Dual Proximal Gradient Method

http://bicmr.pku.edu.cn/~wenzw/opt-2018-fall.html

Acknowledgement: this slides is based on Prof. Lieven Vandenberghes lecture notes

Outline

Dual methods

subgradient method : slow, step size selection difficult

gradient method : requires differentiable dual cost function

- often dual cost is not differentiable, or has nontrivial domain
- dual can be smoothed by adding small strongly convex term to primal

augmented Lagrangian method

- equivalent to gradient ascent on a smoothed dual problem
- however smoothing destroys separable structrue

proximal gradient method(this lecture): dual cost split in two terms

- one term is differentiable with Lipschitz continuous gradient
- other term has an inexpensive prox-operator

Composite structure in the dual

$$\min\ f(x) + g(Ax) \qquad\qquad \max\ -f^*(-A^T z) - g^*(z)$$

dual has the right structure for the proximal gradient method if

- prox-operator of g (or g^*) is cheap (closed form or simple algorithm)
- f is strongly convex ($f(x) - (\mu/2)x^T x$ is convex)

  implies $f^*(-A^T z)$ has Lipschitz continuous gradient ($L = ||A||_2^2/\mu$):

$$||A\nabla f^*(-A^T u) - A\nabla f^*(-A^T v)||_2 \leq \frac{||A||_2^2}{\mu}||u - v||_2$$

  because ∇f^* is Lipschitz continuous with constant $1/\mu$

Dual proximal gradient update

$$z^+ = \mathrm{prox}_{tg^*}(z + tA\nabla f^*(-A^T z))$$

equivalent expression in terms of f:

$$z^+ = \mathrm{prox}_{tg^*}(z + tA\hat{x}) \qquad \text{where } \hat{x} = \operatorname*{argmin}_x (f(x) + z^T Ax)$$

- if f is separable, calculation of $\hat{x}$ decomposes into independent problems
- step size t constant or from backtracking line search
- can use accelerated proximal gradient methods

Alternating minimization interpretation

- Moreau decomposition:

$$
\begin{aligned}
x &= \mathrm{prox}_h(x) + \mathrm{prox}_{h^*}(x) \\
x &= \mathrm{prox}_{th}(x) + t\mathrm{prox}_{t^{-1}h^*}(x/t) \\
x &= t\mathrm{prox}_{t^{-1}h}(x/t) + \mathrm{prox}_{th^*}(x)
\end{aligned}
$$

- Let $\hat{y} = \mathrm{prox}_{t^{-1}g}(z/t + A\hat{x})$:

$$
\begin{aligned}
& z^+ = \mathrm{prox}_{tg^*}(z + tA\hat{x}) \\
\Longleftrightarrow \quad & z + tA\hat{x} = z^+ + t\mathrm{prox}_{t^{-1}g}(z/t + A\hat{x}) \\
\Longleftrightarrow \quad & z^+ = z + t(A\hat{x} - \hat{y}),
\end{aligned}
$$

- The computation of $\hat{y}$ is equivalent to

$$
\begin{aligned}
& \min_y \quad g(y) + \frac{t}{2}\|y - (z/t + A\hat{x})\|_2^2 \\
\Longleftrightarrow \quad & \min_y \quad g(y) + \langle z, A\hat{x} - y\rangle + \frac{t}{2}\|A\hat{x} - y\|_2^2
\end{aligned}
$$

Alternating minimization interpretation

Moreau decomposition gives alternate expression for z-update

$$z^+ = z + t(A\hat{x} - \hat{y})$$

where

$$\begin{aligned}
\hat{x} &= \underset{x}{\operatorname{argmin}}(f(x) + z^T Ax) \\
\hat{y} &= \operatorname{prox}_{t^{-1}g}(z/t + A\hat{x}) \\
&= \underset{y}{\operatorname{argmin}}(g(y) + z^T(A\hat{x} - y) + \frac{t}{2}||A\hat{x} - y||_2^2)
\end{aligned}$$

in each iteration, an alternating minimization of:

- Lagrangian $f(x) + g(y) + z^T(Ax - y)$ over x
- augmented Lagrangian $f(x) + g(y) + z^T(Ax - y) + \frac{t}{2}||Ax - y||_2^2$ over y

Alternating minimization method

Consider the equivalent problem:

$$\min_{x,y} f(x) + g(y), \qquad \text{s.t.} \qquad Ax = y$$

Define the Lagrange function:

$$L(x, y, z) = f(x) + g(y) + \langle z, Ax - y \rangle$$

Define the augmented Lagrangian function:

$$L_t(x, y, z) = L(x, y, z) + \frac{t}{2}\|Ax - y\|_2^2.$$

The equivalent alternating minimization scheme is

$$\begin{aligned} x^{k+1} &= \arg\min_x L(x, y^k, z^k) \\ y^{k+1} &= \arg\min_y L_t(x^{k+1}, y, z^k) \\ z^{k+1} &= z^k + t(Ax^{k+1} - y^{k+1}) \end{aligned}$$

Outline

Regularized norm approximation

$$\min\ f(x) + \|Ax - b\| \qquad \text{(with } f \text{ strongly convex)}$$

a special case of Page 4 with $g(y) = ||y - b||$

$$g^*(x) = \begin{cases} b^T z & ||z||_* \leq 1 \\ +\infty & \text{otherwise} \end{cases} \qquad \text{prox}_{tg^*}(z) = P_C(z - tb)$$

C is unit norm ball for dual norm $|| \cdot ||_*$.

dual gradient projection update

$$\hat{x} = \underset{x}{\text{argmin}}(f(x) + z^T Ax)$$

$$z^+ = \text{prox}_{tg^*}(z + tA\hat{x}) = P_C(z + t(A\hat{x} - b))$$

Regularized norm approximation

Consider an equivalent problem

$$\min_{x,y} f(x) + \|y\|, \quad \text{s.t. } Ax - b = y$$

The alternating minimization scheme is

$$\begin{aligned} x^+ &= \operatorname*{argmin}_x \quad f(x) + \|y\| + \langle z, Ax - b - y \rangle \\ y^+ &= \operatorname*{argmin}_y \quad f(x^+) + \|y\| + \left\langle z, Ax^+ - b - y \right\rangle + \frac{t}{2}\|Ax - b - y\|_2^2 \\ z^+ &= z + t(Ax^+ - b - y^+) \end{aligned}$$

Example

$$\min f(x) + \sum_{i=1}^{p} ||B_i x||_2 \qquad \text{(with } f \text{ strongly convex)}$$

a special case of Page 4 with $g(y_1, \ldots, y_p) = \sum_{i=1}^{p} ||y_i||_2$ and

$$A = \begin{bmatrix} B_1^T & B_2^T & \cdots & B_p^T \end{bmatrix}^T$$

dual gradient projection update

$$\hat{x} = \underset{x}{\operatorname{argmin}} (f(x) + (\sum_{i=1}^{p} B_i^T z_i)^T x)$$

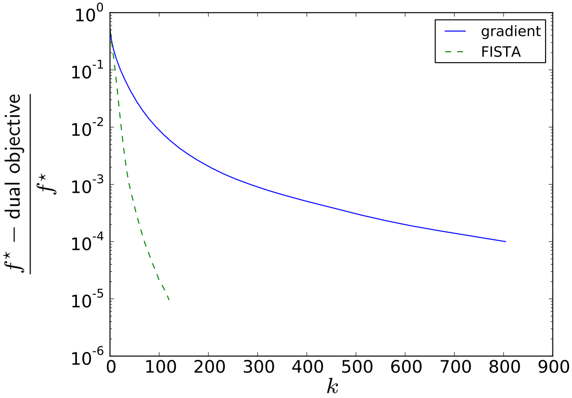
$$z_i^+ = \operatorname{prox}_{tg^*}(z_i + tA\hat{x}) = P_{C_i}(z_i + tB_i\hat{x}), \quad i = 1, \ldots, p$$

C_i is unit Euclidean norm ball in $\mathbb{R}^{m_i}$, if $B_i \in \mathbb{R}^{m_i \times n}$

numerical example

$$f(x) = \frac{1}{2}||Cx - d||_2^2$$

with random generated $C \in \mathbb{R}^{2000\times 1000}$, $B_i \in \mathbb{R}^{10\times 1000}$, $p = 500$



Minimization over intersection of convex sets

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in C_1 \cap \cdots \cap C_m \end{array}$$

- f strongly convex; $e.g.$, $f(x) = ||x - a||_2^2$ for projecting a on intersection
- sets C_i are closed, convex, and easy to project onto
- this is a special case of Page 4 with g a sum of indicators

$$g(y_1, \ldots, y_m) = I_{C_1}(y_1) + \cdots + I_{C_m}(y_m), \qquad A = \begin{bmatrix} I & \cdots & I \end{bmatrix}^T$$

dual proximal gradient update

$$\hat{x} = \underset{x}{\operatorname{argmin}}(f(x) + (z_1 + \cdots + z_m)^T x)$$

$$z_i^+ = z_i + t\hat{x} - tP_{C_i}(z_i/t + \hat{x}), \quad i = 1, \ldots, m$$

Decomposition of separable problems

$$\min \quad \sum_{j=1}^{n} f_j(x_j) + \sum_{i=1}^{m} g_i(A_{i1}x_1 + \cdots + A_{in}x_n)$$

each f_i is strongly convex; g_i has inexpensive prox-operator

dual proximal gradient update

$$\hat{x}_j = \underset{x_j}{\operatorname{argmin}}(f_j(x_j) + \sum_{i=1}^{m} z_i^T A_{ij} x_j), \quad j = 1, \ldots, n$$

$$z_i^+ = \operatorname{prox}_{tg_i^*}(z_i + t \sum_{j=1}^{n} A_{ij}\hat{x}_j), \quad i = 1, \ldots, m$$

Outline

1. proximal gradient method applied to the dual

2. Examples

3. alternating minimization method

Primal problem with separable structure

composite problem with separable f

$$\min \quad f_1(x_1) + f_2(x_2) + g(A_1x_1 + A_2x_2)$$

we assume f_1 strongly convex, but not necessarily f_2

dual problem

$$\max \quad -f_1^*(-A_1^Tz) - f_2^*(-A_2^Tz) - g^*(z)$$

- first term is differentiable with Lipschitz continuous gradient
- prox-operator $h(z) = f_2^*(-A_2^Tz) + g^*(z)$ was discussed

Dual proximal gradient method

$$z^+ = \text{prox}_{th}(z + tA_1 \nabla f_1^*(-A_1^T z))$$

- equivalent form using f_1:

$$z^+ = \text{prox}_{th}(z + tA_1 \hat{x}_1) \quad \text{where } \hat{x}_1 = \underset{x_1}{\text{argmin}}(f_1(x_1) + z^T A_1 x_1)$$

- prox-operator of $h(z) = f_2^*(-A_2^T z) + g^*(z)$ is given by

$$\text{prox}_{th}(w) = w + t(A_2 \hat{x}_2 - \hat{y})$$

where $\hat{x}_2, \hat{y}$ minimize an augmented Lagrangian

$$(\hat{x}_2, \hat{y}) = \underset{x_2, y}{\text{argmin}}(f_2(x_2) + g(y) + \frac{t}{2}||A_2 x_2 - y + w/t||_2^2)$$

Proof: $\text{prox}_{th}(w) = w + t(A_2\hat{x}_2 - \hat{y})$

- $h(z) = f_2^*(-A_2^T z) - g^*(z)$ and

$$\begin{aligned} h^*(y) &= \sup_z y^T z - f_2^*(-A_2^T z) - g^*(z) \\ &= \sup_{z,w} y^T z - f_2^*(w) + g^*(z), \text{ s.t. } w = -A_2^T z \\ &= \inf_v \sup_{z,w} y^T z - f_2^*(w) - g^*(z) + v^T(w + A_2^T z) \\ &= \inf_v f_2(v) + g(A_2 v + y) \end{aligned}$$

- Moreau decomposition: $w = \text{prox}_{th}(w) + t\text{prox}_{t^{-1}h^*}(w/t)$

$$\begin{aligned} & \min \quad t^{-1}h^*(y) + \frac{1}{2}\|y - w/t\|_2^2 \\ \Longleftrightarrow & \min_{y,v} \quad f_2(v) + g(A_2 v + y) + \frac{t}{2}\|y - w/t\|_2^2 \\ \Longleftrightarrow & \min_{u,v} \quad f_2(v) + g(u) + \frac{t}{2}\|u - A_2 v - w/t\|_2^2 \quad \text{using } u = A_2 v + y \end{aligned}$$

- $\text{prox}_{th}(w) = w - ty = w - t(u - A_2 v)$

Alternating minimization method

starting at some initial z, repeat the following iteration

1. minimize the Lagrangian over x_1:

$$\hat{x_1} = \underset{x_1}{\operatorname{argmin}}(f_1(x_1) + z^T A_1 x_1)$$

2. minimize the augmented Lagrangian over $\hat{x_2}, \hat{y}$:

$$(\hat{x_2}, \hat{y}) = \underset{x_2, y}{\operatorname{argmin}}(f_2(x_2) + g(y) + \frac{t}{2}||A_1\hat{x_1} + A_2 x_2 - y + z/t||_2^2)$$

3. update dual variable:

$$z^+ = z + t(A_1\hat{x_1} + A_2\hat{x_2} - \hat{y})$$

Comparison with augmented Lagrangian method

augmented Lagrangian method

1. compute minimizer $\hat{x}_1, \hat{x}_2, \hat{y}$ of the augmented Lagrangian

$$\min_{x_1, x_2, y} \quad f_1(x_1) + f_2(x_2) + g(y) + \frac{t}{2}||A_1 x_1 + A_2 x_2 - y + z/t||_2^2$$

2. update dual variable:

$$z^+ = z + t(A_1 \hat{x}_1 + A_2 \hat{x}_2 - \hat{y})$$

differences with alternating minimization

- more general: AL method does not require strong convexity of f_1
- quadratic penalty in step 1 destroys separability

References

alternating minimization method

- P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control and Optimization (1991)
- P. Tseng, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Mathematical Programming (1990)