UIUC IE510 Applied Nonlinear Programming

Lecture 12: Duality

Ruoyu Sun

Review Question for Lecture 11b

- **Question 1**: Judge: For (smooth) inequality constrained problems, at a local-min the gradient of the objective and gradients of all constraints are always linearly dependent.

Proposal: False, since "active constraints"..
or "regular".

Answer: TRUE. If $\nabla f, \nabla g_1$ are linearly dependent, then $\nabla f, \nabla g_1, \nabla g_2$ are also linearly dependent (assume g_1 is active at x^*, g_2 is not active).

- **Question 2**: Judge: For (smooth) inequality constrained problems, at a regular local-min the Hessian of $L(x, \mu^*)$ is positive semi-definite on the null space defined by the gradients of all constraints ∇g_j's.

If "active constraints": TRUE
If just "constraints": also TRUE.

If $\nabla^2_{xx} L$ is PSD in $\{y : y \perp \nabla g_1\}$
then $\nabla^2_{xx} L$ is also PSD in $\{y : y \perp \nabla g_1, y \perp \nabla g_2\}$.

- **Question 3**: What is complementary slackness?

$\mu_i g_i(x) = 0, \ \forall i.$

Summary of Last Lecture

- Optimality condition for

  $$\min_x f(x), \text{ s.t. } h_i(x) = 0, i = 1, \ldots, m; g_j(x) \le 0, j = 1, \ldots, r.$$

  where f and h_i's, g_j's are cts-differentiable.

- Lagrangian function $L(x; \lambda, \mu) =$ $f(x) + \sum_i \lambda_i h_i(x) + \sum_j \mu_j g_j(x), \quad \mu \ge 0.$

- **KKT conditions:** at a "regular" local-min x^* (gradients of active constraints are linearly independent), we have

  - The gradient of the Lagrangian function is zero
    - ◇ $\nabla f(x^*)$ lies in span$\{\nabla h_i(x^*), \forall i; \nabla g_j(x^*), j \in \mathcal{A}(x^*)\}$, with coefficients λ_i^*'s, μ_j^*'s
    - ◇ $h_i(x^*) = 0, \forall i; g_j(x^*) = 0, \forall j \in \mathcal{A}(x^*)$
  - Complementary slackness:
  - $\nabla^2_{xx} L$ is positive semidefinite in the nullspace defined by the gradients of all active constraints

- If $\nabla^2_{xx} L$ is PD in that nullspace, and strict complentarity holds, then sufficient condition.

Summary of Last Lecture

- Optimality condition for

  $\min_x f(x)$, s.t. $h_i(x) = 0, i = 1, \ldots, m; g_j(x) \leq 0, j = 1, \ldots, r.$

  where f and h_i's, g_j's are cts-differentiable.

- Lagrangian function $L(x; \lambda, \mu) =$.

- **KKT conditions:** at a "regular" local-min x^* (gradients of active constraints are linearly independent), we have
  - The gradient of the Lagrangian function is zero
    - ◇ $\nabla f(x^*)$ lies in span$\{\nabla h_i(x^*), \forall i; \nabla g_j(x^*), j \in \mathcal{A}(x^*)\}$, with coefficients λ_i^*'s, μ_j^*'s
    - ◇ $h_i(x^*) = 0, \forall i; g_j(x^*) = 0, \forall j \in \mathcal{A}(x^*)$
  - Complementary slackness:
  - $\nabla_{xx}^2 L$ is positive semidefinite in the nullspace defined by the gradients of all active constraints

- If $\nabla_{xx}^2 L$ is PD in that nullspace, and strict complentarity holds, then sufficient condition.

This Lecture

- Today: duality
- After the lecture, you should be able to
  - Write down sufficient condition for a convex problem
  - Construct a dual problem of SVM
  - Distinguish "weak duality" and "strong duality"
- Advanced goals: derive the duality theory

Outline

Sufficient Condition for Convex Problems (4 slides)

Duality (8 slides)

Example: SVM (3 slides)

Convex Case?

Motivation: We learned that for convex function, $\nabla f(x) = 0$ is sufficient for global optimality.

Question: Theory for convex constrained problems?

Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \quad g_i(x) \leq 0, \; i = 1, ..., r. \end{array} \tag{1}$$

In Section I, $X = \mathbb{R}^n$; In Section II, X can be any set, or polyhedral.

For simplicity, we focus on inequality constraints in this lecture.

Convex Case: Sufficiency and Necessary Condition

Prop 12.1 (1st order condition is enough for convex problem)
Consider (1); suppose f and $g_j's$ are convex and $X = \mathbb{R}^n$.
Assuming x^* is regular. The following statements are equivalent:

- (i) x^* is a global minimum of (1).
- (ii) There exists μ^* such that

1st order condition
$$\nabla_x L(x^*, \mu^*) = 0, \tag{2a}$$
$$\mu_j^* \geq 0 \perp g_j(x^*) \leq 0, \quad j = 1, \ldots, r, \tag{2b}$$

- (iii) There exists μ^* such that

$$x^* = \text{argmin}_{x \in \mathbb{R}^n} L(x, \mu^*), \tag{3a}$$
$$\mu_j^* \geq 0 \perp g_j(x^*) \leq 0, \quad j = 1, \ldots, r, \tag{3b}$$

Proof:

- (i) $\Rightarrow$ (ii): KKT condition.
- (ii) $\Rightarrow$ (iii): $L = f + \sum_i \mu_i g_i$, $\mu_i \geq 0$, $\Rightarrow$ L is convex. $\nabla_x L = 0 \Rightarrow x^* = \arg\min$
- (iii) $\Rightarrow$ (i): Next few slides. (no need of convexity)

A General Sufficiency Condition

Proposition 12.2: Consider the problem (1). Define

$$L(x,\mu) = f(x) + \underbrace{\sum_{j=1}^{r} \mu_j g_j(x)}_{\langle \mu,\, g(x) \rangle},\ x \in X.$$

$$g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_r(x) \end{pmatrix}$$

If there exists μ^* such that

$$x^* = \text{argmin}_{x \in X} L(x,\mu^*), \tag{4a}$$

$$\mu_j^* \geq 0 \perp g_j(x^*) \leq 0, \quad j = 1, \ldots, r, \tag{4b}$$

then x^* is a global minimum of the problem (1).

Conditions $\rightarrow$ Conclusion

Bulid bridge between "conditions" & "conclusion".

Proof of Sufficiency Condition

- **Proof strategy**: See how far you get from conditions.
- First condition $x^* = \text{argmin}_{x\in X} L(x,\mu^*)$ implies

$$f(x^*) + \underbrace{\langle \mu^*, g(x^*)\rangle}_{(\mathrm{I})} = L(x^*,\mu^*) \le L(x,\mu^*) = f(x) + \underbrace{\langle \mu^*, g(x)\rangle}_{(\mathrm{II})}, \quad \forall\, x. \quad ①$$

- Suppose $g(x) = (g_1(x);\ldots;g_r(x)) \in \mathbb{R}^{r\times 1}$.

  Second condition $\mu_j^* \ge 0 \perp g_j(x^*) \le 0, \forall j$ implies

$$\underbrace{\langle \mu^*, g(x^*)\rangle}_{(\mathrm{I})} = 0 \underset{\substack{\text{since } \mu^*\ge 0,\\ g(x)\le 0}}{\ge} \underbrace{\langle \mu^*, g(x)\rangle}_{(\mathrm{II})}, \forall \underset{\text{meaning } g(x)\le 0}{\text{feasible}}\ x, \quad ②$$

- Combine both, get

By ①: $f(x^*) + (\mathrm{I}) \le f(x) + (\mathrm{II})$

By ②: $(\mathrm{I}) \ge (\mathrm{II})$

$\Rightarrow f(x^*) \le f(x), \ \forall$ feasible x.

$\square$

Proof (standard presentation) (reading)

Proof of Prop 12.2: We have

$$\begin{aligned} f(x^*) &= f(x^*) + \langle \mu^*, g(x^*) \rangle = \min_{x \in X} \{ f(x) + \langle \mu^*, g(x^*) \rangle \} \\ &\leq \min_{x \in X, g(x) \leq 0} \{ f(x) + \langle \mu^*, g(x^*) \rangle \} \\ &\leq \min_{x \in X, g(x) \leq 0} f(x), \end{aligned}$$

where the first equality follows from the hypothesis, which implies that $(\mu^*)' g(x^*) = 0$, and the last inequality follows from the nonnegativity of μ^*.
Q.E.D.

Outline

Duality

- What is one magical word that proves you are "optimization expert"?
  - Candidate: "dual"
- Two things are "dual" if their roles can be switched
- Duality theory is central to linear programming IE411.
- Duality theory also important for nonlinear constrained optimization
- It constructs dual problems, that
  - are easier to solve see this in SVM example
  - provide lower bound see this in "weak duality"
  - help design algorithms see this in SVM example

The Dual Problem

Consider a problem

$$\underset{x \in X, g_i(x) \leq 0, i=1,\ldots,r}{\overset{\text{black-box}}{\min}} f(x)$$

where f is cts-differentiable over $\mathbb{R}^n$.

fix $\mu \xrightarrow{\text{solve } \inf_x L(x,\mu)}$ value $\triangleq q(\mu)$.

- Define the dual function $q : \mathbb{R}^r \mapsto [-\infty, \infty)$

$$q(\mu) = \inf_{x \in X} L(x, \mu) = \inf_{x \in X}\{f(x) + \textstyle\sum_{j=1}^r \mu_j g_j(x)\}.$$

keep μ fixed.

  and the dual problem

$$\max_{\mu \geq 0} \quad q(\mu), \quad \text{i.e.,} \quad \max_{\mu \geq 0} \Big[\underbrace{\inf_{x \in X} L(x, \mu)}_{q(\mu)}\Big]$$

- If X is bounded, the dual function takes real values. In general, $q(\mu)$ can take the value $-\infty$. The effective constraint set of the dual is

$$Q = \{\mu \mid \mu \geq 0, q(\mu) > -\infty\}$$

Min-max and Max-min

Primal = min-max, dual = max-min.

Consider the min-max problem

$$\min_{x \in X} \max_{\mu \geq 0} L(x, \mu). \tag{5}$$

Claim: The original problem is equivalent to (5).

Proof: For simplicity, consider 1-constraint case.

(P1) $\min_{x \in \mathbb{R}^n} f(x)$, s.t. $g(x) \leq 0$. → Constrained

$L(x, \mu) = f(x) + \mu g(x)$, $\mu \geq 0$.

(P2) $\min_{x \in \mathbb{R}^n} \max_{\mu \geq 0} f(x) + \mu g(x)$. $\Longleftrightarrow$ (P1). → Unconstrained.

$\max_{\mu \geq 0} 2\mu = \infty$.

Fix $\bar{x}$. Consider $\max_{\mu \geq 0} \cancel{f(\bar{x})} + \mu \underbrace{g(\bar{x})}_{c}$. $= \begin{cases} \infty, & g(\bar{x}) > 0 \\ 0, & g(\bar{x}) = 0 \\ 0, & g(\bar{x}) < 0. \end{cases}$

$= \begin{cases} \infty, & g(\bar{x}) > 0 \\ f(\bar{x}), & g(\bar{x}) \leq 0. \end{cases}$

(P2) $\Leftrightarrow$ $\min_{\bar{x}} f(\bar{x})$, s.t. $g(\bar{x}) \leq 0$.

Weak Duality

What is the relation of the dual problem and the primal problem?

Prop 12.2 (Weak Duality) For any $\mu \geq 0$ and primal feasible x_F, we have (Universal)

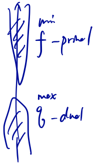
$\to g_i(x_F) \leq 0.$

$$q(\mu) \leq f(x_F).$$

$$q(\mu) = \inf_x L(x,\mu)$$

This implies

$$q^* = \sup_{\mu \geq 0} q(\mu) \leq \inf_{x_F \text{ is primal feasible}} f(x_F) = f^*.$$



Proof (2-line): $q^* = \sup_{\mu \geq 0} \underbrace{\inf_x L(x,\mu)} \leq \sup_{\mu \geq 0} L(x_F, \mu) = \sup_{\mu \geq 0} f(x_F) + \langle \underbrace{\mu}_{\geq 0}, \underbrace{g(x_F)}_{\leq 0 \text{ primal fea.}} \rangle$

$\leq f(x_F)$

$\Rightarrow q^* \leq f^*.$

Implication: min-max $\geq$ max-min.

When Does Strong Duality Hold?

Rewrite the proof: suppose x^* is a global-min, and $\mu \geq 0$, then

$$q(\mu) = \inf_{x \in X} L(x, \mu) \leq f(x^*) + \langle \mu, g(x^*) \rangle \leq f(x^*).$$

- Second inequality: $\exists \mu^*$ s.t. $\langle \mu^*, g(x^*) \rangle = 0$.
- First inequality: when f, g are convex and $X = \mathbb{R}^n$, and x^* is regular, then $x^* = \mathrm{argmin}_{x \in \mathbb{R}^n} L(x, \mu^*)$ holds.
  - Note: When f, g are non-convex, x^* is not necessarily the minimizer of $L(x, \mu^*)$

When Does Strong Duality Hold?

Rewrite the proof: suppose x^* is a global-min, and $\mu \geq 0$, then

$$q(\mu) = \inf_{x \in X} L(x, \mu) \underset{x^* = \arg\min_x L(x,\mu^*)}{\leq} f(x^*) + \langle \mu, g(x^*) \rangle \underset{\mu = \mu^*}{\leq} f(x^*).$$

- Second inequality: $\exists \mu^*$ s.t. $\langle \mu^*, g(x^*) \rangle = 0$. 3rd statement of Prop 12.1.
- First inequality: when f, g are convex and $X = \mathbb{R}^n$, and x^* is regular, then $x^* = \text{argmin}_{x \in \mathbb{R}^n} L(x, \mu^*)$ holds.
  - Note: When f, g are non-convex, x^* is not necessarily the minimizer of $L(x, \mu^*)$

Duality Gap and Strong Duality

f

duality gap.

q

- **Duality gap**: the difference $f^* - q^*$.
- We say that **strong duality** holds for (1) if $f^* = q^*$.
- When does strong duality hold?
  - $f, g_j's$ are convex and $X = \mathbb{R}^n$ and there exists regular global-min
  - f convex, $g_j's$ are linear, X is polyhedral (defined by linear equalities/inequalities)
  - Convex + (weak) Slater condition: $f, g_j's$ are convex, but X may not be $\mathbb{R}^n$.
    - ◇ Here, Slater condition means that there is a strict feasible point $\bar{x}$, i.e., $g_j(\bar{x}) < 0$ for all j and $\bar{x} \in X$.
    - ◇ "Weak" Slater condition removes the inequality requirement for affine constraints.

Duality Theorem

Prop 12.3 (Duality Theorem, for linearly constrained problems)
Suppose f is convex cts-differentiable and $X \subseteq \mathbb{R}^n$ is a polyhedral.
Consider the problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & x \in X, \quad a_i^T x \le b_i, \ i = 1, ..., r. \end{aligned} \tag{6}$$

(a) (strong duality) If the primal problem has an optimal solution, the dual problem also has an optimal solution and the optimal values are equal.

$f^* = q^*$

(b) $x^* \in X$ is primal-optimal and μ^* is dual-optimal if and only if

$$\left.\begin{aligned} \mu_j^* \ge 0 \perp a_j^T x^* - b_j \le 0, \forall j; \\ x^* = \operatorname{argmin}_{x \in X} L(x, \mu^*). \end{aligned}\right\} \text{3rd statement in Prop 12.1.}$$

Linear Equality Constraints

- Linear equality constraints are treated similar to inequality constraints, except that the sign of the Lagrange multipliers is unrestricted:

$$\text{Primal}: \quad \min_{x \in X,\, e_i' x = d_i,\, i=1,\ldots,m,\, a_j' x \le b_j,\, j=1,\ldots,r} f(x)$$

$$\text{Dual}: \quad \max_{\underbrace{\lambda \in \mathbb{R}^m}_{\text{equality}},\, \underbrace{\mu \ge 0}_{\text{inequality}}} \inf_{x \in X} L(x, \lambda, \mu).$$

Outline

Example: Support Vector Machine

- **Training data set**: two classes $\{x_i, y_i\}_{i=1}^{n}$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1,\ 1\}$.
- Suppose the training data are linearly separable
- **Objective**: find a hyperplane to separate the data points, i.e., find w such that $x_i^T w \geq 1, \forall i$.
- Formulation:

$$\min_w \frac{1}{2} \underbrace{\|w\|^2}_{\text{quadratic}}, \quad \text{s.t. } \underbrace{y_i w^T x_i}_{\text{linearly}} \geq 1, i = 1, \dots, n.$$

- Issue: Can we use SGD to solve it? Pick one sample/constraint, and then what?

SVM: Derive Dual

Write dual problem:

- First, let $f(w) = \frac{1}{2}\|w\|^2$, $g_i(w) = 1 - y_i x_i^T w$, and

$$L(w, \mu) = \frac{1}{2}\|w\|^2 + \sum_i \mu_i \cdot (1 - y_i x_i^T w), \quad \mu \geq 0.$$

- Second, write the dual problem

$$\frac{\partial L}{\partial w} = w + \sum_i \mu_i \cdot (-y_i x_i) = 0$$

$$\Rightarrow w = \sum_i \mu_i x_i y_i.$$

$$q(\mu) = \inf_w L(w, \mu). = \inf_w L(w, \mu). \text{ quadratic in } w.$$

  Solve the optimal $w^*(\mu) = \sum_{i=1}^{n} \mu_i x_i y_i$.

  – Representer theorem in ML
  RKHS
  optimal separator is linear combination of x_i's.

- Third, plugging in the expression of $w^*(\mu)$ into q, and ~~minimize~~ maximize q:

$$\max_{u \geq 0} q(\mu) = -\frac{1}{2}\|\sum_i \mu_i x_i y_i\|^2 + \sum_i \mu_i.$$

SVM: Solve Dual Problem

Dual problem of SVM

$$\max_{\mu \geq 0} \sum_i \mu_i - \frac{1}{2} \| \sum_i \mu_i x_i y_i \|^2.$$

What is the nature of this problem? quadratic-min over $\mathbb{R}_+^n$.
(similar to nonnegative least squares)

How to solve it?

- Gradient projection
- (Dual) coordinate ascent: method of choice in libsvm
- Dual CD uses one sample at a time! (like SGD)

- Side remark: in ML, people use Kernal trick, i.e., write $\| \sum_i \mu_i x_i y_i \|^2$ as $\sum_{i,j} \mu_i \mu_j y_i y_j \langle x_i, x_j \rangle$

SVM: Solve Dual Problem

Dual problem of SVM

$$\max_{\mu \geq 0} \sum_i \mu_i - \frac{1}{2} \| \sum_i \mu_i x_i y_i \|^2.$$

What is the nature of this problem?

How to solve it?

- Gradient projection
- (Dual) coordinate ascent: method of choice in libsvm
- Dual CD uses one sample at a time! (like SGD)

- Side remark: in ML, people use Kernal trick, i.e., write $\| \sum_i \mu_i x_i y_i \|^2$ as $\sum_{i,j} \mu_i \mu_j y_i y_j \langle x_i, x_j \rangle$

SVM: Solve Dual Problem

Dual problem of SVM

$$\max_{\mu \geq 0} \sum_i \mu_i - \frac{1}{2} \| \sum_i \mu_i x_i y_i \|^2.$$

What is the nature of this problem?

How to solve it?

- Gradient projection
- (Dual) coordinate ascent: method of choice in libsvm
- Dual CD uses one sample at a time! (like SGD)
- Side remark: in ML, people use Kernal trick, i.e., write $\| \sum_i \mu_i x_i y_i \|^2$ as $\sum_{i,j} \mu_i \mu_j y_i y_j \langle x_i, x_j \rangle$

Summary

In this lecture, we learned the following (think yourself before reading):

- Sufficient and necessary conditions for convex constrained problems
- Duality: dual problem, weak/strong duality
- Application of duality theory to solve SVM

Summary

In this lecture, we learned the following (think yourself before reading):

- Sufficient and necessary conditions for convex constrained problems
- Duality: dual problem, weak/strong duality
- Application of duality theory to solve SVM