Duality and optimality conditions

Optimality conditions

We do *not* assume convexity by default here.

Unconstrained problems

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$$

- Zeroth-order optimality condition
$$f(\mathbf{x}^\star) \le f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$
    - Global optimality
    - Not easy to check
- First-order optimality condition: assume f is continuously differentiable.
$$\nabla f(\mathbf{x}^\star) = 0.$$
    - Also known as the stationarity condition.
    - *Necessary* condition: $\mathbf{x}^\star$ may *locally* minimize/maximize f, or even is a saddle point.
    - If f is convex, then necessary and sufficient.
- Second-order optimality condition: assume f is twice continuously differentiable.
$$\nabla^2 f(\mathbf{x}^\star) \succeq 0.$$
    - *Necessary* condition: $\mathbf{x}^\star$ *locally* minimize f (if the first-order condition also holds).

Constrained problems

$$\begin{aligned}
&\text{minimize } f_0(\mathbf{x}) \\
&\text{subject to } f_i(\mathbf{x}) \le 0, \quad i = 1, \dots, m \\
&\phantom{\text{subject to }} h_i(\mathbf{x}) = 0, \quad i = 1, \dots, p.
\end{aligned}$$

Assume all of these functions share some open set $U \subset \mathbb{R}^d$ as their domain.

- Zeroth-order optimality condition
$$f(\mathbf{x}^\star) \le f(\mathbf{x}), \quad \forall \mathbf{x} \in C = \{\mathbf{x} : f_i(\mathbf{x}) \le 0,\ i = 1, \dots, m,\ h_i(\mathbf{x}) = 0,\ i = 1, \dots, p\}.$$
- First-order optimality condition: Karush-Kuhn-Tucker (KKT) conditions

> Kuhn, H. W.; Tucker, A. W., Nonlinear Programming. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 481--492, University of California Press, Berkeley, Calif., 1951. https://projecteuclid.org/euclid.bsmsp/1200500249
>
> Karush, W., Minima of Functions of Several Variables with Inequalities as Side Constraints. M.Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois, 1939.

KKT conditions

- Define the **Lagrangian**
$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}).$$
The vectors $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)^T$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_p)^T$ are called the **Lagrange multiplier vectors** or **dual variables**.
- Let $\mathbf{x}^\star$ be a local minimizer of f_0 in the feasible region C.
- Assume $f_0, f_1, \dots, f_m, h_1, \dots, h_p$ are continuously differentiable near $\mathbf{x}^\star$.
- (*Mangasarian-Fromovitz constraint qualification*). Assume
    1. the gradients $\nabla f_i(\mathbf{x}^\star)$, $i = 1, \dots, m$, be linearly independent, and
    2. there exists a vector $\mathbf{v}$ with $\langle \nabla h_i(\mathbf{x}^\star), \mathbf{v} \rangle = 0$ for $i = 1, \dots, p$, and $\langle \nabla f_i(\mathbf{x}^\star), \mathbf{v} \rangle < 0$ for all inequality constraints active at $\mathbf{x}^\star$, i.e., for i with $f_i(\mathbf{x}^\star) = 0$.

    (The vector $\mathbf{v}$ is a tangent vector in the sense that infinitesimal motion from $\mathbf{x}^\star$ along $\mathbf{v}$ stays within the feasible region.)
- (**Lagrange multiplier rule**). Suppose that the objective function f_0 of the constrained optimization problem above has a local minimum at the feasible point $\mathbf{x}^\star$. If f_0 and the constraint functions are continuously differentiable near $\mathbf{x}^\star$, and the Mangasarian-Fromovitz constraint qualification holds at $\mathbf{x}^\star$, then there exist Lagrange multipliers $\lambda_1^\star, \dots, \lambda_m^\star$ and $\nu_1^\star, \dots, \nu_p^\star$ such that
$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) = \nabla f_0(\mathbf{x}^\star) + \sum_{i=1}^m \lambda_i^\star \nabla f_i(\mathbf{x}^\star) + \sum_{i=1}^p \nu_i^\star \nabla h_i(\mathbf{x}^\star) = \mathbf{0}.$$
Moreover, each of the multipliers $\lambda_i^\star$ associated with the inequality constraints is nonnegative, and $\lambda_i^\star = 0$ whenever $f_i(\mathbf{x}^\star) < 0$.
- We can summarize the **KKT conditions** as
$$\begin{aligned}
f_i(\mathbf{x}^\star) &\le 0, \quad i = 1, \dots, m \\
h_i(\mathbf{x}^\star) &= 0, \quad i = 1, \dots, p \\
\lambda_i^\star &\ge 0, \quad i = 1, \dots, m \\
\lambda_i^\star f_i(\mathbf{x}^\star) &= 0, \quad i = 1, \dots, m \\
\nabla f_0(\mathbf{x}^\star) + \sum_{i=1}^m \lambda_i^\star \nabla f_i(\mathbf{x}^\star) + \sum_{i=1}^p \nu_i^\star \nabla h_i(\mathbf{x}^\star) &= \mathbf{0}.
\end{aligned}$$
- The fourth condition $\lambda_i^\star f_i(\mathbf{x}^\star) = 0$ is called the **complementary slackness**.
- Remember that the KKT conditions are collectively a necessary condition for local optimality.
- If the optimization problem is *convex*, then they become a necessary and sufficient condition, i.e., finding a triple $(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ that satisfies the KKT conditions guarantees global optimiality of the triple.
    - Many algorithms for convex optimization are conceived as, or can be interpreted as, methods for solving the KKT conditions.
- (*Slater's constraint qualification condition*). Assume there exists a feasible point $\mathbf{x}_0$ such that $f_i(\mathbf{x}_0) < 0$ for active inequality constraints, i.e., for all i with $f_i(\mathbf{x}^\star) = 0$.
    - If f_i are convex, then Slater's condition implies the Mangasarian-Fromovitz condition.

Example: MLE of multinomial probabilities

The likelihood of observation $(n_1, \dots, n_r)$ from $\text{mult}(n, (p_1, \dots, p_r))$ is $\prod_{i=1}^r p_i^{n_i}$. The MLE problem is

$$\begin{aligned}
&\text{minimize } -\sum_{i=1}^r n_i \log p_i \\
&\text{subject to } -p_i \le 0, \quad i = 1, \dots, r \\
&\sum_{i=1}^r p_i = 1.
\end{aligned}$$

This is a convex optimization problem. The Lagrangian is

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \nu) = -\sum_{i=1}^r n_i \log p_i - \sum_{i=1}^r \lambda_i p_i + \nu\left(\sum_{i=1}^r p_i - 1\right).$$

Differentiate with respect to $\mathbf{p}$ to have

$$0 = -\frac{n_i}{p_i} - \lambda_i + \nu, \quad i = 1, \dots, r.$$

Multiply p_i on both sides and sum on i to have

$$0 = -n - \sum_{i=1}^r \lambda_i p_i + \nu \sum_{i=1}^r p_i = -n - 0 + \nu$$

due to complementary slackness. Setting all $\lambda_i = 0$ gives $p_i = n_i/n$. The triple $((n_1/n, \dots, n_r/n), \mathbf{0}, n)$ satisfies the KKT conditions and hence is a global solution (Even if some $n_i = 0$).

Second-order optimality condition

Suppose the Lagrangian multiplier rule is satisfied at a point $\mathbf{x}^\star$ with associated multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$. Assume the Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is twice continuously differentiable in $\mathbf{x}$ near $\mathbf{x}^\star$. If $\mathbf{v}^T \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\nu}) \mathbf{v} > 0$ for all $\mathbf{v} \ne \mathbf{0}$ satisfying $\langle \nabla h_i(\mathbf{x}^\star), \mathbf{v} \rangle = 0$ for all i and $\langle \nabla f_i(\mathbf{x}^\star), \mathbf{v} \rangle \le 0$ for all active constraints, then $\mathbf{x}^\star$ is a local minimum of f_0.

- That is, $\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is positive definite for all tangent vectors at $\mathbf{x}^\star$.

Duality

- The **Lagrange dual function** is the minimum value of the Langrangian over $\mathbf{x}$
$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) \right).$$
    - No constraints are imposed on $\mathbf{x}$ in defining $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$.
    - Dual function $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is **concave** (subject to $\boldsymbol{\lambda} \ge \mathbf{0}$) regardless of the convexity of the primal (original) problem (why?)
- Denote the optimal value of original problem by $p^\star = \inf_{\mathbf{x}\in C} f_0(\mathbf{x})$. For any $\boldsymbol{\lambda} \ge \mathbf{0}$ and any $\boldsymbol{\nu}$, we have
$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \le p^\star.$$
Proof: For any feasible point $\tilde{\mathbf{x}}$,
$$\mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^m \lambda_i f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^p \nu_i h_i(\tilde{\mathbf{x}}) \le f_0(\tilde{\mathbf{x}})$$
because the second term is non-positive (since $\boldsymbol{\lambda} \ge \mathbf{0}$) and the third term is zero. Then
$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \le L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \le f_0(\tilde{\mathbf{x}}).$$
- Since each pair $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ with $\boldsymbol{\lambda} \ge \mathbf{0}$ gives a lower bound to the optimal value $p^\star$. It is natural to ask for the *best possible lower bound* for $p^\star$ the Lagrange dual function can provide. This leads to the **Lagrange dual problem**
$$\begin{aligned}
&\text{maximize } g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\
&\text{subject to } \boldsymbol{\lambda} \ge \mathbf{0},
\end{aligned}$$
which is a convex problem whether or not the primal problem is convex.
- Let $d^\star = \sup_{\boldsymbol{\lambda}\ge\mathbf{0}, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu})$. Then we have **weak duality**
$$d^\star \le p^\star.$$
The difference $p^\star - d^\star$ is called the **duality gap**.
- **Strong duality** refers to zero duality gap. This occurs when the problem is convex and the Lagrange multiplier rule holds at a feasible point $\hat{\mathbf{x}}$. Let the corresponding Lagrange multipliers be $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\nu}})$. In this case, $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is convex in $\mathbf{x}$. Since the Lagrange multiplier rule states that $\nabla_{\mathbf{x}} \mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\nu}}) = \mathbf{0}$, point $\hat{\mathbf{x}}$ furnishes an unconstrained minimum of the Lagrangian. Hence
$$p^\star \le f(\hat{\mathbf{x}}) = \mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\nu}}) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\nu}}) = g(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\nu}}) \le d^\star.$$
(Why the first equality?)

    That is, $d^\star = p^\star$.
- For convex problems, conditions for the Lagrange multiplier rule holds also yields strong duality, e.g., **Slater's condition**.

Example: QP dual

Consider the dual of QP

$$\begin{aligned}
&\text{minimize } \frac{1}{2}\mathbf{x}^T \mathbf{P} \mathbf{x} + \mathbf{q}^T \mathbf{x} \\
&\text{subject to } \mathbf{A}\mathbf{x} = \mathbf{b}
\end{aligned}$$

($\mathbf{P} \succ \mathbf{0}$).

The Lagrangian is

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\nu}) = \frac{1}{2}\mathbf{x}^T \mathbf{P} \mathbf{x} + \mathbf{q}^T \mathbf{x} + \boldsymbol{\nu}^T (\mathbf{A}\mathbf{x} - \mathbf{b}).$$

The dual function is

$$\begin{aligned}
g(\boldsymbol{\nu}) &= \inf_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T \mathbf{P} \mathbf{x} + \mathbf{q}^T \mathbf{x} + \boldsymbol{\nu}^T (\mathbf{A}\mathbf{x} - \mathbf{b}) \\
&= \inf_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T \mathbf{P} \mathbf{x} + (\mathbf{A}^T \boldsymbol{\nu} + \mathbf{q})^T \mathbf{x} - \boldsymbol{\nu}^T \mathbf{b} \\
&= -\frac{1}{2}(\mathbf{A}^T \boldsymbol{\nu} + \mathbf{q})^T \mathbf{P}^{-1} (\mathbf{A}^T \boldsymbol{\nu} + \mathbf{q}) - \boldsymbol{\nu}^T \mathbf{b}.
\end{aligned}$$

The infimum is attained with

$$\mathbf{x} = -\mathbf{P}^{-1}(\mathbf{A}^T \boldsymbol{\nu} + \mathbf{q}).$$

Thus the dual problem is an **unconstraint** QP, which is easier to solve than the primal.

The maximum of $g(\boldsymbol{\nu})$ occurs at

$$\mathbf{0} = -\mathbf{b} - \mathbf{A}\mathbf{P}^{-1}(\mathbf{A}^T \boldsymbol{\nu} + \mathbf{q})$$

or

$$\boldsymbol{\nu} = -(\mathbf{A}\mathbf{P}^{-1}\mathbf{A})^{-1}(\mathbf{b} + \mathbf{A}\mathbf{P}^{-1}\mathbf{q}).$$

Therefore the primal optimum is attained at

$$\mathbf{x}^\star = -\mathbf{P}^{-1}(-\mathbf{A}^T(\mathbf{A}\mathbf{P}^{-1}\mathbf{A})^{-1}(\mathbf{b} + \mathbf{A}\mathbf{P}^{-1}\mathbf{q}) + \mathbf{q}) = \mathbf{P}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{P}^{-1}\mathbf{A})^{-1}(\mathbf{b} + \mathbf{A}\mathbf{P}^{-1}\mathbf{q}) - \mathbf{P}^{-1}\mathbf{q}.$$

Min-max inequality

Note

$$\sup_{\boldsymbol{\lambda}\ge\mathbf{0}, \boldsymbol{\nu}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\lambda}\ge\mathbf{0}, \boldsymbol{\nu}} f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) = \begin{cases} f_0(\mathbf{x}) & \text{if } \mathbf{x} \in C \\ \infty & \text{otherwise} \end{cases} = f_0(\mathbf{x}) + \iota_C(\mathbf{x})$$

where

$$C = \{\mathbf{x} : f_i(\mathbf{x}) \le 0,\ i = 1, \dots, m,\ h_i(\mathbf{x}) = 0,\ i = 1, \dots, p\}.$$

Thus weak duality is

$$d^\star = \sup_{\boldsymbol{\lambda}\ge\mathbf{0}, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\lambda}\ge\mathbf{0}, \boldsymbol{\nu}} \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \le \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda}\ge\mathbf{0}, \boldsymbol{\nu}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}\in C} f_0(\mathbf{x}) = p^\star$$

or min-max inequlity.

Strong duality amounts to prove the minimax theorem

$$\sup_{\boldsymbol{\lambda}\ge\mathbf{0}, \boldsymbol{\nu}} \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda}\ge\mathbf{0}, \boldsymbol{\nu}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

- For some *nonconvex* problems strong duality may hold.
- Not all *convex* problems have strong duality.

References

Boyd & Vandenberghe (Ch. 5)

Lange, K., 2016. MM optimization algorithms (Vol. 147). SIAM. (Ch. 3)

Lange, K., 2010. Numerical analysis for statisticians. Springer Science & Business Media. (Ch. 11)


```
In [ ]:
```