10-725: Optimization **Fall 2012**

Lecture 16: October 18

Lecturer: Geoff Gordon/Ryan Tibshirani *Scribes: Deyang Zhao and Xiaofei Liu*

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.*

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

16.1 Review on duality

16.1.1 Construction

Given a general minization problem,

$$\min_{x\in\mathbb{R}^n} f(x)$$
$$s.t. h_i(x) \leq 0, i = 1, ..m$$
$$l_i(x) = 0, j = 1, ...r$$

These do not need to be convex functions. We can write the Lagrangian:

$$L(x, u, v) = f(x) + \sum u_i h_i(x) + \sum v_i l_i(x)$$

Contrain u_i to be non-negative. One property of the Lagrangian is:

$$f(x) \geq L(x, u, v)$$

where x is feasible. So it's a lower bound of $f(x)$ over the feasible set.
Minimize the both sides, we have

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

Here $g(u, v)$ is called the dual funciton, which provides a lower bound of the primal optimal value f^*. To get the best lower bound, g is maximized over u, v, yielding the dual problem

$$\max_{v\in\mathbb{R}^m, v\in\mathbb{R}^r} g(u, v)$$
$$s.t. u \geq 0$$

16.1.2 Properties

Weak Duality By construction, $f^* \geq g^*$ is always true (even the primal problem is not convex).
Another key property is that the dual problem is always convex (again even the primal problem is not convex).
Notes for proof

- Minimizing a convex function and maxmizing a concave function over a convex set are both convex problems
- Minimizing a convex f is maximizing $-f$, which is concave.
- Dual problem is concave because it is affine of u, v

Example:

$$\min f(x) = x^4 - 50x^2 + 100x, s.t. x \geq -4.5$$

Minimizing the Lagrangian over x involves the differential of f, which is a cubic function. It has a closed-form solution of the roots.

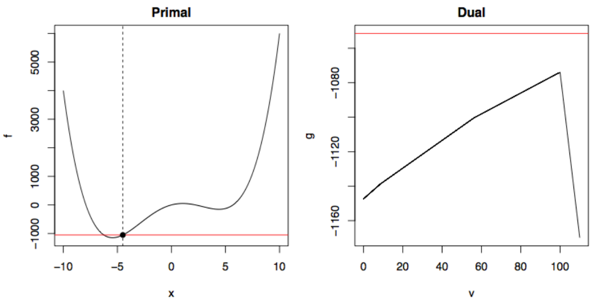

Figure 16.1: Nonconvex quartic minimization

The red line in the right figure is the primal optimal value and the black line is the dual function value. Here the strong duality is not satisfied.

16.1.3 More on strong duality

The stong duality holds when Slater's condition is satisfied. Then back to LP with duality. Since all the constraints are linear, if the primal LQ is feasible, then strong duality holds. In addition, if the primal LP is not feasible but the dual LP is, strong duality holds as well.

16.1.4 Duality gap

Defined as on feasible x, u, v:

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

Since

$$f(x^*) \geq g(u, v)$$
$$f(x) \geq g(u^*, v^*)$$

if the duality gap is zero, then x is primal optimal u, v are dual optimal.
Or we can say

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

This provides a upper bound of how far away from the primal optimal solution. If the dual is easy to evaluate, this is a good stopping criterion for gradient descent.

16.2 KKT conditions

Setup:

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$s.t. h_i(x) \leq 0, i = 1, ..m$$

$$l_i(x) = 0, j = 1, ...r$$

KKT conditions:

- **Stationarity** $0 \in \partial L(x, u, v)$, i.e. zero is subgradient of Lagrangian. This indicates x minimizes Lagrangian over all x.
- **Complementary slackness** $u_i h_i(x) = 0, \forall i$. If $h_i(x)$ is strictly less than zero, then $u_i = 0$. If $h_i(x)$ equals zero, then u_i can be any value.
- **Primal feasibility** $h_i(x) \leq 0, l_i(x) = 0, \forall i, j$
- **Dual feasibility** $u_i \geq 0, \forall i$

Lemma 16.1 *If x^* and u^*, v^* are primal and dual solutions, with zero duality gap, then x^* and u^*, v^* satisfy the KKT conditions*

Proof: Let x^*, u^*, v^* be primal and dual solutions with zero duality gap. Then $f(x^*) = g(u^*, v^*)$.
By the dual fucntion definition,

$$g(u^*, v^*) = \min_{x \in \mathbb{R}^n} f(x) + \sum u_i^* h_i(x) + \sum v_j^* l_j(x)$$

Since it's minimum, we can plug in x^*,

$$\min_{x \in \mathbb{R}^n} f(x) + \sum u_i^* h_i(x) + \sum v_j^* l_j(x) \leq f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*)$$

Since x^* is primal feasible,

$$\sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) \leq 0$$

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

The conclusion is that all these inequalities are actually equalities.
More specifically, look at the first inequality. When it is equality, we know x^* also minimizes the Lagrangian. $\rightarrow 0 \in \partial f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) \rightarrow$ stationarity condition.
For the second inequality, we have $\sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) = 0$. $\sum v_j^* l_j(x^*) = 0$ because x^* is feasible. So $\sum u_i^* h_i(x^*) = 0$. Given $u_i^* h_i(x^*) \leq 0, \forall i$, we know actually $u_i^* h_i(x^*) = 0, \forall i$. $\rightarrow$ complementary slackness.
Primal feasibility and dual feasibility are obvious. ■

Lemma 16.2 *If x^* and u^*, v^* satisfy the KKT conditions, then x^* and u^*, v^* are primal and dual solutions.*

Proof: By the dual fucntion definition,

$$g(u^*, v^*) = \min_{x \in \mathbb{R}^n} f(x) + \sum u_i^* h_i(x) + \sum v_j^* l_j(x)$$

By stationarity, x^* minimizes the Lagrangian, i.e.

$$g(u^*, v^*) = L(x^*, u^*, v^*) = f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*)$$

Given complementary slackness and primal feasibility,

$$\sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) = 0$$

$$g(u^*, v^*) = f(x^*)$$

Therefore duality gap is zero andx^* and u^*, v^* are primal and dual feasible so x^* and u^*, v^* are primal and dual optimal ■

Then we can say

Theorem 16.3 *For a problem with strong duality (e.g., assume Slaters condition: convex problem and there exists x strictly satisfying non- affine inequality contraints),*
x^* and u^*, v^* are primal and dual solutions $\Longleftrightarrow$ x^* and u^*, v^* satisfy the KKT conditions

16.3 Examples

16.3.1 Quadratic with equality constraints

Let's consider for $Q \succeq 0$, and form the following optimization problem,

$$\min_{x \in \mathbb{R}^n} \tfrac{1}{2} x^T Q x + c^T x$$
$$\text{subject to } Ax = 0$$

An example of this problems comes up is Newton's method.When we compute Newton step for $\min_{x \in \mathbb{R}^n} f(x)$ subject to $Ax = b$
It is a convex problem with no inequality constraints, so we will have only one set of new variables corresponding to the equality constraints, called u.
We can define the **Lagrangian**

$$L(x, u) = \tfrac{1}{2} x^T Q x + c^T x + u(Ax - 0)$$

For x and u which satisfies KKT conditions, we need to have:
Stationarity: $0 = Qx + c + A^T u$
Primal feasibility: $Ax = 0$
There is no complementary slackness and dual feasibility conditions because no inequality constraints exist.
We can write these conditions in a matrix form:

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix}$$

16.3.2 Lasso

Let's consider the lasso problem:
Given response $y \in \mathbb{R}^n$, predictors $A \in R^{n \times p}$(columns $A_1, ..., A_p$), solve the optimization problem

$$\min_{x\in\mathbb{R}^p} \frac{1}{2}\|y-Ax\|^2+\lambda\|x\|_1$$

Since there is no constraints, the KKT conditions will only contain stationarity condition:

$$0=-A^T(y-Ax)+\lambda s$$

where $s\in\partial\lambda\|x\|_1$, i.e.,

$$s_i\in\begin{cases}1 & \text{if } x_i>0\\ -1 & \text{if } x_i<0\\ [-1,1] & \text{if } x_i=0\end{cases}$$

Rewrite the condition as:

$$A^T(y-Ax)=\lambda s$$

We can derive from this condition that if $|A_i^T(y-Ax)|<\lambda$, then $x_i=0$. According to the KKT conditions, $s\in(-1,1)$ strictly. And the only time this can happen is when $x_i=0$.

16.3.3 Group Lasso

Group lasso problem appears in the situation where instead of individual predictors, we want to select entire groups such that each group is scientific meaningful.
This problem is similar to lasso problem,while predictors A is split up into groups, i.e., $A=[A_{(1)}A_{(2)}...A_{(G)}]$, Also, the coefficient vector is split up into the same group, i.e., $x=[x_{(1)}x_{(2)}...x_{(G)}]$
The group lasso problem can be written as the following:

$$\min_{x=(x_{(1)}x_{(2)}...x_{(G)})\in\mathbb{R}^p}\frac{1}{2}\|y-Ax\|^2+\lambda\sum_{i=1}^{G}\sqrt{p_{(i)}}\|x_{(i)}\|_2$$

The difference between the lasso problem is that in group lasso problem the 1-norm penalty is replaced by a sum of 2-norm penalty. $\sqrt{p_{(i)}}$ is a term that counts for group sizes, where $p_{(i)}$ is the number of variables in the i_st group.
Group lasso smooths the 1-norm ball of lasso problem in some direction, as shown in the picture below (From Yuan and Lin (2006), Model selection and estimation in regression with grouped variables").

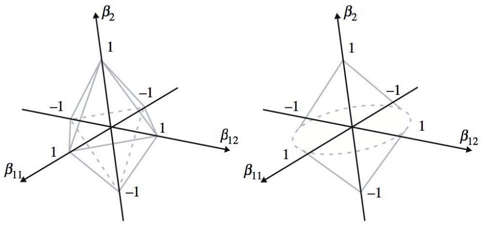

Figure 16.2: Group Lasso

The thought is that by doing this, it is possible to choose some group to be zero entirely, which forces some coefficient to be all zero. And in the group chosen to be non-zero, all the component are non-zero.
To prove the thought, let's look at the KKT condition of this problem. Since there is no constraint, just look at the stationarity condition which respect to every group of $x_{(i)}$.

$$A_{(i)}^T(y-Ax) = \lambda\sqrt{p_{(i)}}s_{(i)}, i = 1, ..., G$$

Where each $s_{(i)} \in \partial\|x_{(i)}\|_2$, i.e.,

$$s_i \in \begin{cases} x_{(i)}/\|x_{(i)}\|_2 & \text{if } x_{(i)} \neq 0 \\ z \in \mathbb{R}^{p_{(i)}} : \|z\|_2 \leq 1 & \text{if } x_{(i)} = 0 \end{cases}, i = 1\text{,...,G}$$

Hence if $\|A_{(i)}^T(y-Ax) = \lambda\sqrt{p_{(i)}}s_{(i)}\|_2 < \lambda\sqrt{p_{(i)}}$, then $x_{(i)} = 0$, else

$$A_{(i)}^T(y-Ax) = \lambda\sqrt{p_{(i)}}s_{(i)} = \lambda\sqrt{p_{(i)}}x_{(i)}/\|x_{(i)}\|_2$$

In this case, we can solve for $x_{(i)}$, and get the following equation:

$$x_{(i)} = (A_{(i)}^T A_{(i)} + \tfrac{\lambda\sqrt{p_{(i)}}}{\|x_{(i)}\|_2}I)^{-1}A_{(i)}^T r_{-(i)}$$

where $r_{-(i)} = y - \sum\limits_{j\neq i} A_{(j)}x_{(j)}$
Hence our previous thought is proved.
This also suggests an algorithm to compute compute group lasso. That is, if the condition holds for $\|A_{(i)}^T(y-Ax) = \lambda\sqrt{p_{(i)}}s_{(i)}\|_2 < \lambda\sqrt{p_{(i)}}$, then we set the group to be zero. And if the condition holds for the second case, then we set:

$$x_{(i)} = (A_{(i)}^T A_{(i)} + \tfrac{\lambda\sqrt{p_{(i)}}}{\|x_{(i)}\|_2}I)^{-1}A_{(i)}^T r_{-(i)}$$

Note that $x_{(i)}$ is on both sides. Iterative method is used in this algorithm. We start with a guess, and plug in $x_{(i)}$ of previous iteration in the right of the equation to compute new $x_{(i)}$.

16.4 Constrained form and Lagrange form

Often in statistics and machine learning we'll switch back and forth between constrained form, where $t \in \mathbb{R}$ is a tuning parameter,

$$\min_{x\in\mathbb{R}^n} f(x) \text{ subject to } h(x) \leq t \tag{C}$$

and **Lagrangian** form, where $\lambda \geq 0$ is a tuning parameter,

$$\min_{x\in\mathbb{R}^n} f(x) + \lambda \cdot h(x) \tag{L}$$

and claim these two forms are equivalent under certain condition. Here we assume that f, h are convex.
To prove that this is true, let's first think about going from constrained form to Lagrange form.
If problem (C) is strictly feasible, then we know that strong duality holds because of the fact that it is strictly convex. Hence there exists some $\lambda \geq 0$, which in this case is the dual solution, such that any solution x^*, which is the primal optimal in (C), is going to minimize

$$f(x) + \lambda \cdot (h(x) - t)$$

This is equivalent to min $f(x) + \lambda \cdot h(x)$ as long as t is not infinite.
Hence we can get that if the problem is convex and strictly feasible, then there exists some $\lambda \geq 0$ such that the constraint problem is the Lagrange problem.
About the other side, let's go from Lagrange form to constrained form.
If we have a solution to

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \cdot h(x)$$

Let's call it x^*. Take $t = h(x^*)$, then go through the KKT conditions for problem:

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } h(x) \leq t$$

The stationarity condition is satisfied because of the fact that we choose x^* to

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \cdot h(x)$$

Which is the same thing to

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \cdot (h(x) - t)$$

Complementary slackness condition, which is $\lambda \cdot (h(x^*) - t)$ is also true because t is defined as $t = h(x^*)$. And primal feasibility condition also holds because we define $t = h(x^*)$. Hence x^* satisfies the KKT conditions, so x^* is also the solution for the problem

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } h(x) \leq t$$

To summarize, we have the following conclusion:

$$\bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} \subseteq \bigcup_{t} \{\text{solutions in (C)}\}$$

$$\bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} \supseteq \bigcup_{t \text{ such that (C) is strictly feasible}} \{\text{solutions in (C)}\}$$

These two forms are almost equivalent, because (C) is not always strictly feasible. However, note that the only value of t that leads to a feasible but not strictly feasible constraint set is $t = 0$, i.e.,

$$x : g(x) \leq t \neq \emptyset, x : g(x) = t = \emptyset \Rightarrow t = 0$$

If $t = 0$ and the problem is not strictly feasible, by setting $\lambda = \infty$, we can enforce the constraint $h(x) = 0$, and achieve the solution in **Lagrange** form.
For example, this is true if g is a norm, and in such situation we do get perfect equivalence. Otherwise it is minor nonequivalent.

16.5 Uniqueness in 1-norm penalized problems

Theorem 16.4 *Let f be differentiable and strictly convex, $A \in \mathbb{R}^{n \times p}, \lambda > 0$. Consider*

$$\underset{x \in \mathbb{R}^p}{min} f(Ax) + \lambda \|x\|_1$$

If the entries of A are drawn from a continuous probability distribution (on $R^{n \times p}$), then with probability 1 the solution $x^* \in \mathbb{R}^p$ is unique and has at most min{n,p} nonzero components.

Here function f must be strictly convex, but there is no restrictions on the dimensions of A (we could have $p \gg n$). Also, this holds for $\forall \lambda > 0$. We can prove this theorem by using the KKT conditions and simple probability argument. **Proof:** The KKT conditions for this problem are:

$$A^T \nabla f(Ax) + \lambda s = 0$$

And we can rewrite this as:

$$-A^T \nabla f(Ax) = \lambda s$$

where $s \in \partial \lambda \|x\|_1$, i.e.,

$$s_i \in \begin{cases} \{sign(x_i)\} & \text{if } x_i \neq 0 \\ [-1, 1] & \text{if } x_i = 0 \end{cases}$$

for $i = 1, 2, ..., n$ First we can note that Ax, s are unique.
Define the set $S = \{j : |A_j^T \nabla f(Ax)| = \lambda\}$, from the KKT conditions we can learn that any solution must satisfies $x_i = 0$ for all $i \notin S$. First, assume that $rank(A_S) < |S|$(here $A \in \mathbb{R}^{n \times |S|}$, which is the submatrix of A corresponding to columns in S). Then for some $i \in S$

$$A_i = \sum_{j \in S\{i\}}$$

For both sides of the above equation, take an inner product with $-\nabla f(Ax)$, we have

$$s_i \lambda = \sum_{j \in S \setminus \{i\}} c_j s_j \lambda$$

and then for both sides of the above equation, multiply by s_i

$$\lambda = \sum_{j \in S \setminus \{i\}} (s_i c_j s_j) \lambda$$

$$1 = \sum_{j \in S \setminus \{i\}} s_i c_j s_j$$

Call $s_i c_j s_j$ as a_j, and we can show that

$$s_i A_i = \sum_{j \in S \setminus \{i\}} a_j s_j A_j$$

where a_j satisfies $1 = \sum\limits_{j \in S \setminus \{i\}} a_j$ This means that $s_i A_i \in aff\{s_j A_j, j \in S \setminus \{i\}\}$ It is straightforward to show that, if the entries of A have a density over $R^{n \times p}$, then A is in general position with probability 1, and the above situation $s_i A_i \in aff\{s_j A_j, j \in S \setminus \{i\}\}$ can't happen with probability 1. The picture below shows this conclusion.

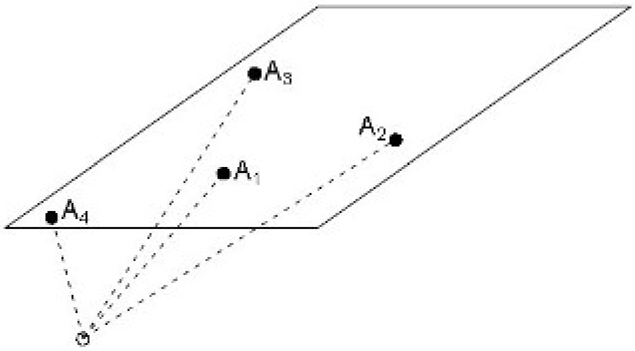

Figure 16.3: Relationships of columns of A

Therefore, if entries of A are drawn from continuous probability distribution, any solution must satisfy $rank(A_S) = |S|$ Since the matrix A_S has the number of columns equal to $|S|$, which is less or equal to p and the number of rows equal to n. Recalling the KKT conditions, this means the number of nonzero components in any solution is $\leq |S| \leq \min n, p$

Furthermore, we can reduce our optimization problem by partially solving. For our problem, plug in 0 for all x_i that $i \notin S$, and we have

$$\min_{x_S \in \mathbb{R}^{|S|}} f(A_S x_S) + \lambda \|x_S\|_1$$

Since A_S has full rank, this problem is strictly convex. Hence the solution in this problem is unique. ■

References

[1] S. Boyd and L. Vandenberghe, "Convex Optimization," *Cambridge University Press*, 2004, Chapter 5.

[2] R. T. Rockafellar, "Convex Analysis," *Princeton University Press*, 1970, Chapters 28-30.