CS292A Convex Optimization: Gradient Methods and Online Learning **Spring 2019**

Lecture 8: May 2

Lecturer: Yu-Xiang Wang *Scribes: Yuxiao Zhang*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

8.1 Last Time: Stochastic Gradient Descent

Consider

$$\min_x \frac{1}{m}\sum_{i=1}^{m} f_i(x).$$

Stochastic gradient descent or SGD: let $x^{(0)} \in \mathbb{R}^n$, repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f_{i_k}(x^{(k-1)}),\ k = 1, 2, 3, \dots$$

where $i_k \in \{1, \dots, m\}$ is chosen uniformly at random. Step sizes t_k is chosen to be fixed and small, or diminishing.

Compare to full gradient, which would use $\frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x)$.

- Upside of SGD: much (potentially much, much) cheaper iterations, optimal for stochastic optimization.
- Downside of SGD: can be slow to converge, suboptimal for finite sum problems.

8.2 Lower Bounds in Linear Programs

Suppose we want to find lower bound on the optimal value in our convex problem, $B \leq \min_x f(x)$.

Example 8.1 *Consider the following simple linear program*

$$\begin{aligned} &\min_{x,y}\ x + y \\ &\textit{subjec to } x + y \geq 2 \\ &\qquad\qquad x, y \geq 0 \end{aligned}$$

It is easy to see that the lower bound is $B = 2$, because one of the constraints is exactly the same as the objective function.

Example 8.2 *Suppose the linear program is*

$$\begin{aligned} \min_{x,y}\ & x+3y \\ \textit{subjec to}\ & x+y \geq 2 \\ & x,y \geq 0 \end{aligned}$$

$$\begin{aligned} & x+3y \geq 2 \\ + \ & 2y \geq 0 \\ = \ & x+3y \geq 2 \end{aligned}$$

Lower bound $B = 2$.

Example 8.3 *More generally, suppose the linear program is*

$$\begin{aligned} \min_{x,y}\ & px+qy \\ \textit{subjec to}\ & x+y \geq 2 \\ & x,y \geq 0 \end{aligned}$$

$$\begin{aligned} a+b =& p \\ a+c =& q \\ a,b,c \geq& 0 \end{aligned}$$

Lower bound $B = 2a$, for any a, b, c satisfying above.

$$\begin{aligned} \min_{x,y}\ & px+qy \\ \textit{subjec to}\ & x+y \geq 2 \\ & x,y \geq 0 \end{aligned}$$

*Called **primal** LP.*

$$\begin{aligned} \max_{a,b,c}\ & 2a \\ \textit{subject to}\ & a+b = p \\ & a+c = q \\ & a,b,c \geq 0 \end{aligned}$$

*Called **dual** LP.*

Note that the number of dual variables is the number of primal constraints.

Example 8.4 *Now let us see another linear programming problem*

$$\begin{aligned} \min_{x,y}\ & px+qy \\ \textit{subjec to}\ & x \geq 0 \\ & y \leq 1 \\ & 3x+y = 2 \end{aligned}$$

***Primal** LP.*

$$\begin{aligned} \max_{a,b,c}\ & 2c-b \\ \textit{subject to}\ & a+3c = p \\ & -b+c = q \\ & a,b \geq 0 \end{aligned}$$

***Dual** LP.*

Note: in the dual problem, c is unconstrained.

We formulate the Duality for general form LP as following. Given $c \in \mathbb{R}^n,\ A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, G \in \mathbb{R}^{r\times n}, h \in \mathbb{R}^r$:

$$\begin{aligned} \min_x \; & c^T x \\ \text{subjec to } & Ax = b \\ & Gx \le h \end{aligned}$$

Primal LP.

$$\begin{aligned} \max_{u,v} \; & -b^T u - h^T v \\ \text{subject to } & -A^T u - G^T v = c \\ & v \ge 0 \end{aligned}$$

Dual LP.

Observe that for any u and $v \ge 0$ and x is primal feasible, we get

$$u^T(Ax - b) + v^T(Gx - h) \le 0 \Rightarrow (-A^T u - G^T v)^T x \ge -b^T u - h^T v.$$

So if $c = -A^T u - G^T v$, we get a bound on primal optimal value.

8.3 Example: Max Flow and Min Cut Problems

Given a directed graph $G = (V, E)$, define f_{ij}, $(i,j) \in E$ as the flow from node i to j. Denote c_{ij} as the capacity of the edge, which is the maximum amount of flow that one can push through that edge. In addition, the flow going into the node has to be equal to the flow coming out of the node. That is true for all nodes except for the source (s) and the sink (t) nodes. These constraints can be formulated as:

$$\begin{aligned} f_{ij} \ge& 0,\ (i,j) \in E \\ f_{ij} \le& c_{ij},\ (i,j) \in E \\ \sum_{(i,k)\in E} f_{ik} =& \sum_{(k,j)\in E} f_{kj},\ k \in V \subset \{s,t\}. \end{aligned}$$

The **max flow problem**: find flow that maximizes total value of the flow from s to t, i.e., as an LP:

$$\begin{aligned} \max_{f\in\mathbb{R}^{|E|}} \; & \sum_{(s,j)\in E} f_{sj} \\ \text{subject to } & 0 \le f_{ij} \le c_{ij} \text{ for all } (i,j) \in E \\ & \sum_{(i,k)\in E} f_{ik} = \sum_{(k,j)\in E} \quad \text{for all } k \in V \setminus \{s,t\} \end{aligned}$$

Follow the steps before, just flip the logic: Find the tightest upper bound of the objective by taking linear combinations of the constraints, subject to the constraints from the primal objectives coefficients.

Dual LP of max flow: The dual problem is (minimize over b, x to get best upper bound)

$$\begin{aligned} \min_{b\in\mathbb{R}^{|E|}, x\in\mathbb{R}^{|V|}} \; & \sum_{(i,j)\in E} b_{ij} c_{ij} \\ \text{subject to } & b_{ij} + x_j - x_i \ge 0 \text{ for all } (i,j) \in E \\ & b \ge 0,\ x_s = 1,\ x_t = 0 \end{aligned}$$

Suppose that at the solution, it just so happened that

$$x_i \in \{0,1\} \text{ for all } i \in V.$$

Let $A = \{i : x_i = 1\}$, $B = \{i : x_i = 0\}$; note $s \in A$ and $t \in B$. Then,

$$b_{ij} \geq x_i - x_j \text{ for } (i,j) \in E,\ b \geq 0$$

imply that $b_{ij} = 1$ if $i \in A$ and $j \in B$, and 0 otherwise. Moreover, the objective $\sum_{(i,j)\in E} b_{ij}c_{ij}$ is the capacity of cut defined by A, B.That is, we have shown that the dual is the LP relaxation of the min cut problem:

$$\begin{aligned} \min_{b\in\mathbb{R}^{|E|}, x\in\mathbb{R}^{|V|}} & \sum_{(i,j)\in E} b_{ij}c_{ij} \\ \text{subject to } & b_{ij} \geq x_i - x_j \\ & b_{ij}, x_i, x_j \in \{0,1\} \text{ for all } i,j \end{aligned}$$

Therefore, from what we have known so far, we have:

$$\text{value of max flow} \leq \text{optimal value for LP relaxed min cut} \leq \text{capacity of min cut}$$

A famous result called **max flow min cut theorem** : value of max flow through a network is exactly the capacity of the min cut. Hence, we have all the equalities in the above equation. In particular, we get that the primal LP and dual LP have exactly the same optimal values, a phenomenon called strong duality.

8.4 Another Perspective on LP Duality

Consider

$$\begin{aligned} \min_x \ & c^T x \\ \text{subjec to } & Ax = b \\ & Gx \leq h \end{aligned}$$	$$\begin{aligned} \max_{u,v} \ & -b^T u - h^T v \\ \text{subject to } & -A^T u - G^T v = c \\ & v \geq 0 \end{aligned}$$
Primal LP.	**Dual** LP.

For any u and $v \geq 0$, and x primal feasible,

$$c^T x \geq c^T x + u^T * (Ax - b) + v^T(Gx - h) := L(x,u,v),$$

and

$$f^* \geq \min_{x\in C} L(x,u,v) \geq \min_x L(x,u,v) := g(u,v),$$

where C denotes primal feasible set and f^* denotes the primal optimal value. In other words, $g(u,v)$ is a lower bound on f^* for any u and $v \geq 0$, $g(u,v) = \begin{cases} -b^T u - h^T v & \text{if } c = -A^T u - G^T v \\ -\infty & \text{otherwise} \end{cases}$

Now we can maximize $g(u, v)$ over u and $v \geq 0$ to get the tightest bound, and this gives exactly the dual LP as before. This last perspective is actually completely general and applies to arbitrary optimization problems (even nonconvex ones).

8.5 Mixed Strategies for Matrix Games

Setup: two players, J and R, and a payout matrix P.

Game: if J chooses i and R choose j, then J must pay R amount P_{ij} . P_{ij} can be either positive or negative. They use mixed strategies, i.e., each will first specify a probability distribution, and then

$$\begin{aligned} x :& \mathbb{P}(J \text{ choose } i) = x_i,\ i = 1, \dots, m \\ y :& \mathbb{P}(R \text{ choose } j) = y_j,\ i = 1, \dots, n \end{aligned}$$

The expected payout from J to R is then:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_i y_j P_{ij} = x^T P y.$$

- Universe 1: Now suppose that J will allow R to know his strategy x ahead of time. In this case, R will choose y to maximize $x^T P y$, which results in J paying off

  $$\max\{x^T P y : y \geq 0, 1^T y = 1\} = \max_{i=1,\dots,n} (P^T x)_i$$

  Js best strategy is then to choose his distribution x according to

  $$\begin{aligned} &\min_x \max_{i=1,\dots,n} (P^T x)_i \\ &\text{subject to } x \geq 0, 1^T x = 1 \end{aligned}$$

- Universe 2: If R allow J to know his strategy y beforehand. By the same logic, R's best strategy is to choose his distribution y according to

  $$\begin{aligned} &\max_y \min_{j=1,\dots,m} (Py)_j \\ &\text{subject to } y \geq 0, 1^T y = 1 \end{aligned}$$

Call Rs expected payout in first scenario f_1^* and expected payout in second scenario f_2^*. Because it is clearly advantageous to know the other players strategy, $f_1^* \geq f_2^*$ But by Von Neummans minimax theorem: we know that $f_1^* = f_2^*$.

Recast first problem as LP

$$\begin{aligned} &\max_{x,t}\ t \\ &\text{subject to } x \geq 0, 1^T x = 1 \\ &P^T x \leq t \end{aligned}$$

Now from the Lagrangian:

$$L(x, t, u, v, y) = t - u^T x + v(1 - 1^T x) + y^T (P^T x - t1)$$

and the Lagrange dual function

$$g(u,v) = \min_{x,t} L(x,t,u,v,y) = \begin{cases} v & \text{if } 1 - 1^T y = 0,\ Py - u - v1 = 0 \\ -\infty & \text{otherwise} \end{cases}$$

Hence the dual problem, after eliminating slack variable u is

$$\max_{y,v}\ v$$
$$\text{subject to } y \geq 0,\ 1^T y = 1$$
$$Py \geq v.$$

This is exactly the second problem, and therefore again we see that strong duality holds. In LPs, strong duality holds unless both the primal and dual are infeasible.

8.6 Duality in General

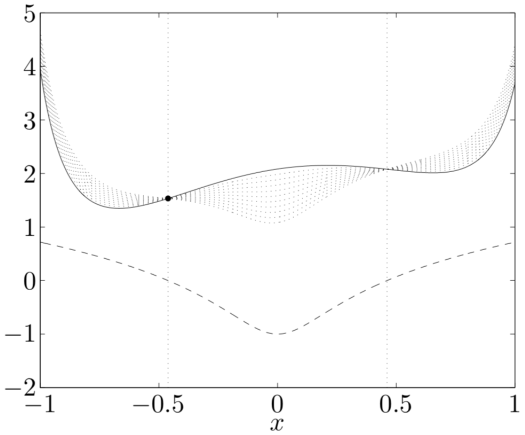
Consider general minimization problem

$$\min_x\ f(x)$$
$$\text{subject to } h_i(x) \leq 0, i = 1, \ldots, m$$
$$l_j(x) = 0, j = 1, \ldots, r.$$

We introduce new variables $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^r$ with $u \geq 0$, and define the Lagrangian to be

$$L(x,u,v) = f(x) + \sum_{i=1}^m u_i \underbrace{h_i}_{\leq 0} + \sum_{j=1}^r v_j \underbrace{l_j(x)}_{=0} \leq f(x)$$

as illustrated in Figure 8.6.



- Solid line is f
- Dashed line is h, hence feasible set $\approx [-0.46, 0.46]$
- Each dotted line shows $L(x,u,v)$ for different choices of $u \geq 0$

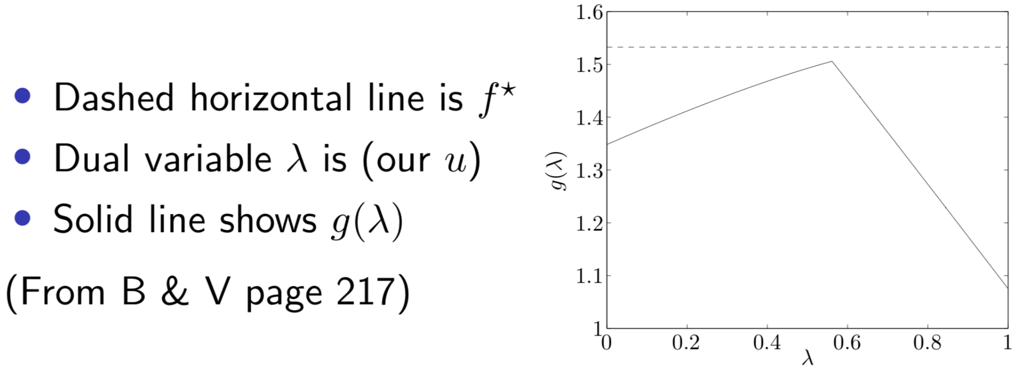
(From B & V page 217)

Figure 8.1:

Let C denote primal feasible set, f^* denote primal optimal value. Minimizing $L(x,u,v)$ over all x gives a lower bound

$$f^* \geq \min_{x \in C} L(x,u,v) \geq \min_x L(x,u,v) := g(u,v).$$

We call $g(u,v)$ Lagrangian dual function, and it gives a lower bound on f^* for any $u \geq 0$ and v, called dual feasible u, v. This is illustrated in Figure 8.1.



- Dashed horizontal line is $f^\star$
- Dual variable λ is (our u)
- Solid line shows $g(\lambda)$

(From B & V page 217)

Figure 8.2:

Example 8.5 *Consider quadratic program:*

$$\min_x \frac{1}{2} x^T Q x + c^T x$$
$$\text{subject to } Ax = b, x \geq 0$$

where $Q \succ 0$. *Lagrangian:*

$$L(x,u,v) = \frac{1}{2} x^T Q x + c^T x - u^T x + v^T (Ax - b).$$

Lagrange dual function

$$g(u,v) = \min_x L(x,u,v) = -\frac{1}{2}(c - u + A^T v)^T Q^{-1} (c - u + A^T v) - b^T v$$

For any $u \geq 0$ *and any* v, *this is lower a bound on primal optimal value* f^*.

Example 8.6 *Consider the same problem:*

$$\min_x \frac{1}{2} x^T Q x + c^T x$$
$$\text{subject to } Ax = b, x \geq 0$$

but now $Q \succeq 0$. *Lagrangian:*

$$L(x,u,v) = \frac{1}{2} x^T Q x + c^T x - u^T x + v^T (Ax - b).$$

Lagrange dual function

$$g(u,v) = \begin{cases} -\frac{1}{2}(c-u+A^Tv)^TQ^+(c-u+A^Tv) - b^Tv & \text{if } c-u+A^Tv \perp \text{ null}(Q) \\ -\infty & \text{otherwise} \end{cases}$$

where Q^+ denotes generalized inverse of Q. For any $u \geq 0$, v, and $c-u+A^Tv \perp$ null(Q), $g(u,v)$ is lower a nontrivial lower bound on f^*.

8.7 Weak Duality

The best lower bound is given by maximizing $g(u,v)$ over all dual feasible, u,v yielding Lagrange dual problem:

$$\max_{u,v} g(u,v)$$
$$\text{subject to } u \geq 0$$

Key property, called weak duality: fi dual optimal value is g^*, then

$$f^* \geq g^*.$$

Note that this always holds even if primal problem is nonconvex.

Another key property: the dual problem is a convex optimization problem (as written, it is a concave maximization problem). Again, this is always true, even when primal problem is not convex. By definition:

$$\begin{aligned} g(u,v) &= \min_x \{f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j l_j(x)\} \\ &= -\max_x \{-f(x) - \sum_{i=1}^m u_i h_i(x) - \sum_{i=1}^r v_j l_j(x)\}, \end{aligned}$$

i.e., g is concave in (u,v) and $u \geq 0$ is a convex constraint, hence dual problem is a concave maximization problem.

Weak duality

$$\min_{x \in H} \max_{y \in G} g(x,y) \geq \max_{x \in H} \min_{y \in G} g(x,y)$$

always true, but when

- Von Neumann (1928) H, G are probabilites, $g(x,y) = x^TPy$. Then strong duality holds.
- Sion (1952) H, G are convex, $g(x,y)$ is quasi-convex in X, quasi-concave in y. Then strong duality holds.
- Ky Fan (1958). $g(x,y)$ convex-concave. H, G one of them is compact. Then strong duality holds.

Example 8.7 *Define*

$$f(x) = x^4 - 50x^2 + 100x,$$

minimize subject to constraint $x \geq -4.5$. Dual function g can be derived explicitly, via closed-form equation for roots of a cubic equation.

$$g(u) = \min_{i=1,2,3} \{F_i^4(u) - 50F_i^2(u) + 100F_i(u)\},$$

where for $i = 1, 2, 3$,

$$F_i(u) = \frac{-a_i}{12 \cdot 2^{1/3}} \Big(432(100-u) - (432^2(100-u)^2) - 4 \cdot 1200^3)^{1/2}\Big)^{1/3} \\ - 100 \cdot 2^{1/3} \frac{1}{\big(432(100-u) - (432^2(100-u)^2 - 4 \cdot 1200^2)^{1/2}\big)^{1/3}},$$

and $a_1 = 1,\ a_2 = (-1 + i\sqrt{3})/2,\ a_3 = (-1 - i\sqrt{3})/2$ *as plotted in Figure 8.2.*

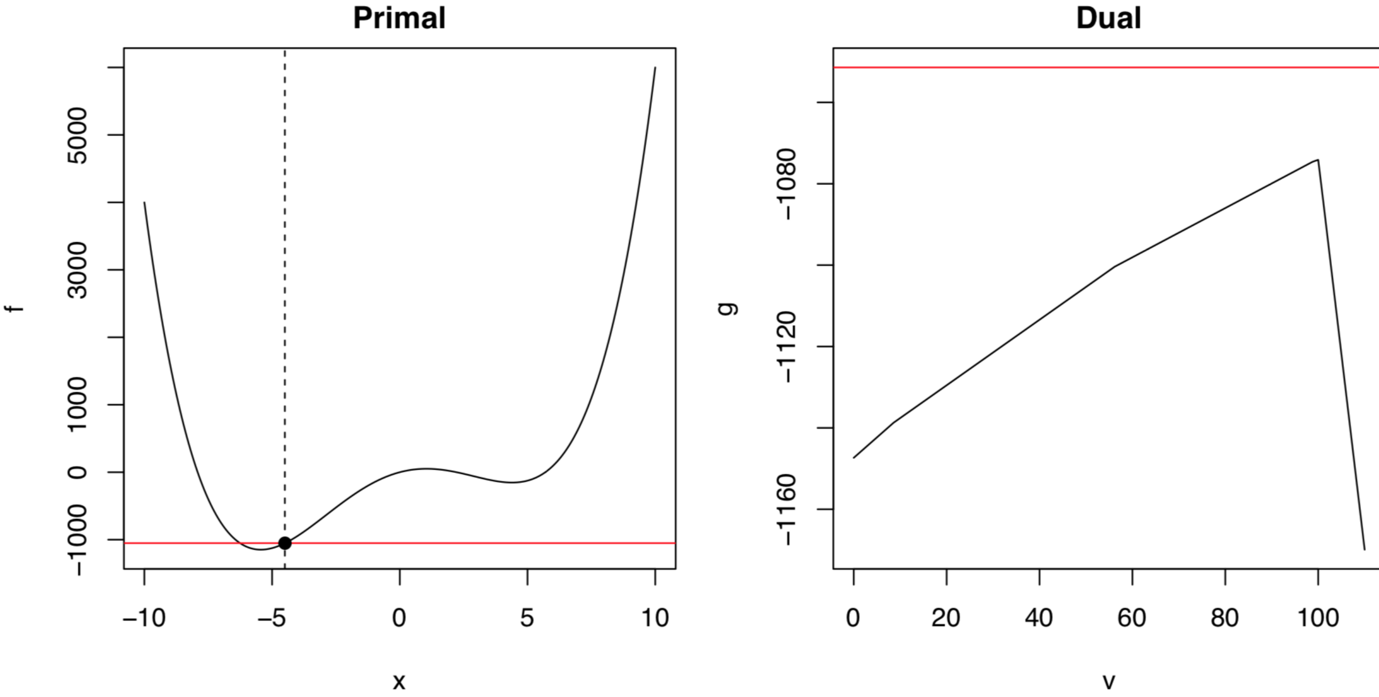


Figure 8.3: On the left, we compare the sample mean of optimal trajectories of $(X_{t_i})_{t_0 \le t_i \le t_N}$. The plot on the right show the comparison of sample mean of trajectories of optimal control $(\alpha_{t_i})_{t_0 \le t_i \le t_N}$ between approach 1 and approach 2.

8.8 Strong duality

Strong duality means that

$$f^* = g^*.$$

Slater's condition: if the primal is a convex problem (i.e., f and $h_1, \ldots, h_m$ are convex, $l_1, \ldots, l_r$ are affine), and there exists at least one strictly feasible $x \in \mathbb{R}^n$, meaning

$$h_1(x) < 0, \ldots, h_m(x) < 0 \text{ and } l_1(x) = 0, \ldots, l_r(x) = 0$$

then strong duality holds.

For linear program: strong duality holds for an LP if it is feasible. Apply same logic to its dual LP. Strong duality holds if it is feasible. Hence strong duality holds for LPs, except when both primal and dual are infeasible.

Example 8.8 *Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$, rows $x_1, \ldots, x_n$, recall the* ***support vector machine*** *problem:*

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \ & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \textit{subject to } & \xi_i \geq 0, \ i = 1, \ldots, n \\ & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots, n \end{aligned}$$

Introducing dual variables $v, w \geq 0$, we form the Lagrangian:

$$L(\beta, \beta_0, \xi, v, w) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n v_i\xi_i + \sum_{i=1}^n w_i(1 - \xi_i - y_i(x_i^T\beta + \beta_0)).$$

Minimizing over β, β_0, ξ gives Lagrange dual function:

$$g(v, w) = \begin{cases} -\frac{1}{2}w^T\tilde{X}\tilde{X}^Tw + 1^Tw \ \textit{if } w = C1 - v, \ w^Ty = 0 \\ -\infty \ \textit{otherwise} \end{cases}$$

where $\tilde{X} = diag(y)X$. Thus SVM dual problem, eliminating slack variable v, becomes

$$\begin{aligned} \max_{w} \ & -\frac{1}{2}w^T\tilde{X}\tilde{X}^Tw + 1^Tw \\ \textit{subject to } & 0 \leq w \leq C1, w^Ty = 0 \end{aligned}$$

We are able to check Slater's condition is satisfied, and we have strong duality. Further, from study of SVMs, might recall that at optimality

$$\beta = \tilde{X}^Tw.$$

This is not a coincidence, as we will later revisit when learning about KKT conditions.

8.9 Duality Gap

Given primal feasible x and dual feasible u, v, the quantity

$$f(x) - g(u, v)$$

is called the **duality gap** between x and u, v. Note that

$$f(x) - f^* \leq f(x) - g(u, v).$$

So if the duality gap is zero, then x is primal optimal (and similarly, u, v are dual optimal). From an algorithmic viewpoint, it can provide a stopping criteria: if $f(x) - g(u, v) \leq \epsilon$, then we are guaranteed that $f(x) - f^* \leq \epsilon$.

References

[] D. Boyd and L. Vandenberghe, *Convex Optimization*, 2004.

[] R.T. Rockafellar, *Convex Analysis*, 1970.