10-425/625: Introduction to Convex Optimization (Fall 2023)

Lecture 18: KKT Conditions & Log-Barrier Analysis

Instructor:[1] *Matt Gormley* *November 1, 2023*

18.1 Newton's Method

Sparse, structured problems When the inner linear systems (in Hessian) can be solved efficiently and reliably, Newton's method can strive

For example, if $\nabla^2 f(x)$ is sparse/structured for all x, say banded, then both memory and computation are $O(n)$ per Newton iteration

What functions admit a structured Hessian? Two examples:

- If $g(\beta) = f(X\beta)$, then $\nabla^2 g(\beta) = X^T \nabla^2 f(X\beta) X$. Hence if X is a structured predictor matrix and $\nabla^2 f$ is diagonal, then $\nabla^2 g$ is structured
- If we seek to minimize $f(\beta) + g(D\beta)$, where $\nabla^2 f$ is diagonal, g is not smooth, and D is a structured penalty matrix, then the Lagrange dual function is $-f^*(-D^T u) - g^*(-u)$. Often $\nabla^2 f^*$ will be diagonal (e.g., when $f(\beta) = \sum_{i=1}^p f_i(\beta_i)$) so the Hessian in dual will be structured

18.1.1 Quasi-Newton methods

If the Hessian is too expensive (or singular), then a quasi-Newton method can be used to approximate $\nabla^2 f(x)$ with $H \succ 0$, and we update according to

$$x^+ = x - tH^{-1}\nabla f(x)$$

- Approximate Hessian H is recomputed at each step. Goal is to make H^{-1} cheap to apply (possibly, cheap storage too)
- Convergence is fast: superlinear, but not the same as Newton. Roughly n steps of quasi-Newton make same progress as one Newton step

[1]These notes were originally written by Ryan Tibshirani for 10-725 Fall 2019 (original version: here) and were edited and adapted for 10-425/625.

- Very wide variety of quasi-Newton methods; common theme is to "propogate" computation of H across iterations

18.2 Strong Duality

We made the (simple) observation that $p^* \geq d^*$, i.e. that *weak duality* always holds. In cases where $p^* = d^*$ we say that *strong duality* holds. We will refer to $p^* - d^*$ as the *duality gap*.

Duality is most useful when strong duality holds, and we will develop several insightful consequences of strong duality. Before we do this, let us explore when strong duality holds.

18.2.1 (Relaxed) Slater's Condition

The basic punchline is roughly that – strong duality holds for most convex problems (except a few pathological ones), and rarely holds for non-convex problems.

To be a bit more precise we'll describe a popular set of conditions which are sufficient for strong duality to hold for a convex optimization problem. The conditions we describe are called weak/relaxed Slater's conditions. (The broader area under which results of this form fall are called either "constraint qualifications" in the special case of Lagrangians, or minimax theorems more generally.)

Suppose we're again interested in a problem of the form:

$$\begin{aligned} \min_x \ & f(x) \\ \text{subject to } & h_i(x) \leq 0 \quad i \in \{1, \ldots, m\} \\ & \ell_j(x) = 0, \quad j \in \{1, \ldots, r\}. \end{aligned}$$

We'll denote by $\mathcal{D}$ the (implicit) domain of the problem, i.e. where all the constraint and objective functions are finite. Some of our inequality constraints h_i may be affine, without loss of generality we'll assume that we re-order the constraints so that for some $k \in \{0, \ldots, m\}$ the constraints $h_1, \ldots, h_k$ are affine.

Slater's Theorem: Suppose that there exists a point $x_0 \in$ relative int$(\mathcal{D})$ such that,

$$\begin{aligned} \ell_j(x_0) &= 0, \quad j \in \{1, \ldots, r\} \\ h_i(x_0) &\leq 0, \quad i \in \{1, \ldots, k\} \\ h_i(x_0) &< 0, \quad i \in \{k+1, \ldots, m\}, \end{aligned}$$

then strong duality holds, i.e. $p^* = d^*$.

In words, Slater's condition simply requires that there is some feasible point x_0, which is *strictly* feasible for the non-affine inequality constraints. This is usually a rather mild assumption (roughly it is saying that the feasible region must have an interior point). It is worth noting that this is not requiring this property to hold for the optimal solution x^*.

An important implication of Slater's condition is the following LP strong duality theorem (in an LP all constraints are affine, so Slater's conditions simply reduces to checking feasibility):

LP Strong Duality: If in an LP, either the primal or dual is feasible then strong duality holds, i.e. $p^* = d^*$.

Some people would add to this a few more cases which are covered by weak duality to conclude the following slightly more general LP strong duality theorem:

1. If both are infeasible, then strong duality fails (but weak duality of course, always holds).
2. If either primal or dual is feasible, then strong duality holds.
3. If the dual is unbounded, then the primal must be infeasible and strong duality holds.
4. If the primal is unbounded, then the dual must be infeasible and strong duality holds.

QP Strong Duality: In a similar vein, for a QP strong duality holds if either the primal or dual is feasible (once again, all the constraints are affine so Slater's conditions just boil down to checking feasibility).

18.2.2 Minimax Formulation

Our treatment of duality seems so far to be a bit asymmetric, i.e. we often treated the primal as special (given to us, and the main object of interest) and the dual as some auxiliary program we derived. However, they are both in fact completely symmetric objects that can be derived from the Lagrangian.

Suppose that we have a Lagrangian, for $v \geq 0$,

$$L(x, u, v) := f(x) + \sum_{j=1}^{r} u_j \ell_j(x) + \sum_{i=1}^{m} v_j h_j(x).$$

We have already seen how to derive the dual from the Lagrangian. Then the following observation shows that we can always derive the primal from the Lagrangian. If you're not familiar with inf and sup you can replace them in your parsing by min and max. Observe that,

$$\sup_{u,v \geq 0} L(x, u, v) = \begin{cases} f(x) & \text{if } x \text{ is feasible} \\ \infty & \text{otherwise.} \end{cases}$$

This is easy to check – if x violates any of the constraints, then if we set the corresponding u or v to $\to \infty$ we obtain that the supremum is ∞. On the other hand when x satisfies all the constraints the supremum is achieved by setting $u, v = 0$, in which case $\sup_{u,v \geq 0} L(x, u, v) = f(x)$.

This in turn means that we can write the primal optimal value in terms of the Lagrangian:

$$p^* = \inf_x \sup_{u,v \geq 0} L(x, u, v).$$

We have already noted that the dual optimal value is simply:

$$d^* = \sup_{u,v \geq 0} \inf_x L(x, u, v).$$

With these definitions in place we can see that weak duality is simply the statement that:

$$\inf_x \sup_{u,v \geq 0} L(x, u, v) \geq \sup_{u,v \geq 0} \inf_x L(x, u, v).$$

This statement is of course always true (i.e. does not require any conditions on L whatsoever). In game-theoretic language, this is the observation that in a two-player, zero-sum game the first player is always at a disadvantage (the second player observes the first players' move and then gets to choose the best response).

On the other strong duality is the (non-trivial) statement that,

$$\inf_x \sup_{u,v \geq 0} L(x,u,v) = \sup_{u,v \geq 0} \inf_x L(x,u,v).$$

When this property holds there is no longer any advantage to going second. Theorems that give conditions under which this equality holds are called minimax theorems. Slater's conditions yield a minimax theorem in the restricted type of Lagrangian game that we were interested in but there are many other minimax theorems that hold in different settings.

An important concept in this setting is that of a *saddle point* (we have encountered saddle points in pure minimization problems before, now we're encountering them in min-max problems). A point $(x^*, (u^*, v^*))$ is a saddle point if we have:

$$L(x, u^*, v^*) \geq L(x^*, u^*, v^*) \geq L(x^*, u, v) \text{ for any } (x,u,v) \in \mathcal{D}.$$

The reason they're an important concept is the following (not difficult to prove) fact: $(x^*, (u^*, v^*))$ is a saddle point if and only if strong duality holds, and in this case,

$$x^* = \arg\inf_x \sup_{u,v \geq 0} L(x,u,v)$$
$$(u^*, v^*) = \arg\sup_{u,v \geq 0} \inf_x L(x,u,v).$$

In words, any saddle point gives us a pair of primal, dual optimal solutions, and conversely any primal-dual optimal pair defines a saddle point.

18.3 KKT Conditions and Optimality

We have already at some point discussed first-order optimality conditions for general convex programs. These are extremely useful (and we already have

used them to reason about properties of projections and proximal operators for instance), but they are not always very transparent to use since we (at that time) didn't know much about the convex set $\mathcal{C}$ for $\min_{x \in \mathcal{C}} f(x)$.

It will turn out for the types of inequality, equality constrained optimization problems we are discussing in this lecture – strong duality will yield some natural first-order optimality conditions which are often useful.

We'll assume throughout this section that our constraint and objective functions are differentiable (things will carry over in the convex case if you replace gradients by subgradients). We have our usual convex optimization problem:

$$\begin{aligned} \min_x \ & f(x) \\ \text{subject to } & h_i(x) \leq 0 \quad i \in \{1, \ldots, m\} \\ & \ell_j(x) = 0, \quad j \in \{1, \ldots, r\}. \end{aligned}$$

Here f, h_i are convex, and ℓ_j are affine. Given a candidate pair of primal-dual points $(\widehat{x}, \widehat{u}, \widehat{v})$, we will say they satisfy the Karush-Kuhn-Tucker (KKT) conditions if:

$$h_i(\widehat{x}) \leq 0, \quad i \in \{1, \ldots, m\} \tag{18.1}$$

$$\ell_j(\widehat{x}) = 0, \quad j \in \{1, \ldots, r\} \tag{18.2}$$

$$\widehat{v} \geq 0, \tag{18.3}$$

$$\widehat{v}_i h_i(\widehat{x}) = 0, \quad i \in \{1, \ldots, m\} \tag{18.4}$$

$$\nabla f(\widehat{x}) + \sum_{i=1}^{m} \widehat{v}_i \nabla h_i(\widehat{x}) + \sum_{j=1}^{r} \widehat{u}_i \nabla \ell_j(\widehat{x}) = 0. \tag{18.5}$$

The two conditions (18.1),(18.2) are called *primal feasibility*, the condition (18.3) is *dual feasibility*, the condition (18.4) is *complementary slackness*, and (18.5) is *stationarity*. We'll refer to points (x, u, v) which satisfy these conditions as *KKT points*.

The KKT conditions characterize optimal solutions to the primal and dual in the following sense:

1. A sufficient condition for x^* to be a primal optimal solution, is that there exists a (u^*, v^*) such that (x^*, u^*, v^*) is a KKT point. Similarly, a sufficient condition for (u^*, v^*) to be a dual optimal solution is that there exists an x^* such that (x^*, u^*, v^*) is a KKT point. Equivalently,

any KKT point gives an optimal solution to the primal and dual problems.

2. When strong duality holds, this is also a necessary condition, i.e. (x^*, u^*, v^*) are optimal primal-dual solutions if and only if they are KKT points.

The proofs of these claims are fairly simple, but insightful so we'll discuss them briefly.

18.3.1 Sufficiency

Suppose that $(\widehat{x}, \widehat{u}, \widehat{v})$ satisfy the KKT conditions. The condition (18.5) is equivalent to the fact that:

$$\nabla_x L(\widehat{x}, \widehat{u}, \widehat{v}) = 0,$$

and since the functions f, h, ℓ are convex this implies that $\widehat{x}$ is a minimizer of $L(x, \widehat{u}, \widehat{v})$, i.e.

$$L(\widehat{x}, \widehat{u}, \widehat{v}) \leq L(x, \widehat{u}, \widehat{v}).$$

This in turn means that, $g(\widehat{u}, \widehat{v}) = L(\widehat{x}, \widehat{u}, \widehat{v})$. We also observe that,

$$g(\widehat{u}, \widehat{v}) = L(\widehat{x}, \widehat{u}, \widehat{v}) = f(\widehat{x}),$$

since by primal feasibility and complementary slackness the other terms in the Lagrangian are 0. We already know (by weak duality), that for any feasible solutions (x, u, v) we have that,

$$g(u, v) \leq f(x),$$

so we conclude that $f(\widehat{x}) = g(\widehat{u}, \widehat{v}) \leq f(x)$ for any feasible x, i.e. $\widehat{x}$ is primal optimal. Similarly, $g(u, v) \leq f(\widehat{x}) = g(\widehat{u}, \widehat{v})$ for any feasibly (u, v), i.e. $(\widehat{u}, \widehat{v})$ is dual optimal.

Note that, along the way we have shown that if there is any KKT point $(\widehat{x}, \widehat{u}, \widehat{v})$ then strong duality holds.

18.3.2 Necessity

Suppose that strong duality holds, and we are given a pair of (feasible) optimal solutions (x^*, u^*, v^*). We already know that they must satisfy (18.1),(18.2),(18.3) since the solutions are feasible.

We also know that,

$$\begin{aligned} f(x^*) &= g(u^*, v^*) \\ &= \inf_x \left[f(x) + \sum_{j=1}^{r} u_j^* \ell_j(x) + \sum_{i=1}^{m} v_j^* h_j(x) \right] \\ &\leq f(x^*) + \sum_{j=1}^{r} u_j^* \ell_j(x^*) + \sum_{i=1}^{m} v_j^* h_j(x^*) \\ &= f(x^*) + \sum_{i=1}^{m} v_j^* h_j(x^*) \\ &\leq f(x^*). \end{aligned}$$

This means that all the inequalities above must in fact be equalities. This in turn means that $\sum_{i=1}^{m} v_j^* h_j(x^*) = 0$, but since each term in the sum is non-positive the only way the sum can be zero is if every term is 0, i.e. that (18.4) holds.

We also observe that, $f(x^*) = g(u^*, v^*) = \inf_x L(x, u^*, v^*)$, i.e. x^* is a minimizer of $L(x, u^*, v^*)$. Since this latter function is convex and differentiable, we know that $\nabla_x L(x^*, u^*, v^*) = 0$ which is precisely our last remaining KKT condition (18.5).

We have thus argued that any optimal solution x^* to the primal and (u^*, v^*) to the dual satisfies the KKT conditions.

18.3.3 KKT Without Convexity

For a general program of the form we described above (without convexity of f, h_i say), the (subgradient) KKT conditions are still sufficient (always) and necessary (when strong duality holds). First, we'll need to modify the stationarity KKT condition to be:

$$0 \in \partial f(\widehat{x}) + \sum_{j=1}^{r} \widehat{u}_j \partial \ell_j(\widehat{x}) + \sum_{i=1}^{m} \widehat{v}_j \partial h_j(\widehat{x}).$$

This is a subtle but extremely important difference. Notice that, even when the functions are differentiable this KKT condition is *not* the same as before (since we're now talking about potentially non-convex functions).

An important thing to notice is that for an unconstrained problem x^* minimizes some function f if and only if $0 \in \partial f(x^*)$ (we showed this before when we talked about sub-gradient optimality conditions) and this doesn't require convexity.

Necessity (under strong duality) and Sufficiency: The proofs of necessity and sufficiency above go through unchanged. The only reason we used convexity before was in reasoning about stationarity, i.e. to say if $\widehat{x}$ minimizes some unconstrained convex, differentiable function, then the gradient at $\widehat{x}$ must be 0. Now, we simply replace this by the subgradient stationarity condition above, and all other steps remain the same.

18.4 Support Vector Machines

Suppose given labeled data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $y_i \in \{-1, +1\}$ and our goal is to learn a linear classifier. One way to do this is by trying to maximize the margin of the classifier. In the case when the data is linearly separable the margin of a (perfect) classifier is the minimum distance of any point to the decision boundary. In case the data is not linearly separable we allow points to violate the margin (by introducing slack variables), but penalize this violation. This results in the following optimization problem:

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

This is a QP (and is clearly feasible) so via the weak Slater's conditions we know that strong duality holds. It is worth noting that have two constraints on ξ_i – i.e. that $\xi_i \geq 0, \xi_i \geq 1 - y_i(x_i^T\beta + \beta_0)$, and we get penalized for large values of ξ_i. Consequently, we can write the SVM optimization equivalently in a reduced form as:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \min\{0, 1 - y_i(x_i^T\beta + \beta_0)\},$$

which can be viewed as a penalized empirical risk where the penalty is the ℓ_2^2-norm of β and the risk corresponds to the average *hinge loss*.

Returning to our original SVM formulation suppose we introduce dual variables $v, w \geq 0$ for the inequality constraints, then we can write the Lagrangian:

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

We can minimize this over ξ and β, β_0 to see that we must satisfy the following conditions:

$$\begin{aligned} \beta &= \sum_{i=1}^{n} w_i y_i x_i, \\ C - v_i - w_i &= 0 \\ \sum_{i=1}^{n} w_i y_i &= 0. \end{aligned}$$

Making these substitutions yields the dual function:

$$g(v, w) = \frac{1}{2}\left(\sum_{i=1}^{n} w_i y_i x_i\right)^T \left(\sum_{i=1}^{n} w_i y_i x_i\right) + \sum_{i=1}^{n} w_i(1 - y_i x_i^T(\sum_{i=1}^{n} w_i y_i x_i)).$$

This yields the dual SVM program:

$$\begin{aligned} &\max_{v \geq 0, w \geq 0} g(v, w), \\ &\text{subject to } C - v_i - w_i = 0, \\ &\sum_{i=1}^{n} w_i y_i = 0. \end{aligned}$$

which can be (after eliminating the v variables) can be written as:

$$\begin{aligned} &\max_{w} \sum_{i=1}^{n} w_i - \frac{1}{2}\left(\sum_{i=1}^{n} w_i y_i x_i\right)^T \left(\sum_{i=1}^{n} w_i y_i x_i\right) \\ &\text{subject to } 0 \leq w_i \leq C, \\ &\sum_{i=1}^{n} w_i y_i = 0. \end{aligned}$$

The SVM dual is also a QP so it is perhaps not immediately obvious why this was a useful sequence of steps to carry out. However, the dual is the entry point to the world of RKHS/kernel machines. We notice that the dual program does not require the actual features x_i to be given but rather only requires the inner products between pairs of features i.e. $x_i^T x_j$. This in turn suggests we could (implicitly) fit a linear classifier in a transformed feature space $\phi(x)$ so long as we know how to evaluate inner products $\phi(x_i)^T \phi(x_j)$ (since we could plug in these values and solve the dual). This is the so-called kernel trick.

Another feature of the dual will be useful in this transformation, which is that given a dual solution w, we can find the primal solution:

$$\beta = \sum_{i=1}^{n} w_i y_i x_i.$$

This idea carries through to the kernelized case, where we observe that to evaluate the classifier at a point x we simply need $\beta^T x$ which in turn can be expressed in terms of the inner products between $\phi(x)$ and the training data $\phi(x_i)$. To find β_0 we need to understand the KKT conditions a bit better.

Furthermore, in this case, the KKT conditions are quite insightful. If we returned to the non-reduced dual, we know that by complementary slackness we must have:

$$\begin{aligned} v_i \xi_i &= 0, \\ w_i(1 - \xi_i - y_i(x_i^T \beta + \beta_0)) &= 0, \end{aligned}$$

but we also have the constraint that $v_i = (C - w_i)$. This yields some facts:

1. If $y_i(x_i^T \beta + \beta_0) > 1$ then we have already observed that $\xi_i = 0$, so it must be the case that $w_i = 0$.
2. If $y_i(x_i^T \beta + \beta_0) < 1$ then we know that $\xi_i > 0$, so $v_i = 0$, i.e. $w_i = C$.
3. If $w_i = 0$, then we know that $\xi_i = 0$, so we know that $y_i(x_i^T \beta + \beta_0) \geq 1$.
4. If $0 < w_i < C$ then we know that $\xi_i = 0$, and therefore that $y_i(x_i^T \beta + \beta_0) = 1$.

The points for which $w_i > 0$ are called *support vectors*. Now, finally, if we can find any point for which $0 < w_i < C$ then we know that, we can write

$\beta_0 = 1/y_i - x_i^T\beta$, (i.e. we can use such points to find β_0). It turns out that if you cannot find such a point the SVM optimization is degenerate, and the optimal $\beta = 0$ and β_0 is either $+1$ or -1 depending on which class is a majority in the training data.