Optimization Notes

Xiaopeng

September 22, 2025

1 Remainders from Multivariable Calculus

1.1 First-Order Conditions

The first-order necessary conditions for optimality can be expressed using the gradient of the objective function and the constraints. Specifically, if x^* is a local minimum of $f(x)$ Then

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

Proof: Writing the Taylor expansion of f around x^* gives

$$f(x) = f(x^*) + \nabla f(x^*)^T(x - x^*) + o(\|x - x^*\|)$$

Since x^* is a local minimum, we have $f(x) \geq f(x^*)$ for all x in a neighborhood of x^*. This implies that the first-order term must vanish, leading to the conclusion that $\nabla f(x^*) = 0$. More rigorously, we can consider the directional derivative of f at x^* in the direction of any vector d:

$$D_f(x^*; d) = \nabla f(x^*)^T d$$

Since x^* is a local minimum, the directional derivative must be non-negative for all feasible directions d (Univariate result: if f is differentiable and x^* is a local minimum, then $f'(x^*) = 0$). Otherwise this will be a decreasing direction, contradicting the local minimality of x^*. Therefore, we have:

$$D_f(x^*; d) \geq 0 \quad \forall d \in \mathcal{D}$$

In particular, if we take $d = -\nabla f(x^*)$, we find that

$$D_f(x^*; -\nabla f(x^*)) = -\|\nabla f(x^*)\|^2 \leq 0$$

This implies that $\nabla f(x^*) = 0$, completing the proof.

1.2 Second-Order Conditions

Aussume f is $\mathscr{C}^2$ and let x^* be a point such that $\nabla f(x^*) = 0$. Then

1. (Necessary condition) If x^* is a local minimum of f, then $\nabla^2 f(x^*) \in S_d^+(\mathbb{R})$ (the set of positive semi-definite matrices);

2. If $\nabla^2 \in S_d^{++}(\mathbb{R})$, then x^* is a strict local minimum of f;

3. If $\nabla^2 f(x^*)$ has at least one negative and one positive eigenvalue, then x^* is a saddle point of f: there exist two orthogonal directions e_1 and e_2 such that $t^* = 0$ is a local minimiser for $t \mapsto f(x^* + te_1)$ and a local maximiser for $t \mapsto f(x^* + te_2)$;

4. If $\nabla^2 f(x^*) \in S_d^+(\mathbb{R})$, but not in $S_d^{++}(\mathbb{R})$, then we cannot conclude and further analysis is required.

Proof: To prove the suffcient condition, we assume that

$$\nabla^2 f(x^*) \in S_d^{++}(\mathbb{R}) \tag{1}$$

Proposition 1.1 (Second-order mean value theorem)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathscr{C}^2$ on an open set containing the segment $[x, y] = \{x + t(y - x) : t \in [0, 1]\}$. Then there exists $\theta \in (0, 1)$ such that*

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \tfrac{1}{2} (y - x)^\top \nabla^2 f(x + \theta(y - x)) (y - x).$$

In particular, if $\nabla f(x) = 0$, then

$$f(y) = f(x) + \tfrac{1}{2} (y - x)^\top \nabla^2 f(x + \theta(y - x)) (y - x).$$

1. Let $d \in \mathbb{R}^d$ be any direction. By Taylor's theorem, we have

$$f(x^* + d) = f(x^*) + \nabla f(x^*)^T d + \frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2)$$

Since $\nabla f(x^*) = 0$, this simplifies to

$$f(x^* + d) = f(x^*) + \frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2)$$

By the minimality of x^*, we have $f(x^* + d) \geq f(x^*)$ for all sufficiently small d. This implies that

$$\frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2) \geq 0$$

for all sufficiently small d. Since the $o(\|d\|^2)$ term becomes negligible as d approaches zero, we conclude that

$$\frac{1}{2} d^T \nabla^2 f(x^*) d \geq 0$$

for all sufficiently small d. This is equivalent to saying that $\nabla^2 f(x^*)$ is positive semi-definite, which proves the necessary condition.

2. (Proof by contradiction, also a bit like contrapositive) Assume that there exists a sequence $\{x_k\}_{k \in \mathbb{N}}$ such that

$$\forall k \in \mathbb{N}, f(x_k) \leq f(x^*)$$

Mean value formula suggests that for x_k there exist $\xi_k \in [x_k, x^*]$ such that

$$f(x_k) = f(x^*) + \frac{1}{2} \langle \nabla^2 f(\xi_k)(x_k - x^*), (x_k - x^*) \rangle$$

The key point here is to use the exact mean value formula then which avoids any $o(\|d\|^2)$ terms. The tricky part is to carefully choose a sequence on the right hand side whose limit's Euclidean norm is strictly positive, and the sequence should also be a simple transformation from the sequence $x_k - x^*$. For this purpose, we define:

$$z_k := \frac{x_k - x^*}{\|x_k - x^*\|}$$

For any $k \in \mathbb{N}$ we have $\|z_k\| = 1$. By the compactness of the unit sphere, we can extract a convergent subsequence $z_{k_j} \to z^*$ for some $z^* \in S^{d-1}$. For the sake of clean notation, we will still denote the subsequence by z_k. We assume that z_k converges to $z_\infty \in \mathbb{S}^{d-1}$. Taking the result from mean value formula, we have (up to multiplying by a suitable factor) :

$$\langle \nabla^2 f(\xi_k) z_k, z_k \rangle \leq 0$$

Since $\xi_k \to x^*$ as $k \to \infty$, by the continuity of $\nabla^2 f$ we have $\nabla^2 f(\xi_k) \to \nabla^2 f(x^*)$. Taking the limit on both sides gives

$$\langle \nabla^2 f(x^*) z_\infty, z_\infty \rangle \leq 0$$

which contradicts the assumption that $\nabla^2 f(x^*)$ is positive definite ($\|z_\infty\|^2 = 1$).

3. Let e_1 and e_2 be the eigenvectors corresponding to the negative and positive eigenvalues of $\nabla^2 f(x^*)$, respectively. Consider the functions,

$$g_1(t) = f(x^* + te_1), \quad g_2(t) = f(x^* + te_2)$$

for $t \in \mathbb{R}$. By the chain rule (**use it or it's proving it myself in a paticular case**),

$$g'(t) = \frac{d}{dt} f(\gamma(t)) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\gamma(t)) \frac{d}{dt}\gamma_k(t) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x^* + tv)\, v_k.$$

$$g_1'(t) = \langle \nabla f(x^* + te_1), e_1 \rangle, \qquad g_2'(t) = \langle \nabla f(x^* + te_2), e_2 \rangle,$$

and

$$\frac{d}{dt}\Big(\frac{\partial f}{\partial x_k}(x^* + tv)\Big) = \sum_{\ell=1}^{n} \frac{\partial^2 f}{\partial x_k \partial x_\ell}(x^* + tv) \frac{d}{dt}(x_\ell^* + tv_\ell) = \sum_{\ell=1}^{n} \frac{\partial^2 f}{\partial x_k \partial x_\ell}(x^* + tv)\, v_\ell.$$

$$g_1''(t) = \langle \nabla^2 f(x^* + te_1)\, e_1, e_1 \rangle, \qquad g_2''(t) = \langle \nabla^2 f(x^* + te_2)\, e_2, e_2 \rangle.$$

In particular,

$$g_1'(0) = \langle \nabla f(x^*), e_1 \rangle = 0, \quad g_2'(0) = \langle \nabla f(x^*), e_2 \rangle = 0,$$

and, since e_i are eigenvectors of $\nabla^2 f(x^*)$ with eigenvalues λ_i,

$$g_1''(0) = \lambda_1 < 0, \qquad g_2''(0) = \lambda_2 > 0.$$

By continuity of $\nabla^2 f$, there exists $\delta > 0$ such that for $|t| < \delta$,

$$g_1''(t) < 0 \quad \text{and} \quad g_2''(t) > 0.$$

Hence $t = 0$ is a strict local maximizer of g_1 and a strict local minimizer of g_2, so x^* is a saddle point.

1.3 Coercive functions

Definition 1.2. We say that a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is coercive if

$$\lim_{\|x\|\to\infty} f(x) = +\infty.$$

Remark

Another definition is for any $M \in \mathbb{R}$ the sub-level set $\{x \in \mathbb{R}^d : f(x) \leq M\}$ is compact (bounded).

Proposition 1.3. *If $f : \mathbb{R}^d \to \mathbb{R}$ is coercive, then it attains a global minimum.*

Proof: Let $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^d$ be a minimizing sequence for f, i.e.,

$$\lim_{k\to\infty} f(x_k) = \inf_{x\in\mathbb{R}^d} f(x) = m.$$

Since f is coercive, we have $f(x_k) \to +\infty$ as $\|x_k\| \to \infty$. Thus, the sequence $(x_k)_{k\in\mathbb{N}}$ must be bounded. By the Bolzano-Weierstrass theorem, we can extract a convergent subsequence $(x_{k_j})_{j\in\mathbb{N}}$ such that

$$\lim_{j\to\infty} x_{k_j} = x^* \in \mathbb{R}^d.$$

By the continuity of f, we have

$$\lim_{j\to\infty} f(x_{k_j}) = f(x^*).$$

Combining these limits gives

$$f(x^*) = \lim_{j\to\infty} f(x_{k_j}) = m,$$

which shows that f attains its global minimum at x^*.