ADMM and Mirror Descent

Geoff Gordon & Ryan Tibshirani
(I am Aaditya Ramdas and I approve this lecture)
Optimization 10-725 / 36-725

Oct 30, 2012

Recap of Dual Ascent

For problems like

$$\min_x f(x)$$
$$s.t. \quad Ax = b$$

We defined Lagrangian

$$L(x, u) = f(x) + u^\top (Ax - b)$$

We defined the lagrange dual function

$$g(u) = \inf_x L(x, u)$$

If x^+ minimizes $L(x, u)$ then $\partial g(u) = Ax^+ - b$

Recap of Dual Ascent

Dual problem : maximize $g(u)$ - use subgradient ascent!

This gives us the algorithm

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

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

If strong duality, $x^* = \arg\min_x L(x, u^*)$, provided it is unique.

For appropriate η^t (and some conditions), x^t, u^t converge to an optimal primal and dual point.

If g not differentiable, convergence not monotone, i.e. sometimes $g(u^{t+1}) \not\geq g(u^t)$.

Recap of Dual Decomposition Ascent

Suppose $f(x) = \sum_i f_i(x_i)$ where $x_i \in R^{n_i}$ are disjoint

Write $Ax = \sum_i A_i x_i$, and so

$$L(x,u) = \sum_i L_i(x_i,u) = \sum_i \left(f_i(x_i) + u^\top A_i x_i - (1/N) u^\top b \right)$$

x-minimization step in dual ascent decomposes

$$x_i^{t+1} = \arg\min_{x_i} L_i(x_i, u^t)$$

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

Recap of Augmented Lagrangian, Method of Multipliers

$$L_\rho(x,u) = f(x) + u^\top(Ax-b) + (\rho/2)\|Ax-b\|_2^2$$

Lagrangian of $\min_x f(x) + (\rho/2)\|Ax-b\|_2^2 \;\; s.t. \;\; Ax = b$

Associated dual function

$$g_\rho(u) = \min_x L_\rho(x,u)$$

Applying dual ascent :

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

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

More robust than dual ascent (converges if f is not strictly convex or when f can be infinite). However, lost decomposability.

Alternating Direction Method of Multipliers

Augmented Lagrangian for $f(x) = f_1(x_1) + f_2(x_2)$ is $L_\rho(x_1, x_2, u)$

$$= f_1(x_1) + f_2(x_2) + u^\top (A_1 x_1 + A_2 x_2 - b) + (\rho/2)\|A_1 x_1 + A_2 x_2 - b\|_2^2$$

"Alternating direction" minimization

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

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

$$u^{t+1} = u^t + \rho(A_1 x_1^{t+1} + A_2 x_2^{t+1} - b)$$

Normal method of multipliers would've done

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

$$u^{t+1} = u^t + \rho(A_1 x_1^{t+1} + A_2 x_2^{t+1} - b)$$

Convergence Guarantees of ADMM

Assumption 1: f_1, f_2 are closed, proper, convex (epigraphs are closed, nonempty, convex)

Assumption 2: Unaugmented Lagrangian $L_0(x_1, x_2, u)$ has saddle

$$L_0(x_1^S, x_2^S, u) \leq L_0(x_1^S, x_2^S, u^S) \leq L_0(x_1, x_2, u^S)$$

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 : $y^t \to y^*$

Primal variables needn't converge (more assumptions needed)

Example: Generalized Lasso with Repeated Ridge

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

In ADMM form,

$$\min_{x,z} \ \|Ax-b\|_2^2 + \lambda\|z\|_1$$

$$s.t. \ \ Fx - z = 0$$

ADMM updates :

$$\begin{aligned} x^{t+1} &= (A^\top A + \rho F^\top F)^{-1}(A^\top b + \rho F^\top (z^t - u^t)) \\ z^{t+1} &= S_{\lambda/\rho}(Fx^{t+1} + u^t) \\ u^{t+1} &= u^t + Fx^{t+1} - z^{t+1} \end{aligned}$$

For group lasso ($\lambda \sum_i \|x_i\|_2$ for disjoint $x_i \in \mathbb{R}^{n_i}$), ADMM uses vector soft thresholding operator

$$S_\kappa(a) = \left(1 - \frac{\kappa}{\|a\|_2}\right)_+ a$$

Break!

Bregman Divergence Δ_g

If g is strongly convex wrt norm $\|.\|$, define

$$\Delta_g(x, y) = g(x) - [g(y) + \nabla g(y)^\top (x - y)]$$

Read "**distance between x and y as measured by function g**".

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

$$\Delta_g(x, y) = \|x - y\|_2^2$$

Eg: $g(x) = \sum_i (x_i \log x_i - x_i)$, strongly convex wrt $\|.\|_1$

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

Properties of Bregman Divergence

For a λ-strongly convex function g, we defined

$$\Delta_g(x, y) = g(x) - [g(y) + \nabla g(y)^\top (x - y)]$$

So $\Delta_g(x, x) = 0$ and by strong convexity,

$$\Delta_g(x, y) \geq \frac{\lambda}{2}\|x - y\|^2 \geq 0$$

Derivatives:

$$\nabla_x \Delta_g(x, y) = \nabla g(x) - \nabla g(y)$$

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

Triangle Inequality (kinda):

$$\Delta_g(x, y) + \Delta_g(y, z) = \Delta_g(x, z) + (\nabla g(z) - \nabla g(y))^\top (x - y)$$

Recap of Gradient Descent

Consider ($S \subseteq \mathbb{R}^n$) the problem $\min_{x \in S} f(x)$

Gradient descent : minimize quadratic approx. of f at x^t ($H^t = I$)

$$x^{t+1} = \arg\min_x \quad f(x^t) + \partial f(x^t)^\top (x - x^t) + \frac{1}{2}\|x - x^t\|_2^2$$

From HW2 (via regret) : for projected subgradient descent,

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

where $\max_{x \in S} \|\partial f(x)\|_2 \leq L_2$, $\max_{x,y \in S} \|x - y\|_2 \leq D_2$

How does this scale with n? Depends on $L_2(f, S)$ and $D_2(S)$

Mirror Descent

Given a norm $\|.\|$ over the domain S,

$$x^{t+1} = \arg\min_x \quad f(x^t) + \partial f(x^t)^\top (x - x^t) + \Delta_g(x, x^t)$$

where g is strongly convex wrt $\|.\|$. Alternatively,

$$x^{t+1} = \arg\min_x \quad x^\top (\partial f(x^t) - \nabla g(x^t)) + g(x)$$

Hence,

$$\partial f(x^t) + \nabla g(x^{t+1}) - \nabla g(x^t) = 0$$

So, we sometimes see

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

Convergence Guarantees

Let $\|\partial f(x)\|_* \leq L_{\|.\|}$ or equivalently

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

If $x^g = \arg\min_{x \in S} g(x)$, let $D_{g,\|.\|} = \sqrt{2 \max_y \Delta_g(x^g, y)/\lambda}$, then

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

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

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

Remember (HW2): $\eta^t = \frac{D_2}{L_2 \sqrt{T}}$ and $D_2 = \sqrt{\max_{x,y} \|x - y\|_2^2}$

Example : Probability Simplex and $\|.\|_1$

n-dimensional simplex : $x \geq 0, 1^\top x = 1$

Functions are Lipschitz wrt $\|.\|_1$: $\max_x \|\partial f(x)\|_\infty \leq L_1$

If $g(x) = \sum_i x_i \log x_i - x_i$, we get exponentiated gradient

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

$D_{g,\|.\|_1} \leq \sqrt{2 \log n}$, yielding a rate $\sqrt{\log n / T}$.

$g(x) = \|x\|_2^2$ (grad. descent) gives $\sqrt{n/T}$ ($D_2 = 1$, $L_2 \leq \sqrt{n} L_1$)

References

Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers - Boyd, Parikh, Chu, Peleato and Eckstein, 2010

Lecture Notes on Modern Convex Optimization - Ben-Tal and Nemirovski, 2012