quad\;\; g_j(\mathbf{x}) \leq 0, \quad j = 1,\ldots,r, \end{aligned}$$

where f, h_i, g_j are continuously differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$. The main idea for generalizing earlier results to ICP is to look at the set of active inequality constraints defined by $A(\mathbf{x}) = \{j : g_j(\mathbf{x}) = 0\}$ and treat them as equality constraints. The inactive constraints $j \notin A(\mathbf{x})$ don't matter at a local minimum. In other words, if $\mathbf{x}^*$ is a local minimum of ICP, then $\mathbf{x}^*$ is also a local minimum for a problem identical to ICP except that the inactive constraints $j \notin A(\mathbf{x}^*)$ have been discarded. On the other hand, if $\mathbf{x}^*$ is a local minimum of ICP, then $\mathbf{x}^*$ is also a local minimum for the equality constrained problem:

$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t. } h_i(\mathbf{x}) = 0, \quad i = 1,\ldots,m, \\ &\quad\;\; g_j(\mathbf{x}) = 0, \quad \forall j \in A(\mathbf{x}^*). \end{aligned}$$

Thus, if $\mathbf{x}^*$ is a regular local minimum for this equality constrained problem, there exist Lagrange multipliers λ^*, μ_j^*, $j \in A(x^*)$ such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j \in A(x^*)} \mu_j^* \nabla g_j(x^*) = 0.$$

Assigning $\mu_j^* = 0$ for $j \notin A(\mathbf{x}^*)$ we obtain an analog of the first-order optimality condition for the equality constrained problems.

Definition 24. *For any feasible point* $\mathbf{x}$*, the set of active inequality constraints is denoted by* $A(\mathbf{x}) = \{j : g_j(\mathbf{x}) = 0\}$*. A feasible vector* $\mathbf{x}$ *is called regular if* $\{\{\nabla h_i(\mathbf{x})\}_{i=1}^m, \{\nabla g_j(\mathbf{x})\}_{j \in A(\mathbf{x})}\}$ *are linearly independent. Finally, we define the Lagrangian function as*

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{r} \mu_j g_j(\mathbf{x}).$$

Theorem 25 (Karush-Kuhn-Tucker (KKT) Conditions). *Let* $\mathbf{x}^*$ *be a local minimum of ICP and assume that* $\mathbf{x}^*$ *is a regular local minimum. Then there exist unique Lagrange multiplier vectors* λ^* *and* μ^* *such that*

$$\nabla_x L(\mathbf{x}^*, \lambda^*, \mu^*) = 0, \quad \mu_j^* \geq 0, \forall j, \quad \mu_j^* = 0 \ \forall j \notin A(x^*).$$

If in addition $f, h,$ *and* g *are twice continuously differentiable, we have*

$$y' \nabla^2_{xx} L(\mathbf{x}^*, \lambda^*, \mu^*) y \geq 0,$$

$\forall y \in V(x^*)$, *where* $V(x^*) = \left\{ y \in \mathbb{R}^n : \nabla h_i(x^*)' y = 0 \ \forall i, \ \nabla g_j(x^*)' y = 0, \ \forall j \in A(x^*) \right\}$.

Remark 7. *The condition $\mu_j^* = 0,\ \forall j \notin A(x^*)$ can also compactly be written as $\mu_j^* g_j(x^*) = 0$, and is called complementary slackness condition (CS).*

Example Consider the optimization problem

$$\min\left\{\frac{1}{2}(x_1^2 + x_2^2 + x_3^2) :\ x_1 + x_2 + x_3 \geq -3\right\}.$$

Then, for a local minimum $\mathbf{x}^*$, the first order necessary condition yields $x_1^* + \mu^* = 0,\ \ x_2^* + \mu^* = 0,\ \ x_3^* + \mu^* = 0$. One approach for solving this system is to consider separately all the possible combinations of constraints being active or inactive. Here, there are two possibilities:

1. The constraint is inactive $\implies \mu^* = 0 \implies x_1^* = x_2^* = x_3^* = 0$, which contradicts $x_1^* + x_2^* + x_3^* > -3$
2. The constraint is active $x_1^* + x_2^* + x_3^* = -3 \implies x_1^* = x_2^* = x_3^* = -1,\ \ \mu^* = 1$. Using the sufficiency condition (where positive semi-definiteness is replaced by positive definiteness in the above theorem), one can see that x^* satisfies the second-order sufficient condition, and is indeed a local minimum.