10-725: Optimization **Fall 2012**

Lecture 18: October 25

Lecturer: Ryan Tibshirani *Scribes: Aniruddha Basak, Mingyu Tang, Lingxue Zhang*

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

18.1 Topics covered in this lecture

- Uses of duality
- Dual gradient methods
- Dual decomposition and Augmented Lagrangian

18.2 Brief review of conjugate functions

Recall, the conjugate function of $f : \mathbb{R}^n \to \mathbb{R}$ is defined as,

$$f^*(y) = \max_{x \in \mathbb{R}^n} y^T x - f(x) \tag{18.1}$$

18.2.1 Important properties of the conjugate function

- It appears frequently in dual programs as,

$$-f^*(y) = \min_{x \in \mathbb{R}^n} f(x) - y^T x \tag{18.2}$$

- If f is closed and convex, $f^{**} = f$.
- If f is closed and convex, $x \in \partial f^*(y) \Leftrightarrow y \in \partial f(x)$
- If f is strictly ocnvex, $\nabla f^*(y) = arg\min_{z \in \mathbb{R}^n}(f(z) - y^T z)$

18.3 Uses of Duality

Two key uses of duality have been discussed in the previous lectures as,

- If x is primal feasible and u, v dual feasible, $f(x) - g(u, v)$ is called the duality gap between x and u, v. Zero duality gap implies optimality as,

$$f(x) - f(x^*) \leq f(x) - g(u, v) \tag{18.3}$$

  Similary $g(u^*, v^*) - g$ can be bounded by the duality gap. Thus this gap can be used as a terminating condition for algorithms while solving primal or dual problem.

- If strong duality condition holds, any primal solution minimizes $L(x, u^*, v^*)$ over $x \in \mathbb{R}^n$ where u^*, v^* are given dual solution. This fact can be used to characterize or compute primal solutions.
  Note any minimizer of $L(x, u^*, v^*)$ is not necessarily a primal solution. However, if there is only one minimizer, it is the primal solution.

18.4 Image of "Lasso" in the mirror of *Duality*

We know the Lasso problem is defined as follows,

$$\min_{x \in \mathbb{R}^p} \frac{1}{2}||y - Ax||^2 + \lambda||x||_1 \tag{18.4}$$

As we are interested in applying the principles of duality, lets rewrite this unconstrained optimization formulation as a constrained problem.

$$\min_{x \in \mathbb{R}^p, z \in \mathbb{R}^n} \frac{1}{2}||y - z||^2 + \lambda||x||_1 \qquad \text{subject to} \quad z = Ax \tag{18.5}$$

As $z \in \mathbb{R}^n$, here are n constraints and thus we construct the Lagrangian using n multipliers as below,

$$\mathcal{L}(x, z, u) = \frac{1}{2}||y - z||^2 + \lambda||x||_1 + u^T(z - Ax) \tag{18.6}$$

Now we can use the KKT stationarity condition which says, at the optimal point the subgradient of $\mathcal{L}(x, z, u)$ with respect to x and z must contain 0. As $\mathcal{L}(x, z, u)$ is differentiable in z, the subgradient with respect to z is unique and equal to the gradient. Thus we have,

$$\begin{aligned} & \frac{\partial \mathcal{L}}{\partial z} = 0 \\ \Rightarrow & \frac{1}{2} \cdot 2(z - y) + 0 + u^* = 0 \\ \Rightarrow & z = y - u^* \\ \Rightarrow & Ax^* = y - u^* \end{aligned} \tag{18.7}$$

Here u^* is the solution to the dual problem and x^* for the primal. Taking the subgradient with respect to x we get,

$$\begin{aligned} & 0 \in 0 + \lambda \partial||x||_1 - ({u^*}^T A)^T \\ \Rightarrow & \lambda s = A^T u^* \qquad \text{where } s \in \partial||x||_1 \end{aligned} \tag{18.8}$$

Recall, from the definition of subgradient and L_1 norm we get

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

Thus equation 18.8 gives,

$$A_i^T u^* \in \begin{cases} \{\lambda\} & \text{if } x_i^* > 0 \\ \{-\lambda\} & \text{if } x_i^* < 0 \\ [-\lambda, \lambda] & \text{if } x_i^* = 0 \end{cases} \tag{18.9}$$

This is a nice relation between primal and dual solution and it says $|A_i^T u^*| < \lambda$ implies $x_i^* = 0$. Thus we can characterize the primal solution looking at the dual solution.

We can get a lot more than this from the "dual image". Lets see..

Now we consider the dual formulation of the Lasso problem which is,

$$\min_{u\in\mathbb{R}} ||y-u||^2 \quad \text{subject to } ||A^T u||_\infty \leq \lambda \tag{18.10}$$

This can be rewritten as,

$$\min_{u\in\mathbb{R}} ||y-u||^2 \quad \text{subject to } u \in C \text{ where } C = \{v : ||A^T v||_\infty \leq \lambda\} \tag{18.11}$$

It is easy to see, the solution u^* is the projection of y onto the set C. That is,

$$\begin{aligned} u^* &= P_C(y) \\ \Rightarrow y - u^* &= y - P_C(y) \\ \Rightarrow Ax^* &= (I - P_C)(y) \end{aligned} \tag{18.12}$$

This equation connects the lasso fit (solution to the primal) to the residual from projecting y onto C. To visualize this, lets take a closer look to C.

$$C = \bigcap_{i=1}^{p} [\{u : A_i^T u \leq \lambda\} \bigcap \{u : A_i^T u \geq \lambda\}] \tag{18.13}$$

We observe C is not just any set, but a special kind of set called "polyhedron". Thus the lasso fit is actually the residual from the projection of y onto a polyhedron !
This fact is not as near obvious from the primal problem as it is from the primal-dual relationship.

Now lets look into the geometry. But before that we rewrite C as,

$$\begin{aligned} C &= \{u : ||A^T u|| \leq \lambda\} \\ &= \{u : A^T u \in \{v : ||v||_\infty \leq \lambda\}\} \\ &= (A^T)^{-1}(\{v : ||v||_\infty \leq \lambda\}) \end{aligned}$$

Note that $v \in \mathbb{R}^p$ and $u \in \mathbb{R}^n$. So the polyhedron C (in $\mathbb{R}^n$) is the inverse image of the hypercube $\{v : ||v||_\infty \leq \lambda\}$ (in $\mathbb{R}^p$) under the linear map A^T. The picture is shown in Figure 18.2.

Now we can get a few properties of the lasso fit from the geometry.

- Considering the lasso fit as a function of y, we can say it is nonexpansive with respect to y. That is,

  $$||Ax^*(y) - Ax^*(y')|| \leq ||y - y'|| \quad \forall y, y' \tag{18.14}$$

  The proof follows from the property of projection onto convex sets, that is the distance of the projected points can not be larger than the original points.

- From Figure 18.2 we observe, each face of the hypercube is mapped to a face of the polyhedron C under a linear transformation. Thus each face of C corresponds to a particular active set of lasso solutions.

- If a small perturbation is applied to y, almost every time it will be projected to same face. Thus the active set remains the same in most of the cases.

- The lasso solution is locally constat and the lasso fit is locally affine projection map.

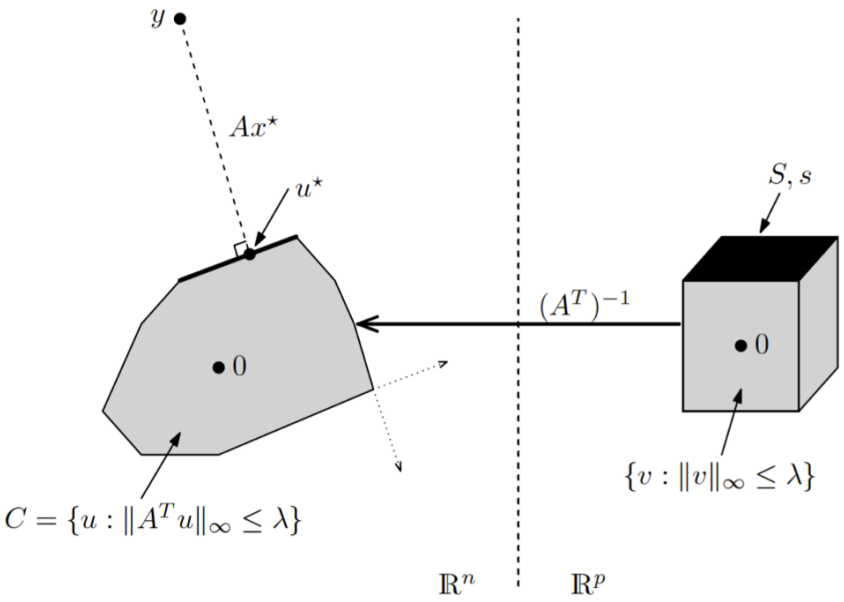

Figure 18.1: Lasso and projection onto polyhedra

18.4.1 Safe rules

In 2010, L. El Ghaoui et al., showed an amazing safe rule for eliminating features apriori in the article, "Safe Feature Elimination in Sparse Supervised Learning". The rule says,

$$|A_i^T y| < \lambda - ||A_i|| \cdot ||y|| \frac{\lambda_{max} - \lambda}{\lambda} \Rightarrow x_i^* = 0 \quad \forall i = 1, 2, ..., p \tag{18.15}$$

This rule comes from the lasso dual which is

$$\max_{u \in \mathbb{R}^n} g(u) \quad \text{subject to } ||A^T u||_\infty \leq \lambda \tag{18.16}$$

where $g(u) = \frac{1}{2}(||y||^2 - ||y - u||^2)$. Suppose, $u_0 = y \cdot \lambda / \lambda_{max}$ is a dual feasible point. Hence, $\gamma = g(u_0)$ is a lower bound to the dual optimal value. Thus we can rewrite 18.16 as

$$\max_{u \in \mathbb{R}^n} g(u) \quad \text{subject to } ||A^T u||_\infty, g(u) \geq \gamma \tag{18.17}$$

Now we compute,

$$m_i = \max\{\max A_i^T u \quad \text{s.t. } g(u) \geq \gamma, \quad \min A_i^T u \quad \text{s.t. } g(u) \geq \gamma\} \tag{18.18}$$

We have just shown from the KKT conditions,

$$m_i < \lambda \Rightarrow |A_i^T u^*| < \lambda \Rightarrow x_i^* = 0$$

Now,

$$m_i < \lambda \Rightarrow |A_i^T y| < \lambda - \sqrt{||y||^2 - 2\gamma} \cdot ||x|| \tag{18.19}$$

Substituting $\gamma = g(y \cdot \lambda / \lambda_{max})$, gives 18.16.

In cases other than lasso where m_i cannot be expressed as a close form as in 18.18, we can upperbound this analytically or try to solve this optimization problem. Is this problem is cheaper in terms of computational cost, we can solve this and discard few features from the original problem apriori.
Note that the safe rule does not give the sufficient condition for $x_i^* = 0$. That is only a few among all possible features can be eliminated by this rule.

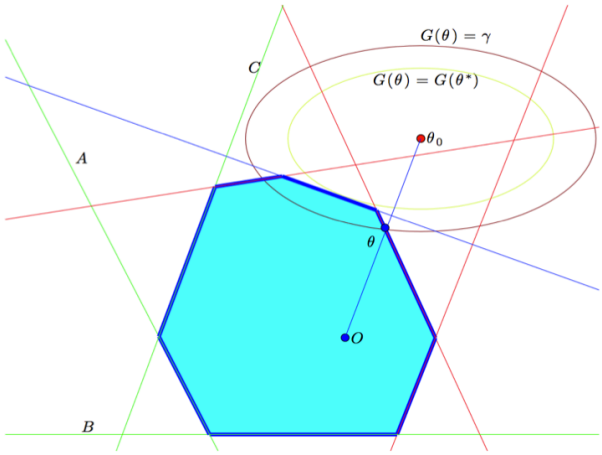

Figure 18.2: Safe rule to eliminate features in Lasso

18.5 Beyond pure sparsity

Lets consider the reverse lasso problem,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}||y - x||^2 + \lambda||Dx||_1 \tag{18.20}$$

where $D \in \mathbb{R}^{m \times n}$ is a given penalty matrix. Note that, if D is invertible, the problem can be transformed into a lasso problem. However, if $rank(D) < m$, this cannot be done.
In the later case, this problem is harder to solve compared to lasso. Observe that we cannot use generalized gradient descent in this case as the prox function would be same as the optimization problem.
The idea behind this reverse lasso problem is that the sparsity of Dx^* implies some structure in x^*.
For instance, in fused lasso, D is chosen as follows,

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & & 0 \\ 0 & -1 & 1 & & & \\ 0 & 0 & -1 & & & \\ & \dots & & .. & & \\ 0 & \dots & & & -1 & 1 \end{bmatrix} \tag{18.21}$$

Thus,

$$||Dx||_1 = \sum_{i=2}^{n} |x_i - x_{i-1}| \tag{18.22}$$

We can see that if lot of elements in Dx^* are 0, lot of x_i and x_j are equal which is a special structure.
Now we can follow the same steps as in lasso dual derivation to get the primal-dual relationship. This time lets look at the geometry first.

Evidently from Figure 18.3, $D^T u^* = P_C(y)$, where $C = \{D^T u : ||u||_\infty \lambda\}$ and $x^* = (I - P_C)(y)$.
As before, similar arguments can be drawn from the geometry.

- Primal solution x^* is 1-Lipschitz continuous as a function of y.
- Each face of polyhedron C corresponds to a nonzero pattern in Dx^*.

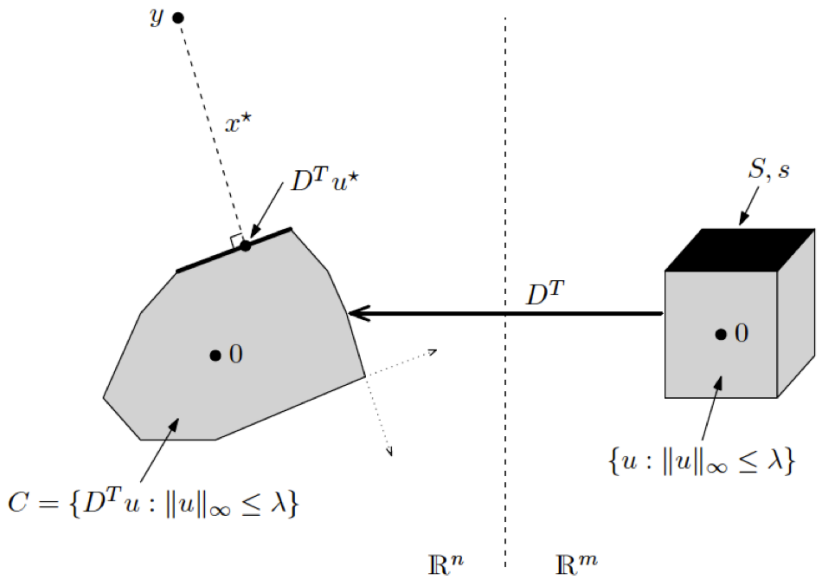

Figure 18.3: Reverse lasso and projection onto polyhedra

- Small perturbations on y will almost everytime result in same structure $S = support(Dx^*)$ and thus is a locally affine projection map.

Lets take a closer look at the dual problem one last time.

$$\min_{u \in \mathbb{R}^m} ||y - D^T u||^2 \quad \text{s.t. } ||u||_\infty \leq \lambda \tag{18.23}$$

We observe the dual disentangles the involvement of linear operator D in the 1-norm. Thus it is much easier to solve the dual problem compare to the primal counterpart. The prox function in the dual is as follows,

$$prox(z) = \min_u ||z - u|| \quad \text{s.t. } ||u||_\infty \leq \lambda \tag{18.24}$$

This is the projection of z onto the ∞-norm ball and very easy to compute. Hence generalized gradient descent or accelerated generalized gradient method can be used to solve the dual problem.

18.6 Dual Gradient Method

Even if we cannot get the closed form of the dual problem, we can still use dual-based gradient and sub-gradient method.

E.g.: Consider the primal problem:

$$min_{x \in R^n} f(x) \; subject \; to \; Ax = b$$

Then the dual problem is:

$$max_{u \in R^m} f^*(-A^T u) - b^T u$$

where f^* is the conjugate of f. Define $g(u) = f^*(-A^T u)$, then $\partial g(u) = -A \partial f^*(-A^T u)$
Recall that $x \in \partial f^*(-A^T u) \Leftrightarrow x \in argmin_{z \in R^n} f(z) + u^T A z$
Thus, we derive the dual subgradient method and dual gradient method

18.6.1 Dual Subgradient Method

Starts with an initial dual guess $u^{(0)}$, and repeats for k = 1,2,3...

$$x^{(k)} \in argmin_{x \in R^n} f(x) + (u^{k-1})^T Ax$$

$$u^{(k)} = u^{(k-1)} + t_k(Ax^{(k-1)-b})$$

where t_k is step size.

18.6.2 Dual gradient Method

Similarly, if f is strictly convex, then $f(*$ is differentiable. Then the dual method is:

$$x^{(k)} = argmin_{x \in R^n} f(x) + (u^{k-1})^T Ax$$

$$u^{(k)} = u^{(k-1)} + t_k(Ax^{(k-1)-b})$$

Here, $x^{(k)}$ is unique.

18.6.3 Dual Method Analysis

Lemma 18.1 *If f is strongly convex with parameter d, then $\bigtriangledown f^*$ is Lipschitz with parameter 1/d.*

Proof: If f strongly convex and x is its minimizer, then

$$f(y) \geq f(x) + \frac{d}{2}||y - x||, \quad ally$$

Hence defining $x_u = \bigtriangledown f^*(u), x_v = \bigtriangledown f^*(v)$,

$$f(v) - u^T x_v \geq f(u) - u^T x_u + \frac{d}{2}||x_u - x_v||^2$$

$$f(u) - v^T x_u \geq f(v) - v^T x_v + \frac{d}{2}||x_u - x_v||^2$$

Adding these together:

$$d||x_u - x_v||^2 \leq (u - v)^T(x_u - x_v)$$

Use Cauchy - Schwartz and rearrange, $||x_u - x_v|| \leq (1/d)||u - v||$

■

In summary

- Converge Rate: If f is strongly convex with parameter d, then the dual gradient ascent with constant step size $t_k < d$ converges at rate $O(1/k)$
- Advantage: Decomposability
- Disadvantage: (1) slow to converge (2) Poor convergence properties: even though we may achieve convergence in dual objective value, convergence of $u^{(k)}, x^{(k)}$ requires strong assumptions.

18.6.4 Dual Decomposition

Consider

$$min_{\in R^n} \sum_{i=1}^{B} f_i(x_i) \ subject \ to Ax = b$$

Here $x = (x_1, ...x_B)$ is division into B blocks of variables, so each $x_i \in R^{n_i}$. We can also partition A accordingly

$$A = [A_1, ...A_B], \ where \ A_i \in R^{m*n_i}$$

Then we can decompose the update of x^+ as:

$$x^+ \in argmin_{x \in R^n} \sum_{i=1}^{B} f_i(x_i) + u^T Ax \Leftrightarrow x_i^+ \in argmin_{x \in R^{n_i}} f_i(x_i) + u^T A_i x_i, \ for \ i = 1, ...B$$

i.e., minimization decomposes into B separate problems. Thus, the entire **Dual decomposition algorithm** is:

$$x_i^k \in argmin_{x_i \in R^{n_i}} f_i(x_i) + (u^{(k-1)})^T A_i x_i, i = 1, ...B$$

$$u^{(k)} = u^{(k-1)} + t_k(\sum_{i=1}^{B} A_i x_i^{(k-1)} - b)$$

18.7 Augmented Lagrangian

We can add one extra term on the primal problem with the constraint that the extra term is 0 (assuming A has full rank):

$$\min_{x \in \mathbb{R}^n} f(x) + \frac{\rho}{2}||Ax - b||^2$$
$$\text{subject to } Ax = b$$

The problem solution is not changed. But now the primal objective is strongly convex with parameter $\rho\sigma_{min}^2(A)$. The dual is differentiable and we can use dual gradient ascent updates:

$$x^{(k)} = \arg\min_{x \in \mathbb{R}^n} f(x) + (u^{(k-1)})^T Ax + \frac{\rho}{2}||Ax - b||^2$$

$$u^{(k)} = u^{(k-1)} + \rho(Ax^{(k-1)} - b)$$

The first update is equivalent to saying 0 is in the set of the sub gradient $\partial f(x^{(k)}) + A^T u^{(k-1)} + \rho(Ax^{(k-1)} - b) = \partial f(x^{(k)}) + A^T u^{(k-1)} + \rho Ax^{(k-1)}$, which is the stationary condition of the primal problem. It can be shown that $Ax^{(k)} - b$ approaches zero with this update rule. The KKT conditions are satisfied as k approaches ∞ and $x^{(k)}$ and multiplier $u^{(k)}$ approach optimality.

The convergence property is better but the problem is no longer decomposable because the augmented term $\frac{\rho}{2}||Ax - b||^2$ couples the variables together.

References

[CW87] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *Proceedings of the 19th ACM Symposium on Theory of Computing*, 1987, pp. 1–6.