Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Soumye Singhal
Instructor: Purushottam Kar
Date: March 26, 2018

LECTURE
20

Alternate Minimization 2

1 Introduction

In the previous lecture we looked at the basics of Alternate Minimization. We also looked at the specific example of Clustering and finding the leading eigenvectors of a Matrix using the Power Method. We also began to look at coordinate minimization/descent. Today we continue with that discussion.

2 Coordinate Descent

It is an algorithm to minimize a function f in d dimensions by iteratively optimization over a single dimension while keeping the others fixed. Formally our objective and algorithm is as follows:

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

Algorithm 1: Coordinate Descent

for $t = 1, 2, \ldots$ **do**
  Choose Coordinate $I_t \in [d]$
  $v^{t+1} \leftarrow \arg\min_{v \in \mathbb{R}^d} f(x_1^t, \cdots, x_{I_{t-1}}^t, v, x_{I_{t+1}}^t, \cdots, x_d^t)$
  $x_{\overline{I_t}}^{t+1} \leftarrow x_{\overline{I_t}}^t$ //$\overline{I_t}$ `denotes all but` I_t `dimension`
  $v_{I_t}^{t+1} \leftarrow v^{t+1}$
end for

There are multiple ways that people adopt to choose the coordinate which were discussed in the previous lecture. We can see that our original problem of optimizing the function over d dimensions gets simplified to optimizing over one dimension(or a block of dimension as we will see later) which is less expensive and feasible in most cases. Sometimes we may face an issue if it is expensive to even evaluate the function f.
The solution to the sub-problem can also sometimes have a closed form solution. For example, in the case of SVM, the problem reduces to an quadratic form which has a closed form solution.

SVM-Dual Form

Let us look at the SVM Dual Objective Function(without the bias term in the linear classifier)

$$\max_{\alpha_i \in [0,c]} \sum_i \alpha_i - \boldsymbol{\alpha}^T Q \boldsymbol{\alpha}$$

Here $\boldsymbol{\alpha} \in \mathbb{R}^N$ and N is the number of data points and $Q_{ij} = \frac{1}{2} y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$. If we apply coordinate descent, our objective function at each step reduces to a constraint quadratic function.

$$\max_{\alpha_{I_t} \in [0,c]} \alpha_{I_t} - \alpha_{I_t}^2 Q_{I_t I_t} - 2\alpha_{I_t} \left\langle Q_{I_t}, Q_{I_t} \right\rangle$$

Note that we have used the inner product notation, $\langle , \rangle$ even though it's a simple product so as to generalize for a block of coordinates as discussed later.
Now, Since the quadratic term has a negative coefficient this means that the function is basically an inverted parabola. So if the unconstrained maximum lies in the constraint, we get the solution otherwise the solution lies on the boundary. So we get an almost closed form solution.
Note the form of the constraint in the above solution. Here the constraints on α_i are independent of one another. Hence these are called box-constraints.
We would have problems carrying out coordinate descent if the constraints are interrelated. Consider the following example

Example 20.1. (SVM with bias terms)
The objective function is of form

$$\max_{\alpha \in \mathbb{R}^n} \boldsymbol{\alpha}^T - \boldsymbol{\alpha}^t Q \boldsymbol{\alpha}$$

subject to the conditions $\boldsymbol{\alpha} \in [0, c]^n$ and $\sum_i \alpha_i y_i = 0$.
Platt (1998) Overcame this problem of solving this by changing the notion of coordinates. They took pairs of coordinates (i, j) and then solved the 2D problem. Note that choosing a pair of coordinates helps us to satisfy the condition $\sum_i \alpha_i y_i = 0$ which would not have been possible in the one dimension case.

Remark 20.1. Block Coordinate Descent : This is a generalization of the above stated idea of optimizing over pair of coordinates. The main idea is that at each time step instead of choosing a single dimension I_d to optimize the function over, we instead choose a block $\mathbf{B}_t$ of coordinates s.t $\mathbf{B}_t \subset [d]$ and $|\mathbf{B}_t| = k$. We keep k relatively small so that the problem is still solvable.

Exercise 20.1. Analyse Coordinate Descent for the case of Ridge Regression.

Next, we look at how Coordinate Descent is related to Gradient Descent. But before proceeding let look at Fenchel Duals.

3 Fenchel Duality

Duality is an alternate representation of a function or an objective. Mostly these alternate representations are equivalent to the original one and are also more usable, a very common example being Fourier transforms.
Fenchel Duals are an alternate representation for functions. We normally use an input output based definition or a function which is basically a recipe or an algorithm to compute a functions value. The Fenchel Dual is a an alternate representation of the form (slope, y-intercept) which

means that it maps a slope to it's corresponding y-intercept for the given function. For a convex function this alternate representation is equivalent to the normal one because slopes never repeat in a convex function i.e the slopes are monotonic. More formally, for every function f, we define it's Fenchel Dual f^* : 'slopes' $\to \mathbb{R}$ (y-intercept)

$$f^*(g) = \max_{x \in \mathcal{X}} \{\langle x, g \rangle - f(x)\}$$

Remark 20.2. f^* is always a convex function irrespective of whether f was convex or not. This is because max of a set of linear functions is convex.

Remark 20.3. If f is a convex function then the biconjugate $(f^*)^* = f$. This means that there is no loss of information on taking the conjugate of f.
On the other hand if f is a non-convex function then the biconjugate $(f^*)^* \neq f$. Instead $(f^*)^*$ gives the lower envelope of f ie the tightest convex function approximation to f.

Exercise 20.2. Consider the function

$$\mathbb{I}\{x\} = \begin{cases} 0 & \|x\| \leq 1 \\ \infty & otherwise \end{cases}$$

Now, prove that the Fenchel dual of $\mathbb{I}\{x\}$ is the dual norm.

Exercise 20.3. Check if $\arg\min f \stackrel{?}{=} \arg\min (f^*)^*$

3.1 Use of Fenchel Dual

Fenchel Dual is used to make convex, non-smooth functions smooth (eg hinge loss).

Remark 20.4. A function can't be made smooth just by adding something unlike strong convexity. For example we can use regularization to make a convex function Strongly Convex. Consider a convex function $f(x)$, then $f(x) + \frac{\lambda}{2}\|x\|_2^2$ is $\lambda-$strongly convex.

Smoothness is of use to us because it implies that the function is differentiable everywhere which in turn ensures faster rates of convergence and possibility of using accelerated descent methods like NAG.
Now we state an curious property of Fenchel Duals without proof.

Theorem 20.1. f is $\alpha - SC$ iff f^* is $\frac{1}{\alpha}$S.S.

Now, suppose we want to make a general non-convex, non-smooth function $f(x)$ smooth, then we proceed with the following transformations

$$f(x) \to \left(f^*(g) + \frac{\alpha}{2}\|g\|_2^2\right)^* \to \tilde{f}(x)$$

Note that $f^*(g)$ will be convex and hence $\left(f^*(g) + \frac{\alpha}{2}\|g\|_2^2\right)$ will be α Strongly Convex. Then using the theorem we conclude that $\tilde{f}(x)$ is smooth.
A common example of such conversion is converting the hinge loss to the smooth huber loss. This procedure of making loss functions smooth is called Huberization of loss-functions.

Exercise 20.4. Prove that $\left|\tilde{f}(x) - f(x)\right| \leq \mathcal{O}(\alpha)$

4 Deriving Lagrangian Dual of Problems

Now we look at how Fenchel Dual helps us in deriving a dual formulation for some optimization problems. Consider an optimization problem with regularization term,

$$\min_w \sum_i \phi(w^T z^i) + \frac{\lambda}{2} \|w\|_2^2$$

Here ϕ may be some loss function. Recall that the usual way we derived the dual to a problem was using the constraints to set the Lagrangian. Since, there are no constraints here, we introduce dummy constraints.

$$\begin{aligned} \min_w \sum_i \phi(\xi_i) + \frac{\lambda}{2} \|w\|_2^2 \\ \text{such that} \quad \xi_i = w^T z^i \end{aligned}$$

Now, we construct the Lagrangian and formulate the primal form of the problem as follows:

$$\min_{w,\xi_i} \left\{ \max_{\alpha_i} \sum_i \phi(\xi_i) + \frac{\lambda}{2} \|w\|_2^2 + \sum_i \alpha_i \left(w^T z^i - \xi_i \right) \right\}$$

The corresponding dual form of the problem is then:

$$\max_{\alpha_i} \left\{ \min_{w,\xi_i} \sum_i \phi(\xi_i) + \frac{\lambda}{2} \|w\|_2^2 \sum_i \alpha_i \left(w^T z^i - \xi_i \right) \right\}$$

Note that the constraint variables w, ξ_i don't talk to each other, hence we can separate the terms as follows.

$$\max_{\alpha_i} \left[\min_w \left(\frac{\lambda}{2} \|w\|_2^2 + w^T \sum_i \alpha_i z^i \right) + \sum_i \min_{\xi_i} \left(\phi(\xi_i) - \alpha_i \xi_i \right) \right]$$

Note that we can use the fact that $\min f(x) = -\max -f(x)$ to simplify the above expression as

$$\max_{\alpha_i} \left[\min_w \left(\frac{\lambda}{2} \|w\|_2^2 + w^T \sum_i \alpha_i z^i \right) - \sum_i \max_{\xi_i} \left(-\phi(\xi_i) + \alpha_i \xi_i \right) \right]$$

Now, note that the term $(\alpha_i \xi_i - \phi(\xi_i))$ is the Fenchel Dual of ϕ. Also solving for the $\min_w$ term the expression simplifies to the following.

$$\max_{\alpha_i} \left[\frac{-1}{2\lambda} \boldsymbol{\alpha}^T Z \boldsymbol{\alpha} - \sum_i \phi^* \left(\alpha_i \right) \right]$$

Exercise 20.5. Verify that the Fenchel Dual of the hinge-loss is

$$h^*(\alpha) = \begin{cases} -\infty & \alpha \notin [0, c] \\ \alpha & \alpha \in [0, c] \end{cases}$$

5 Choosing the coordinate

At each step of the coordinate descent, there are multiple ways of choosing the coordinate. For example,

- Random coordinate at each step.
- Cyclic Order : It is great to minimize disk-thrashing. For details see Wright and Lee (2017)
- Random Cyclic: Take a random permutation and then cycle over it.
- Gauss-Southwell: $I_t \in \arg\max_i \left|\nabla_i f(x^t)\right|$

6 Analysis of Coordinate Descent on Marginal SS Functions

Definition 20.1. A function f is called Marginally Strongly Smooth in it's i^{th} variable if

$$f(x + \lambda e_i) \leq f(x) + \langle \nabla_i f(x), \lambda e_i \rangle + \frac{\beta \lambda^2}{2}$$

Consider a function f which is Marginally Strongly Smooth in all it's dimension (Note that this doesn't imply Strong Smoothness) and has global Convexity. Coordinate Descent is helpful to us in this case as marginal problems would be easier to solve due to marginal smoothness. Consider the t-th stage in Coordinate Descent. We got x^{t+1} by complete minimization over the chose coordinate I_t. Now we use marginal strong smoothness between $t+1$ and t time steps.

$$f(x^{t+1}) \leq f(x^t) + \left\langle \nabla_{I_t} f(x^t), x^{t+1} - x^t \right\rangle + \frac{\beta}{2} \left\| x^{t+1} - x^t \right\|_2^2$$

Since we have $x^{t+1} = x^t + \lambda e_{I_t}$, we get

$$f(x^t + \lambda e_{I_t}) \leq f(x^t) + \left\langle \nabla_{I_t} f(x^t), \lambda e_{I_t} \right\rangle + \frac{\beta \lambda^2}{2}$$

Taking min over λ on both sides we get

$$\min_\lambda f(x^t + \lambda e_{I_t}) \leq \min_\lambda \left[f(x^t) + \left\langle \nabla_{I_t} f(x^t), \lambda e_{I_t} \right\rangle + \frac{\beta \lambda^2}{2} \right]$$

Note that by definition, $f(x^{t+1}) = \min_\lambda f(x^t + \lambda e_{I_t})$. Also, the minimum of RHS is obtained for $\lambda = -\dfrac{\nabla}{\beta}$. Substituting we get,

$$f(x^{t+1}) \leq f(x^t) - \frac{1}{2\beta} \left\| \nabla_{I_t} f(x^t) \right\|_2^2$$

Since we have global convexity, we want our result to be in the form of global ∇ and not ∇_{I_t}. By pigeonhole principle largest $\geq$ avg. Since we choose our I_t such that $I_t \in \arg\max_i \left|\nabla_i f(x^t)\right|$, we get

$$\frac{1}{2\beta} \left\| \nabla_{I_t} f(x^t) \right\|_2^2 \geq \frac{\frac{1}{2\beta} \left\| \nabla f(x^t) \right\|_2^2}{d}$$

$$\implies f(x^{t+1}) \leq f(x^t) - \frac{1}{2\beta} \left\| \nabla_{I_t} f(x^t) \right\|_2^2 \leq f(x^t) - \frac{1}{2\beta d} \left\| \nabla f(x^t) \right\|_2^2$$

Now we use global convexity to lower bound the $\frac{1}{2\beta d}\left\|\nabla f(x^t)\right\|_2^2$, so as to ensure that the function value decreases by at-least a certain amount in each time step. By convexity, we have

$$f^* \geq f(x^t) + \left\langle \nabla f(x^t), x^* - x^t \right\rangle$$

Choose an R such that $R \geq \left\|x^0 - x^*\right\|_2 \geq \left\|x^t - x^*\right\|_2$, we get the following by applying Cauchy-Swartz.

$$\|\nabla f\| R \geq \left\langle \nabla f(x^t), x^t - x^* \right\rangle > \Phi_t$$

where $\Phi_t = f(x^t) - f(x^*)$ is the potential function we want to show decreases.
Now, combining global convexity with marginal SS, the recursive formulae for Φ_t we get is

$$\Phi_{t+1} \leq \Phi_t - \frac{\Phi_t^2}{2\beta d R^2}$$

Let $c = \frac{1}{2\beta d R^2}$. Then,

$$\Phi_{t+1} \leq \Phi_t - c\Phi_t^2$$

Notice that the above recurrence closely corresponds to a differential equation of the form

$$df(x) = -f(x)^2$$

A possible solution to the above equation is the polynomial $f(x) = 1/x$. Since, the decrease in potential $(\Phi_t - \Phi_{t+1})$ is proportional to Φ_t^2, the potential cannot remain large for long. Since, Φ is a decreasing function,

$$\begin{aligned}
&\Phi_t \geq \Phi_{t+1} \\
\implies &\Phi_t - \Phi_{t+1} \geq c\Phi_t^2 \geq c\Phi_t\Phi_{t+1} \\
\implies &\frac{1}{\Phi_{t+1}} - \frac{1}{\Phi_t} \geq c \\
\implies &\frac{1}{\Phi_t} + c \leq \frac{1}{\Phi_{t+1}} \\
\implies &\frac{1}{\Phi_t} - \frac{1}{\Phi_1} = \sum_{i=2}^{t}\left[\frac{1}{\Phi_i} - \frac{1}{\Phi_{i-1}}\right] \geq (t-1)c
\end{aligned}$$

Hence $\Phi_T = \mathcal{O}\left(\frac{1}{T}\right)$. This shows that convex function with marginal strong smoothness converges.

References

John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

S. J. Wright and C.-p. Lee. Analyzing Random Permutations for Cyclic Coordinate Descent. *ArXiv e-prints*, June 2017.