Large Scale Optimization for Machine Learning: Lecture 10

Lecturer: Meisam Razaviyayn Scribe: Ziyu He

Sept 26, 2017

1 Geometric Interpretation

Assume that we are working with the following problem in this section:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \quad \forall i = 1, 2, \ldots, m. \\ & \mathbf{x} \in \mathcal{X} \end{aligned}$$

1.1 Some Setup

Here we define a set $\mathcal{A}$ as follow

$$\mathcal{A} = \left\{(\mathbf{u}, t) \in \mathbb{R}^{m+1} | \exists \mathbf{x} \in \mathcal{X} \text{ with } f_i(\mathbf{x}) \leq u_i, f_0(\mathbf{x}) \leq t\right\}$$

Intuitively this set can be treated as some kind of epigraph of set $\mathcal{G} \triangleq \{(\mathbf{u}, t) \in \mathbb{R}^{m+1} | f_i(\mathbf{x}) = u_i, i = 1...m, f_0(\mathbf{x}) = t, \mathbf{x} \in \mathcal{X}\}$. Based on the definition of $\mathcal{A}$, we can reformulate the primal optimal value p^* and dual function hence the dual problem. First, the primal problem can be rewritten as the following equivalent problem:

$$\begin{aligned} p^* = \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \ \forall i. \\ & \mathbf{x} \in \mathcal{X} \end{aligned} \iff \begin{aligned} p^* = \min_{\mathbf{x}, \mathbf{u}, t} \quad & t \\ \text{s.t.} \quad & f_0(\mathbf{x}) \leq t. \\ & f_i(\mathbf{x}) \leq u_i, \forall i. \\ & \mathbf{u} \preceq \mathbf{0}, \ \mathbf{x} \in \mathcal{X}. \end{aligned}$$

We can easily justify that the above two problems are equivalent in terms of having the same optimal cost, by assuming an optimal solution for one problem and showing that we can construct an optimal solution of the other problem with the same cost. By definition the equivalent problem can be reformulated as:

$$\begin{aligned} p^* = \min_{\mathbf{u},t} \quad & t \\ \text{s.t.} \quad & (\mathbf{u}, t) \in \mathcal{A}. \\ & \mathbf{u} \preceq \mathbf{0}. \end{aligned}$$

Similarly we can also reformulate the dual function $g(\boldsymbol{\lambda})$ by geometric interpretation:

$$\begin{aligned} & g(\boldsymbol{\lambda}) = \inf_{\mathbf{x}\in\mathcal{X}} \left\{ f_0(\mathbf{x}) + \textstyle\sum_{i=1}^m \lambda_i f_i(\mathbf{x}) \right\} \\ \iff & g(\boldsymbol{\lambda}) = \inf_{\mathbf{x},\mathbf{u},t} \left\{ \boldsymbol{\lambda}^T \mathbf{u} + t | f_0(\mathbf{x}) \le t,\ f_i(\mathbf{x}) \le u_i, \forall i,\ \mathbf{x} \in \mathcal{X} \right\} \end{aligned}$$

By definition, the second expression can be equivalently written as follow:

$$g(\boldsymbol{\lambda}) = \inf \left\{ \begin{bmatrix} \boldsymbol{\lambda} \\ 1 \end{bmatrix}^T \begin{bmatrix} \mathbf{u} \\ t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{u} \\ t \end{bmatrix} \in \mathcal{A} \right\}$$

With this formulation, for a specific $\boldsymbol{\lambda}$, we have:

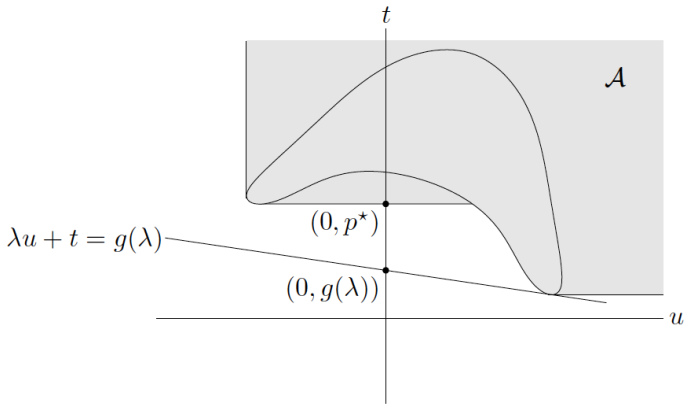
$$(\boldsymbol{\lambda}, 1)^T (\mathbf{u}, t) \ge g(\boldsymbol{\lambda}),\ \forall (\mathbf{u}, t) \in \mathcal{A}$$

This actually defines a halfspace in $\mathbb{R}^{m+1}$. To be more accurate, we define a supporting hyperplane of set $\mathcal{A}$ with normal vector $(\boldsymbol{\lambda}, 1)$ and offset $g(\boldsymbol{\lambda})$ (i.e., $\mathcal{L}_{\boldsymbol{\lambda}} \triangleq \{(\mathbf{u}, t) | (\boldsymbol{\lambda}, 1)^T (\mathbf{u}, t) \ge g(\boldsymbol{\lambda})\}$). Note that the last entry of normal vector $(\boldsymbol{\lambda}, 1)$ is $1 \neq 0$, hence such hyperplane is *non-vertical*.

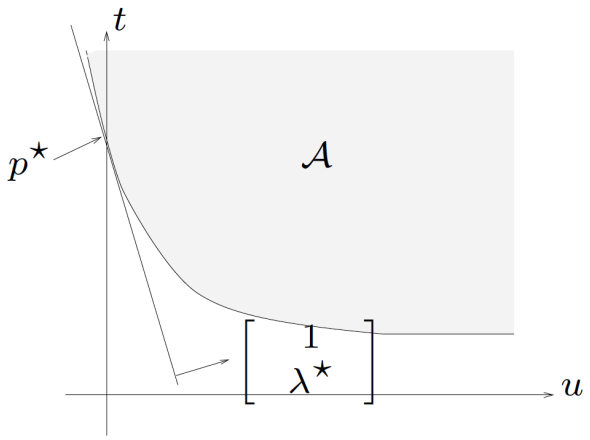
We can easily see that when $\mathbf{u} = \mathbf{0}$ we have $\mathcal{L}_{\boldsymbol{\lambda}}$ intersect with $t-$axis at $(\mathbf{0}, g(\boldsymbol{\lambda}))$. Thus the process of solving the dual problem can be equally represented as moving $\mathcal{L}_{\boldsymbol{\lambda}}$ in the space by changing $\boldsymbol{\lambda} \succeq \mathbf{0}$ and find the maximum "intercept" on $t-$axis.

1.2 Strong Duality and Slater's Condition

With the tools we have constructed so far, we have the following example visualizing weak duality (Fig 5.5 in Chapter 5 of [1]):



This figure is for the case when we only have one inequality constraint in our primal problem. The boomerang-shape set is the set $\mathcal{G}$ and the shaded area is the set $\mathcal{A}$. For any particular $\lambda \geq 0$ we can define a *non-vertical* supporting hyperplane for set $\mathcal{A}$ with intercept on $t-$axis as $g(\lambda)$. Though absent in this figure, we may as well define d^* (i.e., the dual optimal value) as the largest intercept on $t-$axis that we can find in this case. We have $d^* < p^*$ which means strong duality does not hold in this case.



The above figure shows the case when our primal problem is convex (hence our set $\mathcal{A}$ is convex) with one inequality constraint. In this particular case, we can define a supporting hyperplane of $\mathcal{A}$ for some $(\lambda^*, 1)$ s.t. it intersect with $t-$axis by $(0, p^*)$ and this is the largest intercept we can actually achieve (hence this indicates strong duality $d^* = p^*$).

We can also use this figure to intuitively show why Slater's condition ensures strong duality. Note that the definition of Slater's condition can be rewritten as $\exists(\mathbf{u}, t) \in \mathcal{A}$ *with* $\mathbf{u} \prec \mathbf{0}$. In this particular case, Slater's condition, along with convexity of $\mathcal{A}$, means that we have $(0, p^*)$ as the intersection of $\mathcal{A}$'s boundary and $t-$axis. Since $\mathcal{A}$ convex, we can define a *non-vertical* hyperplane at $(0, p^*)$ hence we have strong duality (this is exactly what we have observed from the previous figure).

A proof of strong duality under Slater's condition which formalizes the idea shown above with separating hyperplane can be found in 5.3.2 of [1].

2 Sensitivity Analysis and Duality

2.1 $p^*(\mathbf{u})$ and Its Lower Bound

Consider a perturbed version of our primal problem:

$$\begin{aligned} p^*(\mathbf{u}) = \min_{\mathbf{u},t} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq u_i, i = 1...m \end{aligned}$$

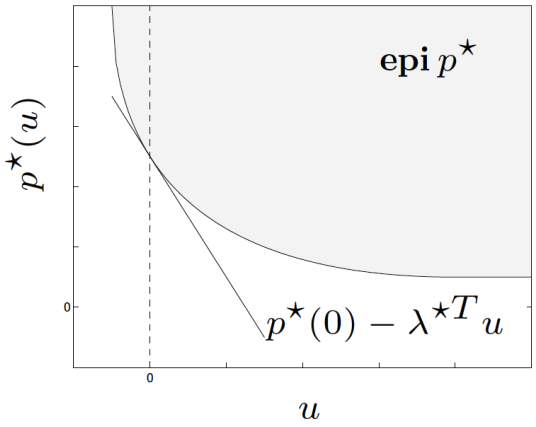
Here, for any specific $\mathbf{u} \in \mathbb{R}^m$ we have a specific perturbed problem and its corresponding optimal cost. Hence we have a function $p^*(\mathbf{u}) : \mathbb{R}^m \to \mathbb{R}$. An interesting fact is that the epigraph of function $p^*(\mathbf{u})$ is just the set $\mathcal{A}$ we defined in the last section.

First, under strong duality, $p^*(\mathbf{u})$ has a lower bound:

$$p^*(\mathbf{u}) \geq p^*(\mathbf{0}) - \boldsymbol{\lambda}^{*T}\mathbf{u}$$

Where $\boldsymbol{\lambda}^*$ is the dual optimal solution and $p^*(\mathbf{0})$ is the original primal optimal cost.

The following figures shows a "pictorial proof" of this bound.



2.2 Sensitivity Analysis

For a particular constraint i, λ_i^* is its corresponding dual optimal solution and u_i is its "perturbation". Note that since $\boldsymbol{\lambda}^* \succeq \mathbf{0}$ we always have $\lambda_i^* \geq 0$. We can interpret $u_i > 0$ as "relaxing" constraint i while $u_i < 0$ as "tightening" constraint i.

When λ_i^* is very large, if we tighten constraint i ($u_i < 0$) a little bit, we will greatly increase the lower bound of $p^*(\mathbf{u})$ hence $p^*(\mathbf{u})$.

If we further have $p^*(\mathbf{u})$ differentiable at $\mathbf{0}$, then we can see from the above figure that:

$$\lambda_i^* = -\frac{\partial p^*(\mathbf{0})}{\partial u_i}$$

In other words, if we have $p^*(\mathbf{u})$ differentiable at $\mathbf{0}$ and strong duality holds, λ_i^* can be seen as the local sensitivities of the optimal value w.r.t. u_i (i.e., some perturbation on constraint i) [1]. Proof of this conclusion can be found in 5.6.3 of [1].

3 Duality with Equality Constraints

So far we have only considered the primal problem with inequality constraints. In this section, we will complete our introduction of Lagrangian du-

ality when our primal problem has both inequality and equality constraints:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \quad \forall i = 1, 2, \ldots, m. \\ & h_i(\mathbf{x}) = 0, \quad \forall i = 1, 2, , \ldots, p. \\ & \mathbf{x} \in \mathcal{X} \end{aligned}$$

3.1 Primal and Dual Problem with Equality Constraints

Here we show the counterparts for Lagrange function, dual function and dual problem when we consider primal problem with equality constraints. Lagrangian function ($L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$):

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \triangleq f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x})$$

Dual function ($g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$):

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \triangleq \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

The dual function is concave regardless of primal convexity. For any $\boldsymbol{\lambda} \succeq \mathbf{0}$ and primal feasible $\mathbf{x}$ we have $g(\lambda, \mu) \leq f_0(\mathbf{x}) \leq p^*$, so to find the best lower bound of primal optimal cost p^*, we have the following dual problem:

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{s.t.} \quad & \boldsymbol{\lambda} \succeq \mathbf{0} \end{aligned}$$

The dual problem is convex regardless of primal convexity. Denote d^* as the dual optimal cost, we have weak duality $d^* \leq p^*$. If $d^* = p^*$, then strong duality holds.

Similar to what we have introduced, if our primal problem is convex, with some special *constraint qualifications* conditions we can conclude strong duality. Slater's condition for primal problems with equality constraints should be changed as:

Definition 1 *Slater's condition:*

$\exists \mathbf{x} \in \text{relint}(\mathcal{D})$ *with* $f_i(\mathbf{x}) < 0, i = 1, \ldots, m, h_i(\mathbf{x}) = 0, j = 1, \ldots, p$

As usual, we do not need to check the affine inequality constraints for the "interior part".

3.2 KKT Optimality Conditions

From strong duality, we can conclude complementary slackness and primal optimal $\mathbf{x}^*$ being minimizer of $\inf_{\mathbf{x}\in\mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$. Similar to the cases when we only consider inequality constraints for primal problems, here we can conclude KKT conditions as necessary optimality conditions for general cases and KKT condition as sufficient optimality conditions for convex primal problems. Here we will present the KKT conditions for general cases without derivation.

Theorem 1 *Assume f_i's are differentiable and $\mathcal{X} = \mathbb{R}^n$ and $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ optimal with zero duality gap. Then,*

1. $f_i(\mathbf{x}^*) \leq 0,\ h_j(\mathbf{x}^*) = 0, \forall ij \quad \Leftarrow$ *Primal Feasibility*
2. $\boldsymbol{\lambda}^* \succeq 0 \quad \forall i \quad \Leftarrow$ *Dual Feasibility*
3. $\lambda_i^* f_i(\mathbf{x}^*) = 0 \quad \forall i \quad \Leftarrow$ *Complementary Slackness*
4. $\nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_j \mu_j^* \nabla h_j(\mathbf{x}^*) = 0$
   $\Leftarrow \mathbf{x}^*$ *minimizer for* $L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$

Where $i = 1...m, j = 1...p$. Conversely, if the problem is convex and primal/dual feasible $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\mu}})$ satisfy KKT, then $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\mu}})$ are primal/dual optimal.

Example

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{x}\|^2 \\ s.t. \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned}$$

- Assume matrix $\mathbf{A}$ is fat and $\mathbf{A}\mathbf{A}^T$ is invertible
- Lagrangian function: $L(\mathbf{x}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{x}\|^2 + \langle \boldsymbol{\mu}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle$
- Dual function: $g(\boldsymbol{\mu}) = \inf_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}\|^2 + \langle \boldsymbol{\mu}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle$

  Checking first order optimality condition implies that the minimizer for

$L(\mathbf{x}, \boldsymbol{\mu})$ should be $-\mathbf{A}^T\boldsymbol{\mu}$

Then $g(\boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{A}^T\boldsymbol{\mu}\|^2 + \langle \boldsymbol{\mu}, -\mathbf{A}\ \mathbf{A}^T\boldsymbol{\mu} - \mathbf{b}\rangle = -\frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\mathbf{A}^T\boldsymbol{\mu} - \mathbf{b}^T\boldsymbol{\mu}$

• Dual problem: $\max_{\boldsymbol{\mu}} -\frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\mathbf{A}^T\boldsymbol{\mu} - \mathbf{b}^T\boldsymbol{\mu}$

By first order optimality condition, $\mathbf{A}\mathbf{A}^T\boldsymbol{\mu}^* + \mathbf{b} = \mathbf{0} \Rightarrow \boldsymbol{\mu}^* = -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b} \Rightarrow \mathbf{x}^* = -\mathbf{A}^T\boldsymbol{\mu} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}$

4 Different formulations can lead to different duals

Consider the optimization problem:

$$\min_{\mathbf{x}} \quad \log\Big(\sum_{i=1}^m \exp(\mathbf{a}_i^T\mathbf{x} - b_i)\Big)$$

log-sum-exp is known to be convex with

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \log\Big(\sum_{i=1}^m \exp(\mathbf{a}_i^T\mathbf{x} - b_i)\Big)$$

The dual function is the constant function $g(\boldsymbol{\lambda}) = p^*$. Therefore, the dual problem can be written as

$$\max_{\boldsymbol{\lambda}} \quad p^*,$$

which has a constant objective and not very helpful. Now consider a different reformulation of the original problem:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & \log\Big(\sum_{i=1}^m \exp(y_i)\Big) \\ \text{s.t.} \quad & y_i = \mathbf{a}_i^T\mathbf{x} - b_i, \quad \forall i \end{aligned}$$

This is also a convex problem and we can write the Lagrangian function as

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \log\Big(\sum_{i=1}^m \exp(y_i)\Big) + \langle \boldsymbol{\mu}, \mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{y}\rangle$$

and the dual function is

$$g(\boldsymbol{\mu}) = \inf_{\mathbf{x},\mathbf{y}} \quad \log\Big(\sum_{i=1}^{m}\exp(y_i)\Big) + \langle \boldsymbol{\mu}, \mathbf{A}\mathbf{x}\rangle - \langle \boldsymbol{\mu}, \mathbf{b}\rangle - \langle \boldsymbol{\mu}, \mathbf{y}\rangle$$
$$= \begin{cases} \inf\limits_{\mathbf{y}} \quad \log\Big(\sum_{i=1}^{m}\exp(y_i)\Big) - \langle \boldsymbol{\mu}, \mathbf{b}\rangle - \langle \boldsymbol{\mu}, \mathbf{y}\rangle & \text{if } \mathbf{A}^T\boldsymbol{\mu} = \mathbf{0} \\ -\infty & \text{Otherwise} \end{cases}$$

Now consider the problem:

$$\inf_{\mathbf{y}} \quad \log\Big(\sum_{i=1}^{m}\exp(y_i)\Big) - \langle \boldsymbol{\mu}, \mathbf{b}\rangle - \langle \boldsymbol{\mu}, \mathbf{y}\rangle \quad \text{when } \mathbf{A}^T\boldsymbol{\mu} = \mathbf{0}$$

First order optimality condition implies $\Rightarrow \frac{\exp(y_j^*)}{\sum_j \exp(y_j^*)} - \mu_j = 0,\ \forall j$, which is solvable if $\mu_j \geq 0$ and $\sum_j \mu_j = 1$. In fact, one can show that if $\mu_j < 0$ or $\sum_j \mu_j \neq 1$, then the above optimization problem is unbounded from below. Let $\sum_j \exp(y_j^*) = \alpha \Rightarrow y_j^* = \log(\alpha\mu_j)$ then we will get $g(\boldsymbol{\mu}) =:$

$$\begin{cases} \log(\alpha) - \langle \boldsymbol{\mu}, \mathbf{b}\rangle - \sum_j \mu_j \log(\alpha\mu_j) & \text{if } \mathbf{A}^T\boldsymbol{\mu} = \mathbf{0};\ \mu_j \geq 0 \forall j;\ \sum_j \mu_j = 1 \\ -\infty & \text{Otherwise} \end{cases}$$

Since $\sum_j \mu_j = 1$ in the first row, we can simplify the above equation as

$$g(\boldsymbol{\mu}) = \begin{cases} -\boldsymbol{\mu}^T\mathbf{b} - \sum_j \mu_j \log(\mu_j) & \text{if } \mathbf{A}^T\boldsymbol{\mu} = \mathbf{0};\ \mu_j \geq 0 \forall j;\ \sum_j \mu_j = 1 \\ -\infty & \text{Otherwise} \end{cases}$$

Thus we have the dual problem:

$$\begin{aligned} \min_{\boldsymbol{\mu}} \quad & -\boldsymbol{\mu}^T\mathbf{b} - \sum_j \mu_j \log(\mu_j) \\ \text{s.t.} \quad & \textstyle\sum_j \mu_j = 1, \quad \mu_j \geq 0 \forall j \\ & \mathbf{A}^T\boldsymbol{\mu} = \mathbf{0} \end{aligned}$$

From this case we can see that even if two primal problems are equivalent, their dual problems might not necessarily be equivalent.

References

[1] S. Boyd, and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.