SDSC6015 - Semester A, 2025

Stochastic Optimization for Machine Learning

3. Lagrangian Duality & Optimality Conditions

1 Lagrange Dual Problem

Consider a generic optimization problem (possibly nonconvex)

$$\begin{aligned} \mathrm{P}: \quad & \underset{\boldsymbol{x} \in \mathbb{R}^n}{\text{minimize}} && f_0(\boldsymbol{x}) \\ & \text{subject to} && f_i(\boldsymbol{x}) \leq 0, \quad \forall i \in [m] \\ & && h_j(\boldsymbol{x}) = 0, \quad \forall j \in [p] \end{aligned}$$

where $f_0 : \mathbb{R}^n \to \mathbb{R}$, $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i \in [m]$ and $h_j : \mathbb{R}^n \to \mathbb{R}$ for $j \in [p]$. We denote $\mathcal{D} \subseteq \mathbb{R}^n$ to be the domain of this problem P. We refer to P as the **primal problem**.

Definition 1 (Lagrangian)**.** *The Lagrangian $L : \mathcal{D} \times \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$ for problem P (where $\mathcal{D} \subseteq \mathbb{R}^n$ is the domain of the problem) is defined as*

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{j=1}^{p} \mu_j h_j(\boldsymbol{x})$$

Definitions and Properties

- $\approx$ weighted sum of objective and constraint functions
- λ_i is Lagrange multiplier corresponding to $f_i(\boldsymbol{x}) \leq 0$
- μ_i is Lagrange multiplier corresponding to $h_i(\boldsymbol{x}) = 0$
- For any fixed $\boldsymbol{x} \in \mathcal{D}$, the Lagrangian is **concave (affine)** in $(\boldsymbol{\lambda}, \boldsymbol{\mu})$

1.1 Equivalent Formulation using Lagrangian

Lagrangian let us to re-express P as a *min-max problem.* To see this, we define

$$f(\boldsymbol{x}) = \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \boldsymbol{\mu}\in\mathbb{R}^p} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \boldsymbol{\mu}\in\mathbb{R}^p} \left(f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{j=1}^{p} \mu_j h_j(\boldsymbol{x}) \right).$$

Note:

$$f(\boldsymbol{x}) = \begin{cases} f_0(\boldsymbol{x}) \quad \text{if} \begin{cases} f_i(\boldsymbol{x}) \le 0, & \forall i \in [m] \\ h_j(\boldsymbol{x}) = 0, & \forall j \in [p] \end{cases} \\ \infty \quad \text{otherwise} \end{cases}$$

Thus, we obtain

$$\inf \mathrm{P} \;=\; \inf_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x}) \;=\; \inf_{\boldsymbol{x}\in\mathcal{D}} \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \boldsymbol{\mu}\in\mathbb{R}^p} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})$$

1.2 Lagrange Dual Function

Definition 2 (Lagrange Dual Function)**.** *The Lagrangian dual function* $g : \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$ *for problem P is defined as*

$$g(\boldsymbol{\lambda},\boldsymbol{\mu}) = \inf_{\boldsymbol{x}\in\mathcal{D}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = \inf_{\boldsymbol{x}\in\mathcal{D}} \left(f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{j=1}^{p} \mu_j h_j(\boldsymbol{x}) \right)$$

where $\mathcal{D} \subseteq \mathbb{R}^n$ *is the domain of the problem P.*

Note: g is a concave function, and it can returns $-\infty$ for some $(\boldsymbol{\lambda},\boldsymbol{\mu})$.

Proposition 1.

$$g(\boldsymbol{\lambda},\boldsymbol{\mu}) \le \inf P, \qquad \forall \boldsymbol{\lambda} \in \mathbb{R}^m_+,\ \boldsymbol{\mu} \in \mathbb{R}^p$$

Example: Least-norm solution of linear equations

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \boldsymbol{x}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

Question: What are the Lagrangian and Lagrange dual function?

Example: Two-way partitioning

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subject to} \quad & x_i^2 = 1, \quad \forall i \in [n] \end{aligned}$$

Question: Is the above a convex problem? What are the Lagrangian and Lagrange dual function?

2 Weak and Strong Duality

2.1 The Dual Problem

Since the Lagrange dual function gives the lower bound of the objective function, one can seek the best lower bound via

$$\mathrm{D}: \quad \underset{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \boldsymbol{\mu}\in\mathbb{R}^p}{\text{maximize}} \quad g(\boldsymbol{\lambda}, \boldsymbol{\mu})$$

The above is called the **dual (maximization) problem**. The problem D, by construction, is equivalent to a *max-min problem* as

$$\sup \mathrm{D} \;=\; \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \boldsymbol{\mu}\in\mathbb{R}^p} g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \;=\; \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \boldsymbol{\mu}\in\mathbb{R}^p} \inf_{\boldsymbol{x}\in\mathcal{D}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

2.2 Weak Duality

Corollary 1 (Weak Duality)**.**

$$\sup D \leq \inf P$$

Question: What happen if D or P is unbounded?

2.3 Significance of Dual Solutions

Every *feasible solution* of P (D) provides an upper (lower) bound on both $\inf \mathrm{P}$ and $\sup \mathrm{D}$.

Question: Assume that $\hat{\boldsymbol{x}}$ is a feasible candidate solution of P, how can we verify its quality $f_0(\hat{\boldsymbol{x}}) - \inf \mathrm{P}$ if $\inf \mathrm{P}$ is unknown?

Answer: A dual feasible solution $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ provides a certificate that

$$\inf \mathrm{P} \geq g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \quad \Longrightarrow \quad f_0(\hat{\boldsymbol{x}}) - \inf \mathrm{P} \leq f_0(\hat{\boldsymbol{x}}) - g(\boldsymbol{\lambda}, \boldsymbol{\mu})$$

Thus, if $f_0(\hat{\boldsymbol{x}}) - g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \epsilon$, $\hat{\boldsymbol{x}}$ is an $\epsilon-$optimal solution.

2.4 Strong Duality

Definition 3 (Strong Duality)**.** *Consider a primal problem P and its dual problem D. Denote the duality gap* $\Delta = \inf P - \sup D$*. By weak duality,* $\Delta \geq 0$*. Strong duality holds if* $\Delta = 0$*.*

Note:

- In general, strong duality does not hold.
- Strong duality always hold if (i) P is convex, and (ii) P satisfies *constraint qualification.*

Definition 4 (Slater's constraint qualification)**.** *Consider the generic convex optimization problem*

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\textit{minimize}} \quad & f_0(\boldsymbol{x}) \\ \textit{subject to} \quad & f_i(\boldsymbol{x}) \leq 0, \quad \forall i \in [m] \\ & \boldsymbol{a}_i^\top \boldsymbol{x} = b_i, \quad \forall i \in [p] \end{aligned}$$

Slater's constraint qualification holds if it is strictly feasible, i.e.there exists $\hat{\boldsymbol{x}} \in \boldsymbol{int}(\mathcal{D})$ *where*

1. $f_i(\hat{\boldsymbol{x}}) < 0, \quad \forall i \in [m]$
2. $\boldsymbol{a}_i^\top \hat{\boldsymbol{x}} = b_i, \quad \forall i \in [p]$

3 Duality: LP & QP

3.1 Linear Program

Consider the following LP

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{aligned}$$

Question: What is the corresponding dual problem?

3.2 Quadratic Program

Consider the following QP

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x} \qquad (\boldsymbol{P} \succ \boldsymbol{0}) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

Question: What is the corresponding dual problem?

4 Duality: SOCP

Definition 5 (Dual Norm). *Let* $\|\cdot\|$ *be a norm on* $\mathbb{R}^n$*. The associated* **dual norm***, denoted* $\|\cdot\|_\star$*, is defined as*

$$\|\boldsymbol{z}\|_\star = \sup\{\boldsymbol{z}^\top \boldsymbol{x} \mid \|\boldsymbol{x}\| \leq 1\}.$$

Euclidean norms:

- If $\|\cdot\| = \|\cdot\|_2$, then $\|\cdot\|_\star = \|\cdot\|_2$
- If $\|\cdot\| = \|\cdot\|_\infty$, then $\|\cdot\|_\star = \|\cdot\|_1$
- If $\|\cdot\| = \|\cdot\|_1$, then $\|\cdot\|_\star = \|\cdot\|_\infty$
- If $\|\cdot\| = \|\cdot\|_p$, then $\|\cdot\|_\star = \|\cdot\|_q$ where $1/p + 1/q = 1$

Consider

$$\begin{aligned}
\underset{\boldsymbol{x}\in\mathbb{R}^n,\, \boldsymbol{y}\in\mathbb{R}^{nm},\, \boldsymbol{t}\in\mathbb{R}^m}{\text{minimize}} \quad & \boldsymbol{c}_0^\top \boldsymbol{x} \\
\text{subject to} \quad & \|\boldsymbol{y}_i\|_2 \le t_i, && \forall i \in [m] \\
& \boldsymbol{y}_i = \boldsymbol{A}_i \boldsymbol{x} + \boldsymbol{b}_i, && \forall i \in [m] \\
& t_i = \boldsymbol{c}_i^\top \boldsymbol{x} + d_i, && \forall i \in [m]
\end{aligned}$$

where $\boldsymbol{c}_0 \in \mathbb{R}^n$, $\boldsymbol{A}_i \in \mathbb{R}^{n_i \times n}$, $\boldsymbol{b}_i \in \mathbb{R}^{n_i}$, $\boldsymbol{c}_i \in \mathbb{R}^n$, $d_i \in \mathbb{R}$, $\forall i \in [m]$.

Question: What is the corresponding dual problem?

5 Optimality Conditions

5.1 Complementary Slackness

Proposition 2. *Consider a primal problem P and its corresponding dual problem D, and assume strong duality holds. Denote $\boldsymbol{x}^\star \in \mathbb{R}^n$ to be the optimal solution in P and $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star) \in \mathbb{R}_+^m \times \mathbb{R}^p$ is optimal solution in D. Then*

(i) $\boldsymbol{x}^\star$ *minimizes* $L(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$

(ii) $\sum_{i=1}^{m} \lambda_i^\star \cdot f_i(\boldsymbol{x}^\star) = 0$, *which implies complementary slackness, i.e.*

$$\lambda_i^\star > 0 \Longrightarrow f_i(\boldsymbol{x}^\star) = 0 \quad \textit{and} \quad f_i(\boldsymbol{x}^\star) < 0 \Longrightarrow \lambda_i^\star = 0, \qquad \forall i \in [m].$$

5.2 Karush–Kuhn–Tucker (KKT) Conditions

Theorem 1. *Consider a primal problem P and its corresponding dual problem D. Assume the functions $f_0, \ldots, f_m, h_1, \ldots, h_p$ are differentiable, $\boldsymbol{x}^\star \in \mathbb{R}^n$ is optimal in P, $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star) \in \mathbb{R}_+^m \times \mathbb{R}^p$ is optimal in D, and strong duality holds (but P may be nonconvex). Then, $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ satisfies the Karush–Kuhn–Tucker (KKT) conditions:*

$$\begin{aligned}
&f_i(\boldsymbol{x}^\star) \leq 0, && \forall i \in [m] && : \textit{primal feasibility}\\
&h_i(\boldsymbol{x}^\star) = 0, && \forall j \in [p] && : \textit{primal feasibility}\\
&\lambda_i^\star \geq 0, && \forall i \in [m] && : \textit{dual feasibility}\\
&\lambda_i^\star \cdot f_i(\boldsymbol{x}^\star) = 0, && \forall i \in [m] && : \textit{complementary slackness}\\
&\nabla f_0(\boldsymbol{x}^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(\boldsymbol{x}^\star) + \sum_{j=1}^{p} \mu_j^\star \nabla h_j(\boldsymbol{x}^\star) = 0 && && : \textit{stationary}
\end{aligned}$$

Theorem 2 (Sufficiency)**.** *Consider a convex primal problem P with differentiable objective and constraint functions and* $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star) \in \mathbb{R}^n \times \mathbb{R}_+^m \times \mathbb{R}^p$ *satisfies the KKT conditions, then* $\boldsymbol{x}^\star$ *is optimal in P and* $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ *is optimal in the corresponding dual problem D.*

Theorem 3 (Necessity)**.** *Consider a convex primal problem P with differentiable objective and constraint functions satisfies Slater's condition. If there exists* $\boldsymbol{x}^\star \in \mathbb{R}^n$ *that is optimal in P, then there exists* $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star) \in \mathbb{R}_+^m \times \mathbb{R}^p$ *such that* $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ *satisfies the KKT conditions.*

Notes:

- The KKT conditions were first named after Harold W. Kuhn and Albert W. Tucker, two famous professors from Princeton, who published them in 1951.
- Later it was discovered that the conditions had been stated in the MSc thesis of the student William Karush in 1939.
- Many algorithms for convex optimization are conceived as methods for solving the KKT conditions.
- Sometimes it is possible to solve the KKT conditions (and thus the optimization problem) analytically.

5.3 Example

Consider the following optimization problem

$$\begin{array}{ll} \underset{\boldsymbol{x} \in \mathbb{R}^n}{\text{minimize}} & \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + c \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{array}$$

where $\boldsymbol{P} \in \mathbb{S}^n_+$, $\boldsymbol{q} \in \mathbb{R}^n$, $c \in \mathbb{R}$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and $\boldsymbol{b} \in \mathbb{R}^m$.

Question: What are the KKT conditions?