Week 5-6

Nicolo Michelusi

I. Lagrange duality and applications

- Duality:

  - convert a convex problem (primal problem) with a set of variables (primal variables) to another convex problem (dual problem) with another set of variables (dual variables)

  - Under mild conditions, the two problems are equivalent!

- Benefits of duality:

  - Reveal structure of the optimal solution (KKT conditions)

  - The dual variables have engineering interpretations; they can be interpreted as the price of the constraints

  - The original problem may be decomposed into sub-problems, when we convert to the dual, leading to distributed solutions: each user can minimize a sub-problem from the Lagrangian independently, as long as the correct "prices" are provided.

  - The constraint set of the dual problem is often simpler (easy to compute projection)

- Primal problem (for now, let us consider also non-convex problems)

$$\min f_0(x)$$

$$\text{s.t. } f_i(x) \leq 0,\ i = 1, \ldots, m$$

$$h_i(x) = 0,\ i = 1, \ldots, p$$

- Lagrangian function

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),\ \ \lambda_i \geq 0, \forall i$$

  - λ can be interpreted as a "price" for violating the inequality constraints; similar interpretation for ν

- Minimize Lagrangian over primal variables x

$$g(\lambda, \nu) = \min_x L(x, \lambda, \nu)$$

$$= \min_x f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x), \;\; \lambda_i \geq 0, \forall i$$

- note: the constraints are eliminated!

- if the problem is convex, $L(x, \lambda, \nu)$ is convex wrt x and the necessary and sufficient optimality condition becomes (case when functions differentiable)

$$\nabla_x L(x, \lambda, \nu) = 0$$

or equivalently (assume equality constraint $Ax = b$ for convex problem)

$$\nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + A^T \nu = 0$$

- Interpretation (valid also for non-convex problems): assume $x^* = \arg\min L(x, \lambda, \nu)$ and let $f_i^* = f_i(x^*)$ and $h_i^* = h_i(x^*)$. Then, x^* is the global optimizer of the following problem:

$$\min f_0(x)$$
$$\text{s.t. } f_i(x) \leq f_i^*, \;\; i = 1, \ldots, m$$
$$h_i(x) = h_i^*, \;\; i = 1, \ldots, p$$

In fact,

$$f_0(x^*) = L(x^*, \lambda, \nu) - \sum_{i=1}^m \lambda_i f_i^* - \sum_{i=1}^p \nu_i h_i^* \leq f_0(x) + \sum_{i=1}^m \lambda_i (f_i(x) - f_i^*) + \sum_{i=1}^p \nu_i (h_i(x) - h_i^*) \leq f_0(x),$$

for every feasible x such that $f_i(x) \leq f_i^*$ and $h_i(x) = h_i^*$.

- Roughly speaking, we can then tune the prices λ and ν to obtain the desired trade-off.

- Properties of dual function $g(\lambda, \nu)$ (even when the problem is not convex):

  - $g(\lambda, \nu)$ is a concave function of (λ, ν): $L(x, \lambda, \nu)$ is linear in (λ, ν) (hence concave) and the minimization over concave functions is concave!

  - Let $x \in \mathcal{F}$, i.e. $f_i(x) \leq 0$ and $h_i(x) = 0, \;\forall i$. Then

$$g(\lambda, \nu) \leq L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) \leq f_0(x), \;\; \forall x \in \mathcal{F}$$

It follows that

$$g(\lambda, \nu) \leq \min_{x \in \mathcal{F}} f_0(x) = p^*$$

and $g(\lambda, \nu)$ is a lower bound to the optimal value!

  - It makes then sense to find the tightest lower bound:

$$\max_{\lambda \geq 0, \nu} g(\lambda, \nu) \leq \min_{x \in \mathcal{F}} f_0(x) \text{ (weak duality)}$$

- Dual problem:

$$\max g(\lambda,\nu)$$
$$\text{s.t. } \lambda \geq 0$$

This is a convex optimization problem since g is concave and constraints are convex.

- We have investigated this problem previously: the optimality conditions are (for simplicity now, assume g differentiable)

$$\nabla_\lambda g(\lambda,\nu) \leq 0$$
$$\lambda \geq 0$$
$$\lambda_i \nabla_{\lambda_i} g(\lambda,\nu) = 0$$
$$\nabla_\nu g(\lambda,\nu) = 0$$

This can be simplified by noticing that $\nabla_{\lambda_i} g(\lambda,\nu) = f_i(x_{\lambda,\nu})$ (not a rigorous argument, but it helps to develop intuition), where

$$x_{\lambda,\nu} = \arg\min_x L(x,\lambda,\nu)$$

To see this, let us compute the directional derivative of $g(\lambda,\nu)$ along the direction $(\lambda,\nu)+t(d,v),\ t>0$:

$$g(\lambda+td,\nu+tv) \leq L(x_{\lambda,\nu},\lambda+td,\nu+tv) = L(x_{\lambda,\nu},\lambda,\nu)+t\sum_{i=1}^m d_i f_i(x_{\lambda,\nu})+t\sum_{i=1}^p v_i h_i(x_{\lambda,\nu})$$

$$= g(\lambda,\nu)+t\sum_{i=1}^m d_i f_i(x_{\lambda,\nu})+t\sum_{i=1}^m v_i h_i(x_{\lambda,\nu})$$

hence

$$\lim_{t\to 0^+}\frac{g(\lambda+td,\nu+tv)-g(\lambda,\nu)}{t} \leq \sum_{i=1}^m d_i f_i(x_{\lambda,\nu})+\sum_{i=1}^m v_i h_i(x_{\lambda,\nu}).$$

Similarly,

$$g(\lambda,\nu) \leq L(x_{\lambda+td,\nu+tv},\lambda,\nu) = g(\lambda+td,\nu+tv)$$
$$-t\sum_{i=1}^m d_i f_i(x_{\lambda+td,\nu+tv})-t\sum_{i=1}^m v_i h_i(x_{\lambda+td,\nu+tv})$$

hence

$$\lim_{t\to 0^+}\frac{g(\lambda+td,\nu+tv)-g(\lambda,\nu)}{t} \geq \sum_{i=1}^m d_i f_i(x_{\lambda+td,\nu+tv})+\sum_{i=1}^m v_i h_i(x_{\lambda+td,\nu+tv})$$

Combining the two inequalities, and assuming that $f_i(x_{\lambda+td,\nu+tv})$ and $h_i(x_{\lambda+td,\nu+tv})$ are continuous in t, it follows that

$$\lim_{t\to 0}\frac{g(\lambda+td,\nu+tv)-g(\lambda,\nu)}{t}=\nabla_\lambda g(\lambda,\nu)^T d+\nabla_\nu g(\lambda,\nu)^T v=\sum_{i=1}^{m} d_i f_i(x_{\lambda,\nu})+\sum_{i=1}^{m} v_i h_i(x_{\lambda,\nu})$$

or equivalently

$$\nabla_{\lambda_i} g(\lambda,\nu)=f_i(x_{\lambda,\nu})$$

$$\nabla_{\nu_i} g(\lambda,\nu)=h_i(x_{\lambda,\nu})$$

- By combining together the optimality conditions of the Lagrangian and dual problems, we obtain a set of Karush-Kuhn-Tucker (KKT) conditions:

  $x^* = \arg\min_{x\in D} f_0(x)+\sum_{i=1}^{m}\lambda_i f_i(x)+\sum_{i=1}^{p}\nu_i h_i(x)$ (Lagrangian optimality)

  $f_i(x^*)\leq 0,\ \forall i$ (primal feasibility, comes from $\nabla_\lambda g(\lambda,\nu)\leq 0$)

  $h_i(x^*)=0,\ \forall i$ (primal feasibility, comes from $\nabla_{\nu_i} g(\lambda,\nu)=0$)

  $\lambda\geq 0$ (dual feasibility)

  $\lambda_i f_i(x^*)=0,\forall i$ (complementary slackness conditions, comes from $\lambda_i\nabla_{\lambda_i} g(\lambda,\nu)=0$)

- In general, for non-convex problems, weak duality holds $g(\lambda,\nu) \leq \min_{x\in\mathcal{F}} f_0(x) = p^*$, but not strong duality $g(\lambda,\nu) = \min_{x\in\mathcal{F}} f_0(x) = p^*$.

  Example: $\min \ln(x)$ s.t. $x \geq 2$; optimal value is $\ln 2$, but dual function is

$$g(\lambda) = \min_{x>0} \ln(x) - \lambda(x-2) = -\infty, \ \forall\lambda \geq 0$$

  However, for convex problems, under mild conditions (more later), strong duality holds. In the above example, by using the change of variables $x = e^y$ we obtain $\min y$ s.t. $y \geq \ln 2$, and the dual function is

$$g(\lambda) = \min y - \lambda(y - \ln 2) = \begin{cases} \ln 2 & \lambda = 1 \\ -\infty & \lambda \neq 1 \end{cases}$$

  and the dual problem is $\max_{\lambda\geq 0} g(\lambda) = \ln 2$, so that strong duality holds.

- In general (convex AND non-convex), KKT conditions are equivalent to global optimality and strong duality:

  - **Theorem**: (x^*, λ^*, ν^*) satisfy KKT conditions $\Leftrightarrow$ x^* is globally optimal and strong duality holds. (valid also for non-convex problems)

  $\Rightarrow$ To see this, assume (x^*, λ^*, ν^*) satisfy all KKT conditions. Then, since $h_i(x^*) = 0$,

$$g(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*).$$

  From complementary slackness conditions, we also have that $\lambda_i^* f_i(x^*), \forall i$, yielding

$$g(\lambda^*, \nu^*) = f_0(x^*)$$

  Since $d^* = g(\lambda^*, \nu^*) = f(x^*) \leq p^* = \min_{x\in\mathcal{F}} f_0(x)$, and x^* is feasible, it must follow that x^* is globally optimal, with optimal value p^*, and strong duality holds.

  $\Leftarrow$ assume that x^* is globally optimal and strong duality holds with $\lambda^* \geq 0$ and ν^* (dual feasibility). Then

$$f(x^*) = g(\lambda^*, \nu^*) \leq L(x^*, \lambda^*, \nu^*) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i f_i(x^*) + \sum_{i=1}^{p} \nu_i h_i(x^*) \leq f_0(x^*)$$

  since x^* is feasible (primal feasibility: $f_i(x^*) \leq 0$, $h_i(x^*) = 0$) and $\lambda^* \geq 0$. Then it follows that all inequalities are, instead, equalities. As a result, we must have

$$\lambda_i f_i(x^*) = 0, \ \forall i \text{ (complementary slackness)}$$

  and

$$g(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*),$$

so that x^* minimizes the Lagrangian.

All KKT conditions are thus satisfied.

- If the problem is convex, under mild assumption, a stronger result holds

  - **Theorem**: if the problem is convex, Slater condition is satisfied, and rank$(A) = p$ (# of equality constraints) then strong duality holds.

  - Slater condition: there exists $x \in D$ such that

$$f_i(x) < 0, \ \forall i = 1, \ldots, m$$

$$Ax = b$$

  In other words, there exist a feasible x which satisfies all inequality constraint strictly.

  - This is a very mild assumption, valid in many applications

  - rank$(A) = p$ is not difficult to attain and is not restrictive

  - Proof via graphical interpretation

- Weak and strong duality: we have seen that

  - $g(\lambda, \nu)$ is concave

  - $d^* = \max_{\lambda \geq 0, \nu} g(\lambda, \nu) \leq p^* = \min_{x \in \mathcal{F}} f_0(x)$ (weak duality)

  - $p^* - d^*$ is the "duality gap"

  - If the problem is convex, under mild assumptions (Slater condition), the duality gap is zero (strong duality)

- Example: rate control

$$\min_x \sum_s U_s(x_s)$$

$$\text{s.t. } \sum_s H_{s,l} x_s \leq R_l, \ \forall l$$

need $R_l > 0, \ \forall l$ for Slater condition to hold.

- What if $R_l = 0$ for some l?

- Graphical interpretation (Boyd p.234):

  - Let

$$G = \{(f_1(x), \dots, f_m(x), h_1(x), \dots, h_p(x), f_0(x))|x \in D\}$$

$$\mathcal{A} = \{(u, v, t)|\exists x \in D : f_i(x) \leq u_i, \forall i, h_i(x) = v_i, \forall i,\ f_0(x) \leq t\}$$

  (like an epigraph, but without the variable x)

  - Let p^* denote the optimal value of the primal problem:

$$p^* = \inf\{t|(u, v, t) \in G, u \leq 0, v = 0\} = \inf\{t|(u, v, t) \in \mathcal{A}, u \leq 0, v = 0\}$$

  - Let the dual for $\lambda \geq 0$

$$g(\lambda, \nu) = \inf\{t + \lambda^T u + \nu^T v|(u, v, t) \in G\} = \inf\{t + \lambda^T u + \nu^T v|(u, v, t) \in \mathcal{A}\}$$

  - $g(\lambda, \nu)$ is the intersection with the vertical axis at the highest line below $\mathcal{A}$. Clearly,

$$p^* \geq g(\lambda, \nu) \text{ (weak duality)}$$

Example: one inequality constraint

Example: one equality constraint

- If problem is convex, then $\mathcal{A}$ is convex:

- In this case, if further $\mathcal{A}$ satisfies Slater's condition: there exists an $x \in$ relint$D$ such that

$$f_i(x) < 0,\ \forall i$$

$$Ax = b$$

Equivalently:

$$\mathcal{A} \cap \{(u, v, t) : u < 0, v = 0\} \neq \emptyset$$

- Then it follows that there exists a hyperplane that supports $\mathcal{A}$ at the point $(0, 0, p^*)$.
- Equivalently, convexity and Slater condition together guarantee strong duality $p^* = g(\lambda^*, \nu^*)$

Theorem (Boyd p.227): if the problem is convex, Slater condition holds and rank$(A) = p$, where p is the number of equality constraints, then strong duality holds

- Slater condition: there exists an $x \in \text{relint} D$ such that

$$f_i(x) < 0, \ \forall i$$

$$Ax = b$$

In other words, there exists a feasible x which satisfies the inequality constraints strictly. This is a mild assumption, satisfied in many applications.

- Proof:

In light of the weak duality, we only need to show that there exists $\lambda^* \geq 0$ and ν^* such that $g(\lambda^*, \nu^*) = p^*$. We will use the separation theorem (handout provided)

Let

$$W(u, v) = \min f_0(x) \tag{1}$$

$$\text{s.t. } f_i(x) \leq u_i, \ \forall i \tag{2}$$

$$h_i(x) = v_i \ \forall i \tag{3}$$

Clearly, $p^* = W(0, 0)$.

- Consider the case $p^* < +\infty$ (the case $p^* = +\infty$ occurs when the problem is unfeasible)
- Equivalently,

$$W(u, v) = \inf\{t | (\bar{u}, \bar{v}, t) \in \mathcal{A}, \bar{u} \leq u, \bar{v} = v\}$$

Claim 1: $W(u, v)$ is a convex function of (u, v)

<u>Claim 2</u>: under the assumptions, $W(u, v)$ is well defined (i.e., not equal to $+\infty$) at a neighborhood of the origine $(0, 0)$, i.e., $\exists \epsilon > 0$ s.t. $W(u, v) < \infty, \ \forall (u, v) : \|u\|^2 + \|v\|^2 \leq \epsilon^2$.

- if u is changed from zero a little bit, we can still find x such that $f_i(x) < u_i, \ \forall i$
- if v is changed from zero a little bit, we can still find x such that $h_i(x) = v_i, \ \forall i$
- Similarly, if both u and v change from zero a little bit, we can still find x such that $f_i(x) < u_i, \ \forall i$ and $h_i(x) = v_i, \ \forall i$

<u>Claim 3</u>: Two cases:

1) $W(0, 0) = -\infty$, in that case $g(\lambda, \nu) = -\infty$ (by weak duality)

2) Now consider the case $W(0, 0) > -\infty$. Then, $W(u, v)$ must also take real values in a neighborhood of $(0, 0)$ (similar to previous claim)

- From the <u>separation theorem</u>, there exists a subgradient $(-\lambda, -\nu)$ of $W(u, v)$ at $(0, 0)$.

<u>Claim 4</u>: (λ, ν) is the Lagrange multiplier such that $g(\lambda, \nu) = W(0, 0) = p^*$

- By definition of subgradients:

$$W(u, v) \geq W(0, 0) + (-\lambda)^T u + (-\nu)^T v, \ \forall u, v$$

- Note that $W(u, 0) \leq W(0, 0)$ for $u \geq 0$ (by definition of W), hence

$$W(0, 0) \geq W(u, 0) \geq W(0, 0) + (-\lambda)^T u, \ \forall u \geq 0$$

i.e.

$$\lambda^T u \geq 0, \ \forall u \geq 0$$

or equivalently $\lambda \geq 0$.

- Let $x \in D$, $u_i(x) = f_i(x)$ and $v_i(x) = h_i(x)$. Then, by definition of W,

$$f_0(x) \geq W(u(x), v(x)), \ \forall x$$

and by definition of subgradient and using the fact that $W(0, 0) = p^*$,

$$f_0(x) \geq W(u(x), v(x)) \geq W(0, 0) + (-\lambda)^T u(x) + (-\nu)^T v(x) = p^* + (-\lambda)^T u(x) + (-\nu)^T v(x).$$

Rearranging the terms,

$$f_0(x) + \lambda^T u(x) + \nu^T v(x) \geq p^*, \ \forall x$$

Now, note that

$$g(\lambda, \nu) = \min_x f_0(x) + \lambda^T u(x) + \nu^T v(x)$$

hence

$$f_0(x) + \lambda^T u(x) + \nu^T v(x) \geq g(\lambda, \nu) \geq p^*$$

However, we must also have that $g(\lambda, \nu) \leq p^*$ (weak duality).

Together, these conditions imply

$$g(\lambda, \nu) = p^*$$

- Note also that complementary slackness conditions immediately follow.

In fact,

$$W(0,0) = g(\lambda, \nu) = \min_x f_0(x)+\lambda^T u(x)+\nu^T v(x) \leq f_0(x^*)+\lambda^T u(x^*)+\nu^T v(x^*) = W(0,0)+\lambda^T u(x^*)$$

where x^* is the minimizer of $p^* = W(0,0)$, which satisfies $u(x^*) \leq 0$ and $v(x^*) = 0$. In other words, we need to simultaneously obey the conditions

$$u(x^*) \leq 0,\ \ \lambda \geq 0,\ \ \lambda^T u(x^*) \geq 0$$

yielding

$$\lambda_i u_i(x^*) = 0,\ \ \forall i$$

- Example: separable problems (see also Boyd p.228)

  Find the discrete probability distribution with max entropy under power constraint

$$\max_p H(p) = -\sum_{i=-n}^{n} p_i \ln(p_i)$$

$$\text{s.t. } \sum_{i=-n}^{n} (i\Delta)^2 p_i \leq \sigma^2$$

$$\sum_{i=-n}^{n} p_i = 1$$

$$p_i \geq 0, \ \forall i \text{ (implicit constraints)}$$

  Solution:

- Note that the Lagrangian is separable in each i!

- This type of convex problems is said to have separable objective functions: although the constraint in the primal problem is coupled, duality helps us to decouple them in the Lagrangian. Hence, the minimization of the Lagrangian can be carried out independently for each variable.

- This property is useful for developing distributed solutions

- Example:

$$\min_{x} x^T x$$

$$\text{s.t. } Ax = b$$

- Linear-program:

$$\min_x c^T x$$

$$\text{s.t. } Ax = b$$

$$x \geq 0$$

- Example:

$$\min x_1^2 + 9x_2^2 \tag{4}$$

$$\text{s.t. } 2x_1 + x_2 \geq 1 \tag{5}$$

$$x_1 + 3x_2 \geq 1 \tag{6}$$

Solution:

- Alternative solution: check directly KKT conditions

- Partial Lagrangian: it is also possible to relax some constraints (by Lagrange multipliers) but not others:

$$\min_x f_0(x)$$
$$\text{s.t. } f_1(x) \le 0$$
$$f_2(x) \le 0$$

- Lagrangian

$$L(x, \lambda) = f_0(x) + \lambda f_1(x)$$

- Dual function:

$$g(\lambda) = \min f_0(x) + \lambda f_1(x), \text{ s.t. } f_2(x) \le 0$$

- Optimality condition of the Lagrangian becomes:

$$(\nabla f_0(x^*) + \lambda \nabla f_1(x^*))^T (x - x^*) \ge 0, \; \forall x : f_2(x) \le 0$$

along with

$$f_2(x^*) \le 0$$

- Dual problem:

$$\max g(\lambda), \text{ s.t. } \lambda \ge 0$$

- This property is useful when some constraints are easy or not coupled.

- Utility Maximization of a single resource:

$$\max_{x_i \geq 0} \sum_{i=1}^{n} U_i(x_i) \tag{7}$$

$$\text{s.t. } \sum_i x_i \leq R \tag{8}$$

$$\tag{9}$$

Solution:

- λ can be interpreted as the price of the resource: each user chooses $x_i(\lambda)$ that maximizes the local utility.

- if $\lambda = 0$ and $\sum_i x_i(\lambda) < R$, it means that the overall demand for the resource is low $\Rightarrow$ price is set to zero

- if $\lambda > 0$, then we must have $\sum_i x_i(\lambda) = R$, i.e., the entire resource pool is being used; the price is strictly positive to enforce that

- Example: Water-filling for fading channels

- Subgradient: the vector h is a subgradient of f at x_0 if

$$f(x) \geq f(x_0) + h^T(x - x_0), \ \forall x$$

  - The set of all subgradients of f at x_0 is called the subdifferential of f at x_0, denoted as $\partial f(x_0)$.

  - A convex function f has non-empty subdifferentials $\partial f(x_0)$ at any $x \in \text{int}(\text{dom} f)$

  - In particular, if f is differentiable, then

$$\partial f(x_0) \equiv \{\nabla f(x_0)\}$$