10-725: Optimization **Fall 2012**

Lecture 19: October 30

Lecturer: Aaditya Ramdas *Scribes: Mu Li, Minli Xu*

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.*

19.1 ADMM

19.1.1 Dual (Decomposition) Ascend

Consider solving the following problem

$$\min_x f(x) \quad \text{s.t.} \quad Ax = b \tag{19.1}$$

We know its Lagrangian form is

$$L(x, u) = f(x) + u^T(Ax - b) \tag{19.2}$$

and the Lagrange dual function is

$$g(u) = \inf_x L(x, u). \tag{19.3}$$

So the dual problem of (19.2) is

$$\max_u g(u). \tag{19.4}$$

A natural way to solve (19.4) is using subgradient ascent. Given u, assume x^+ minimizes $L(x, u)$, then the subgradient is $\partial g(u) = Ax^+ - b$. Choosing learning rate sequence $\eta_1, \ldots,$, the subgradient ascent is defined as following:

For time $t = 1, \ldots,$

$$x^{t+1} = \arg\min_x L(x, u^t) \tag{19.5}$$

$$u^{t+1} = u^t + \eta^t(Ax^{t+1} - b) \tag{19.6}$$

If strongly duality holds, and u^* is the optimal solution of the dual problem (19.4), then the optimal primal point can be computed as $x^* = \arg\min_x L(x, u^*)$.

For appropriate learning rate η^t and under certain conditions, x^t, u^t converge to optimal primal and dual points, respectively. However, when g is not differentiable, we cannot assure the ascent of the dual objective value for each iteration, namely $g(u^{t+1}) \not\geq g(u^t)$.

Now assume the objective function $f(x)$ is decomposable, that is, it can be written as $f(x) = \sum_i f_i(x_i)$, where $x_i \in \mathbb{R}^{n_i}$ are disjoint sets subject to $x = \{x_1, \ldots, x_N\}$. Denote by $Ax = \sum_i A_i x_i$, we rewrite the Lagrange form (19.2) as

$$L(x, u) = \sum_i L_i(x_i, u) = \sum_i \left(f_i(x_i) + u^T A_i x_i - \frac{1}{N} u^T b \right) \tag{19.7}$$

For fixed u, (19.7) consists of N independent components, which can be optimized simultaneously. So the dual ascent algorithm turns to be

$$x_i^{t+1} = \arg\min_{x_i} L_i(x_i, u^t) \text{ for } i = 1, \ldots, N \tag{19.8}$$

$$u^{t+1} = u^t + \eta^t(Ax^{t+1} - b) \tag{19.9}$$

Comparing to (19.5), which minimize L on the whole x, now we solve N smaller problems.

19.1.2 Augmented Lagrangian

Augmented Lagrangian explicitly adds an additional strictly convex term to make the problem easier for solving. Choose $\rho > 0$, it has the form:

$$L_\rho(x, u) = f(x) + u^T(Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2. \tag{19.10}$$

The additional term equals to 0 at the optimal point x^*, since we have $Ax^* = b$. So it does not affect the result.

Now the associated dual function is

$$g_\rho(u) = \min_x L_\rho(x, u), \tag{19.11}$$

and the dual ascent comes to

$$x^{t+1} = \arg\min_x L_\rho(x, u^t) \tag{19.12}$$

$$u^{t+1} = u^t + \rho(Ax^{t+1} - b) \tag{19.13}$$

Easy to see $L_\rho(x, u)$ is strictly convex with respect to x, even though it is not true for the original $L(x, u)$. However, $\|Ax - b\|_2^2$ cannot be decomposed as a serial of functions on x_i, so the decomposition trick we used before cannot be applied here.

19.1.3 ADMM

ADMM extends the decomposition idea to augmented Lagrangian. It iteratively solves a smaller problem with respect to x_i by fix the variable x_j for $j \neq i$. We consider the case $N = 2$ for simplicity; it can be easily extend to general N. Now $f(x)$ has form $f(x) = f_1(x_1) + f_2(x_2)$ and the augmented Lagrangian is

$$L_\rho(x_1, x_2, u) = f_1(x_1) + f_2(x_2) + u^T(A_1x_1 + A_2x_2 - b) + \frac{\rho}{2}\|A_1x_1 + A_2x_2 - b\|_2^2. \tag{19.14}$$

ADMM solves each direction alternatively

$$x_1^{t+1} = \arg\min_{x_1} L_\rho(x_1, x_2^t, u^t) \tag{19.15}$$

$$x_2^{t+1} = \arg\min_{x_2} L_\rho(x_1^{t+1}, x_2, u^t) \tag{19.16}$$

$$u^{t+1} = u^t + \rho(A_1x_1^{t+1} + A_2x_2^{t+1} - b) \tag{19.17}$$

On the first step, we fix x_2 and u by values on the last iteration and obtain a new x_1 by solving the smaller problem with respect to only x_1. Next we use the new x_1 to obtain new x_2. Finally we update the dual variable u. Different to the dual decomposition ascent, ADMM updates x_i sequentially. The reason is that the additional augment term makes we can not decompose the Lagrangian form into N conditionally independent components (conditioned on u) as we did on (19.7). So solving step 1 and 2 simultaneously may give different results than solving it sequentially. The latter strategy using the most up-to-date information about x_i potentially accelerates the convergence.

Given assumptions:

1. Function f_1, f_2 are closed, proper, and convex (which means their according epigraphs are closed, nonempty, and convex)
2. The un-augmented Lagrangian $L_0(x_1, x_2, u)$ has saddle points x_1^S, x_2^S and u^S subject to

$$L_0(x_1^S, x_2^S, u) \le L_0(x_1^S, x_2^S, u^S) \le L_0(x_1, x_2, u^S) \tag{19.18}$$

Then when $t \to \infty$, we have

Residual convergence: $r^t = A_1 x_1^t + A_2 x_2^t - b \to 0$

Objective convergence: $f_1(x_1^t) + f_2(x_2^t) \to f^*$

Dual variable convergence: $u^t \to u^*$

For the proof, please refer to [BPCPE10]. Note that we cannot assure the convergence of the primal variables without further assumptions. As for non-strictly convex $f(x)$, though there is unique optimal values f^*, the solution x^* is not necessary unique.

Finally we gives an example on how to apply ADMM. Consider the generalized Lasso with repeated ridge:

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|Fx\|_1. \tag{19.19}$$

F can be arbitrary form, for example, let

$$F = \begin{bmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & \cdots & \\ & & -1 & 1 \end{bmatrix}, \tag{19.20}$$

the regularizer has the form $\sum_i |x_i - x_{i+1}|$, which makes the adjacent pair (x_i, x_{i+1}) be similar. Rewrite (19.19) as a constraint problem

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|z\|_1 \quad \text{s.t.} \quad Fx - z = 0. \tag{19.21}$$

The augmented Lagrangian form is

$$L(x, z, u) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|z\|_1 + \rho u^T(Fx - z) + \frac{\rho}{2}\|Fx - z\|_2^2, \tag{19.22}$$

where we add a scalar ρ before u to make things simple without affecting the results. First we solve $x^{t+1} = \arg\min_x L(x, z^t, u^t)$, it's a quadratic function about x and has the following closed form solution

$$x^{t+1} = \left(A^T A + \rho F^T F\right)^{-1} \left(A^T b + \rho F^T(z^t - u^t)\right). \tag{19.23}$$

Next calculate $z^{t+1} = \arg\min_z L(x^{t+1}, z, u^t)$, it is a soft-shrinkage problem whose solution is also of closed form:

$$z^{t+1} = S_{\lambda/\rho}(Fx^{t+1} + u^t), \tag{19.24}$$

where the soft-shrinkage operator is defined as $S_\kappa(a) = \left(1 - \frac{\kappa}{\|a\|_2}\right)_+ a$ and $(\cdot)_+$ is the positive part. Finally we update u by $u^{t+1} = u^t + Fx^{t+1} - z^{t+1}$.

Solving the original problem (19.19) directly is not easy. By dividing it into two components, we can solve each one with closed form solution. ADMM decomposes complex optimization function elegantly and can be extend to distributed version easily. It converges fast at the early stage, but requires a large number of iterations for high precision solution.

19.2 Mirror Descent

Mirror Descent (MD) is a descent method just like Gradient Descent (GD). MD is better than GD when in high dimension. Say if the dataset is in n-dimensional, MD scales well with n, better than GD.

19.2.1 Bergman Divergence

1. When we say a function g is strongly (or λ-strongly) convex wrt norm $||.||$, we mean:

$$g(y) \geq g(x) + \nabla g_x^T(y - x) + \frac{\lambda}{2}||y - x||^2 \tag{19.25}$$

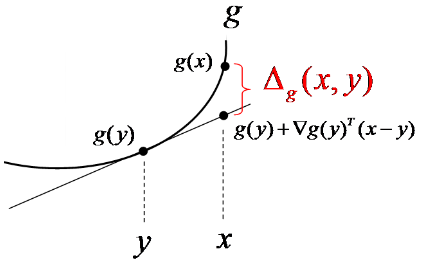
   For example, if function g is strongly convex wrt norm $||.||_2$,then that means:

$$g(y) \geq g(x) + \nabla g_x^T(y - x) + \frac{\lambda}{2}||y - x||_2^2 \tag{19.26}$$

2. If a function g is strongly convex wrt norm $||.||$, we define Bregman Divergence Δ_g to be:

$$\Delta_g(x, y) = g(x) - [g(y) + \nabla g(y)^T(x - y)] \tag{19.27}$$

3. Δ_g is the distance between x and y as meansured by function g



4. Eg 1: $g(x) = |||x||_2^2$ was strongly convex wrt $||.||_2$,

$$\Delta_g(x, y) = ||x - y||_x^2 \tag{19.28}$$

5. Eg 2: $g(x) = \sum_i(x_i log x_i - x_i)$ strong ly convex wrt $||.||_1$,

$$\Delta_g(x, y) = \sum_i(x_i log(\frac{x_i}{y_i}) + y_i - x_i) \tag{19.29}$$

   - $g(x)$ is unromalized entropy.

- This result form should remind you of KL divergence.

6. Some properites of Bregman Divergence.

   (a) By definition, $\Delta_g(x,x) = 0$

   (b) By definition (19.27), and by (19.25), $\Delta_g(x,y)\ ge\frac{\lambda}{2}^2 \geq 0$

   (c) When we take the derivatives of (19.25), we have:

$$\nabla_x \Delta_g(x,y) = \nabla g(x) - \nabla g(y) \tag{19.30}$$

$$\nabla_x^2 \Delta_g(x,y) = \nabla^2 g(x) \succeq \lambda I \tag{19.31}$$

   (d) Triangle Inequality (kinda):

$$\Delta_g(x,y) + \Delta_g(y,z) = \Delta_g(x,z) + (\nabla g(z) - \nabla g(y))^T(x-y) \tag{19.32}$$

19.2.2 Mirror Descent - Updates

Remarks: MD is same to GD, just use a different distance meansure (Bregman).

1. In Gradient Descent, we minimize quadratic approx. of f at $x^t (H^t = I)$

$$x^{t+1} = argmin_x f(x^t) + \partial f(x^t)^T(x - x^t) + \frac{1}{2}||x - x^t||_2^2 \tag{19.33}$$

2. In Mirror Descent, instead of using $\frac{1}{2}||x - x^t||_2^2$, we use Bergman Divergence. Gvien a norm $||.||$ over the domain S,

$$x^{t+1} = argmin_x f(x^t) + \partial f(x^t)^T(x - x^t) + \nabla_g(x, x^t) \tag{19.34}$$

   where g is strongly convex wrt $||x||$.

3. Alternatively,

$$\begin{aligned}
x^{t+1} &= \arg\min_x \quad f(x^t) + \partial f(x^t)^T(x - x^t) + \nabla_g(x, x^t) && (19.35)\\
&= \arg\min_x \quad f(x^t) + \partial f(x^t)^T(x - x^t) + g(x) - g(x^t) - \nabla g(x^t)^T(x - x^t) && (19.36)\\
&= \arg\min_x \quad \cancel{f(x^t)} + \partial f(x^t)^T(x \cancel{- x^t}) + g(x) - \cancel{g(x^t)} - \nabla g(x^t)^T(x \cancel{- x^t}) && (19.37)\\
&= \arg\min_x \quad \partial f(x^t)^T x + g(x) - \nabla g(x^t)^T x && (19.38)\\
&= \arg\min_x \quad x^T(\partial f(x^t) - \nabla g(x^t)) + g(x) && (19.39)
\end{aligned}$$

   In some literature, you will actually see the updating in this form.

4. If we take the derivative of (19.39) wrt x, we have,

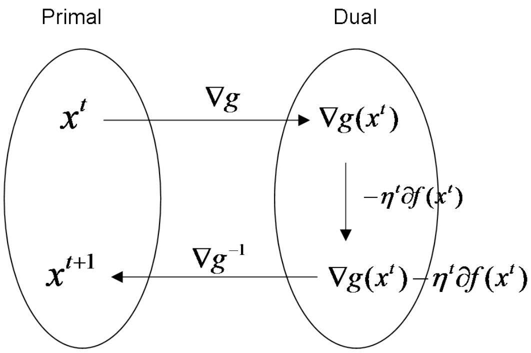
$$\partial f(x^t) - \nabla g(x^t) + \nabla g(x^t) = 0 \tag{19.40}$$

5. Sometimes, we see people write $x^{t+1} = \nabla g^{-1}(..)$, that is because we can rearrange (19.40),

$$\nabla g(x^t) = \nabla g(x^t) - \eta^t \partial f(x^t) \tag{19.41}$$

   Hence,

$$x^{t+1} = \nabla g^{-1}(\nabla g(x^t) - \eta^t \partial f(x^t)) \tag{19.42}$$


Primal
Dual
x^t
∇g
$\nabla g(x^t)$
$-\eta^t \partial f(x^t)$
x^{t+1}
∇g^{-1}
$\nabla g(x^t) - \eta^t \partial f(x^t)$

6. Another way people describe Mirror Descent is draw some diagram like this:

   - Start from x^t in the Primal space, you take gradient of g, and got $\nabla g(x^t)$ in Dual space.
   - Then in the Dual spece, you take a step, in the direction of $-\eta^t \partial f(x^t)$.
   - Then you take a inverse step, ∇g^-1, and got your $x^(t + 1)$ in the Primal space.
   - Somehow I see a mirror here between Primal and Dual spaces ...

19.2.3 Convergence Guarantees

1. Let $||\partial f(x)||_* \leq L_{||x||}$. We say that function f is in Lipshits L, and here L is the Lipshits constant. For different norms $||.||$, we should have different L ($L_{||.||}$). Equivalently,

$$f(x) - f(y) \leq L_{||.||} ||x - y|| \tag{19.43}$$

2. Let $x^g = \arg\min_{x \in S} g(x)$. You can imagine x^g being some kind of center of set S.

   Let diameter $D_{g,||.||} = \sqrt{2 \max_y \Delta_g(x^g, y)/\kappa}$, then

$$||x - x^g|| \leq D_{g,||.||} \tag{19.44}$$

3. Chooing $\eta^t = \frac{\lambda D_{g,||.||}}{||\partial f(x^t)||_* \sqrt{T}}$,

$$f(x^T) - f(x^*) \leq \frac{L_{||.||} D_{g,||.||}}{\sqrt{T}} \tag{19.45}$$

4. The above (19.45) is a very generalized form. For the case of $||.||_2$ norm, you will get what is similar to the form from HW2 (via regret) for projected subgradient descent.

$$f(x^T) - f(x^*) \leq \frac{L_2 D_2}{\sqrt{T}} \tag{19.46}$$

   where both $L_2 f, S$ and $D_2 S$ depends on set S

   - $L_2 = \sqrt{\max_x ||\partial f(x)||_2^2}$
   - $D_2 = \sqrt{\max_{x,y} ||x - y||_2^2}$

19.2.4 Convergence Example: Probability Simplex and $||.||_1$

1. for n-dimensional simplex: $x \geq 0$, $1^T x = 1$
2. Functions are Lipschitz wrt $||.||_1$: $max_x ||\partial f(x)||_{\text{inf}} \leq L_1$
   This is actually a very week condition (subgradient bounded in infinity norm by L1), allowing very large subgradient.
3. We choose g to be this (kind of) unnormalized entropy (sum of x_i is 1):

   $$g(x) = \sum_i x_i log x_i - x_i \tag{19.47}$$

   The updating is like this, what we call exponentiated gradient:

   $$x^{t+1} = x^t \circ \exp(-\eta^t \nabla f(x^t)) \tag{19.48}$$

   This is because:

   $$g(x) = \sum_i x_i \log x_i - x_i \tag{19.49}$$
   $$\nabla g(x) = \log x \;\; \text{(log applied element-wise)} \tag{19.50}$$
   $$\nabla g^{-1}(x) = \exp(x) \;\; \text{(element-wise)} \tag{19.51}$$
   $$\nabla g(x^{t+1}) = \nabla g(x^t) - \eta^t \nabla f(x^t) \;\; \text{(updating in Dual space)} \tag{19.52}$$
   $$\log(x^{t+1}) = log(x^t) - \eta^t \nabla f(x^t) \tag{19.53}$$
   $$x^{t+1} = \exp(\log(x^t) - \eta^t \nabla f(x^t)) \tag{19.54}$$
   $$x^{t+1} = x^t \circ \exp(-\eta^t \nabla f(x^t)) \tag{19.55}$$

4. Diameter $D_{g,||.||} \leq \sqrt{2 \log n}$, yielding a rate $\sqrt{\log n / T}$
5. Suppose you use gradient descent, with $g(x) = ||x||_2^2$. Then
   - Then the diameter bound in L_2 norm is $D_2 = 1$;
   - the Lipschitz bound in L_2 norm is $L_2 \leq \sqrt{n} L_1$.
   - Put them together you get a rate of $\sqrt{n/T}$.

References

[BPCPE10] Boyd, Parikh, Chu, Peleato, and Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Methodof Multipliers."

[BN12] Ben-Tal and Nemirovski, "Lecture Notes on Modern Convex Optimization."