10-725/36-725: Convex Optimization **Fall 2016**

# Lecture 20: November 9

*Lecturer: Ryan Tibshirani* *Scribes: Yilin Yang, Rui Zhu, Jilong Yu*

**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.

## 20.1 Dual (sub)gradient methods

When we cannot find a closed-form solution of the original problem, we can still apply subgradient or gradient methods to its dual problem.

Considering the primal problem as

$$\min_x f(x) \text{ subject to } \; Ax = b$$

We write it dual problem with the conjugate of $f$ as

$$\max_u -f^*(-A^T u) - b^T u$$

We define $g(u) = -f^*(-A^T u) - b^T u$, given the definition of subgradients

$$\partial g(u) = A \partial f^*(-A^T u) - b$$

Given $x \in \partial f^*(-A^T u) \Leftrightarrow x \in \arg\min_z f(z) - (-A^T u)^T z \Leftrightarrow x \in \arg\min_z f(z) + u^T Az$, we write the formal equation as

$$\partial g(u) = Ax - b$$

where $x \in \arg\min_z f(z) + u^T Az$

We solve this with dual subgradient method by repeating two steps for $k = 1, 2, 3, ...$(starting with an initial $u^{(0)}$):

- $x^{(k)} \in \arg\min_x f(x) + (u^{(k-1)})^T Ax$
- $u^{(k)} = u^{(k-1)} + t_k(Ax^{(k)} - b)$

where step sizes $t_k$ are randomly chosen in standard ways. When $f$ is strictly convex, the first item becomes an equation $x^{(k)} = \arg\min_x f(x) + (u^{(k-1)})^T Ax$. Proximal gradients and acceleration are also applicable when necessary.

### 20.1.1 Convergence Analysis

**Lipschitz gradients and strong convexity**

**Conclusion**: Assume that $f$ is a closed and convex function. Then $f$ is strongly convex with parameter $d \Leftrightarrow \nabla f^*$ Lipschitz with parameter $1/d$.

**Proof**: We give proof for "$\Longrightarrow$": Given the definition of strong convexity, if $g$ strongly convex with minimizer $x$, we have

$$g(y) \geq g(x) + \frac{d}{2}||y-x||_2$$

for all $y$. Also define two minimizers $x_u = \nabla f^*(u)$ minimizing $g(x) = f(x) - u^T x$, and $x_v = \nabla f^*(v)$ minimizing $g(x) = f(x) - v^T x$, we have

$$f(x_v) - u^T x_v \geq f(x_u) - u^T x_u + \frac{d}{2}||x_u - x_v||_2^2$$

$$f(x_u) - v^T x_u \geq f(x_v) - v^T x_u v + \frac{d}{2}||x_u - x_v||_2^2$$

Adding these two together, we have

$$f(x_v) - u^T x_v + f(x_u) - v^T x_u \geq f(x_u) - u^T x_u + f(x_v) - v^T x_v + d||x_u - x_v||_2^2$$

Then $d||x_u - x_v||_2^2 \leq u^T x_v + u^T x_u + v^T x_v + v^T x_u = (u-v)^T(x_u - x_v) \leq ||u-v||_2||x_u - x_v||_2$, so we end up with

$$||x_u - x_v||_2 \leq \frac{1}{d}||u-v||_2$$

**Convergence Guarantees** Givne the last facts we have, we have the convergence rate for Dual (sub)gradient methods (dual objective):

- If $f$ is strongly convex with parameter $d$, then dual gradient ascent with fixed step sizes $t_k = d$, $k = 1, 2, 3, ...$, converges at the rate $O(1/\epsilon)$
- If $f$ is strongly convex with parameter $d$, and $\nabla f$ is Lipschitz with parameter $L$, then dual gradient ascent with fixed step sizes $t_k = 2/(1/d + 1/L)$, $k = 1, 2, 3, ...$, converges at the rate $O(\log(1/\epsilon))$

## 20.2 Dual Decomposition

Consider

$$\min_x \sum_{i=1}^{B} f_i(x_i)$$
$$\text{subject to} \quad Ax = b$$

Where $x = (x_1, x_2, ..., x_B) \in \mathbb{R}^n$ with each $x_i \in \mathbb{R}^{n_i}$

It is easy to observe that while calculating the (sub)gradient, the minimization problem can be transformed to B separate problems.

$$x^+ \in \arg\min_x \sum_{i=1}^{B} f_i(x_i) + u^T A x$$

$$\iff x_i^+ \in \arg\min_{x_i} f_i(x_i) + u^T A_i x_i, \quad i = 1, 2, ..., B$$

### 20.2.1 Dual Decomposition with Equality Constraints

Dual Decompostion Algorithm:

$$\text{Broadcast: } x_i^{(k)} \in \arg\min_{x_i} f_i(x_i) + (u^{(k-1)})^T A_i x_i, \quad i = 1, 2, ...B$$

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

### 20.2.2 Dual Decomposition with Inequality Constraints

$$\min_x \sum_{i=1}^{B} f_i(x_i)$$

$$\text{subject to } \sum_{i=1}^{B} A_i x_i \leq b$$

Dual decomposition (projected subgradient method): repeat for k = 1,2,3,...

$$\text{Broadcast: } x_i^{(k)} \in \arg\min_{x_i} f_i(x_i) + (u^{(k-1)})^T A_i x_i, \quad i = 1, 2, ...B$$

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

where u+ denotes the positive part of u, i.e., $(u_+)_i = \max\{0, u_i\}, i = 1, ..., m$

Price coordination interpretation (Vandenberghe):

- Have B units in a system, each unit chooses its own decision variable $x_i$ (how to allocate its goods)
- Constraints are limits on shared resources (rows of A), each component of dual variable $u_j$ is price of resource $j$
- Dual Update:

  $$u_j^+ = (u_j - t s_j)_+, \quad j = 1, 2, ...m$$

  where $s = b - \sum_{i=1}^{B} A_i x_i$ are slacks

  - Increase price $u_j$ if resource $j$ is over-utilized, $s_j < 0$
  - Decrease price $u_j$ if resource $j$ is under-utilized, $s_j > 0$
  - Never let prices get negative.

## 20.3 Augmented Lagrangian Method (Method of Multipliers)

Transfer the primal problem to an equivalent problem, where $\rho > 0$ is a parameter:

$$\begin{aligned} \min_x \quad & f(x) + \frac{\rho}{2}||Ax - b||_2^2 \\ \text{subject to} \quad & Ax = b \end{aligned}$$

Dual gradient ascent:

$$\begin{aligned} x^{(k)} &= \arg\min_x \; f(x) + (u^{(k-1)})^T Ax + \frac{\rho}{2}||Ax - b||_2^2 \\ u^{(k)} &= u^{(k-1)} + \rho(Ax^{(k)} - b) \end{aligned}$$

Choose step size $t_k = \rho$:

$$\begin{aligned} 0 &\in \partial f(x^{(k)} + A^T(u^{(k-1)} + \rho(Ax^{(k)} - b)) \\ &= \partial f(x^{(k)} + A^T u^{(k)} \end{aligned}$$

This is the stationarity condition for the original primal problem. $Ax^{(k)} - b$ approaches zero under mild conditions, hence $x^{(k)}, u^{(k)}$ approach optimality.

Advantage: Much better convergence properties. Objective is strongly convex when $A$ has full column rank. Disadvantage: Lose decomposability.

## 20.4 Alternating Direction Method of Multipliers (ADMM)

### 20.4.1 ADMM

Consider

$$\begin{aligned} \min_{x,z} \quad & f(x) + g(z) \\ \text{subject to} \quad & Ax + Bz = c \end{aligned}$$

Augment the objective with $\rho > 0$:

$$\begin{aligned} \min_{x,z} \quad & f(x) + g(z) + \frac{\rho}{2}||Ax + Bz - c||_2^2 \\ \text{subject to} \quad & Ax + Bz = c \end{aligned}$$

Define the augmented Lagrangian as:

$$L_\rho(x, z, u) = f(x) + g(z) + u^T(Ax + Bz - c) + \frac{\rho}{2}||Ax + Bz - c||_2^2$$

Steps for ADMM are:

$$\begin{aligned} x^{(k)} &= \arg\min_x \; L_\rho(x, z^{(k-1)}, u^{(k-1)}) \\ z^{(k)} &= \arg\min_z \; L_\rho(x^{(k)}, z, u^{(k-1)}) \\ u^{(k)} &= u^{(k-1)} + \rho(Ax^{(k)} + Bz^{(k)} - c) \end{aligned}$$

### 20.4.2 Convergence Guarantees

Under modest assumptions on $f, g$, for any $A, B, \rho > 0$:
**Residual convergence:** $r^{(k)} = Ax^{(k)} + Bz^{(k)} - c \to 0$
**Objective convergence:** $f(x^{(k)}) + g(z^{(k)}) \to f^*$
**Dual convergence:** $u^{(k)} \to u^*$

Convergence rate is not known in general. Roughly, it behaves like a first-order method or a bit faster.

### 20.4.3 ADMM in Scaled Form

Replace the dual variable $u$ by a scaled variable $w = u/\rho$:

$$x^{(k)} = \arg\min_x \; f(x) + \frac{\rho}{2}||Ax + Bz^{(k-1)} - c + w^{(k-1)}||_2^2$$

$$z^{(k)} = \arg\min_z \; g(x) + \frac{\rho}{2}||Ax^{(k)} + Bz - c + w^{(k-1)}||_2^2$$

$$w^{(k)} = w^{(k-1)} + Ax^{(k)} + Bz^{(k)} - c$$

$w^{(k)}$ is given by a sum of residuals:

$$w^{(k)} = w^{(0)} + \sum_{i=1}^{k}(Ax^{(i)} + Bz^{(i)} - c)$$

### 20.4.4 Example: Alternating Projections

Find a point in intersection of convex sets $C, D \subseteq \mathbb{R}^n$:

$$\min_x \; I_C(x) + I_D(x)$$

Express it as:

$$\min_{x,z} \; I_C(x) + I_D(z) \quad \text{subject to} \quad x - z = 0$$

ADMM cycle involves two projections:

$$x^{(k)} = \arg\min_x \; P_C(z^{(k-1)} - w^{(k-1)})$$

$$z^{(k)} = \arg\min_z \; P_D(x^{(k)} + w^{(k-1)})$$

$$w^{(k)} = w^{(k-1)} + x^{(k)} - z^{(k)}$$