12. Dual proximal gradient method

- proximal gradient method applied to the dual
- examples
- alternating minimization method

Dual methods

Subgradient method: converges slowly, step size selection is difficult

Gradient method: requires differentiable dual cost function

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

Augmented Lagrangian method

- equivalent to gradient ascent on a smoothed dual problem
- quadratic penalty in augmented Lagrangian destroys separable primal structure

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 primal and dual problem

$$\begin{array}{lll} \text{primal:} & \text{minimize} & f(x) + g(Ax) \\ \text{dual:} & \text{maximize} & -g^*(z) - f^*(-A^T z) \end{array}$$

the dual problem has the right structure for the proximal gradient method if

- f is strongly convex: this implies $f^*(-A^T z)$ has Lipschitz continuous gradient

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

  μ is the strong convexity constant of f (see page 7-16)

- prox-operator of g (or g^*) is inexpensive (closed form or simple algorithm)

Dual proximal gradient update

$$\text{minimize} \quad g^*(z) + f^*(-A^T z)$$

- proximal gradient update:

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

- ∇f^* can be computed by minimizing partial Lagrangian (from p. 7-15, 7-16):

$$\begin{aligned} \hat{x} &= \operatorname*{argmin}_x \left(f(x) + z^T A x\right) \\ z^+ &= \mathrm{prox}_{tg^*}(z + tA\hat{x}) \end{aligned}$$

- partial Lagrangian is a separable function of x if f is separable
- step size t is constant ($t \leq \mu/\|A\|_2^2$) or adjusted by backtracking
- faster variant uses accelerated proximal gradient method of lecture 11

Dual proximal gradient update

$$\begin{aligned}
\hat{x} &= \operatorname*{argmin}_x \left(f(x) + z^T A x\right) \\
z^+ &= \mathrm{prox}_{tg^*}(z + tA\hat{x})
\end{aligned}$$

- Moreau decomposition gives alternate expression for z-update:

$$z^+ = z + tA\hat{x} - t\mathrm{prox}_{t^{-1}g}\big(t^{-1}z + A\hat{x}\big)$$

- right-hand side can be written $z + t(A\hat{x} - \hat{y})$ where

$$\begin{aligned}
\hat{y} &= \mathrm{prox}_{t^{-1}g}\big(t^{-1}z + A\hat{x}\big) \\
&= \operatorname*{argmin}_y \left(g(y) + \frac{t}{2}\left\|A\hat{x} - t^{-1}z - y\right\|_2^2\right) \\
&= \operatorname*{argmin}_y \left(g(y) + z^T(A\hat{x} - y) + \frac{t}{2}\|A\hat{x} - y\|_2^2\right)
\end{aligned}$$

Alternating minimization interpretation

$$
\begin{aligned}
\hat{x} &= \operatorname*{argmin}_{x} \left(f(x) + z^T A x\right) \\
\hat{y} &= \operatorname*{argmin}_{y} \left(g(y) - z^T y + \frac{t}{2}\|A\hat{x} - y\|_2^2\right) \\
z^+ &= z + t(A\hat{x} - \hat{y})
\end{aligned}
$$

- first minimize Lagrangian over x, then augmented Lagrangian over y
- compare with augmented Lagrangian method:

$$
(\hat{x}, \hat{y}) = \operatorname*{argmin}_{x,y} \left(f(x) + g(y) + z^T(Ax - y) + \frac{t}{2}\|Ax - y\|_2^2\right)
$$

- requires strongly convex f (in contrast to augmented Lagrangian method)

Outline

- proximal gradient method applied to the dual

- **examples**

- alternating minimization method

Regularized norm approximation

$$\begin{array}{lll} \text{primal:} & \text{minimize} & f(x) + \|Ax - b\| \\ \\ \text{dual:} & \text{maximize} & -b^T z - f^*(-A^T z) \\ & \text{subject to} & \|z\|_* \leq 1 \end{array}$$

(see page 7-20)

- we assume f is strongly convex with constant μ, not necessarily differentiable
- we assume projections on unit $\|\cdot\|_*$-ball are simple
- this is a special case of the problem on page 12-3 with $g(y) = \|y - b\|$:

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

Dual gradient projection

$$\begin{array}{lll} \text{primal:} & \text{minimize} & f(x) + \|Ax - b\| \\[1em] \text{dual:} & \text{maximize} & -b^T z - f^*(-A^T z) \\ & \text{subject to} & \|z\|_* \leq 1 \end{array}$$

- dual gradient projection update:

$$z^+ = P_C\left(z + t(A\nabla f^*(-A^T z) - b)\right)$$

- gradient of f^* can be computed by minimizing the partial Lagrangian:

$$\begin{array}{rcl} \hat{x} & = & \operatorname*{argmin}_x \left(f(x) + z^T A x\right) \\ z^+ & = & P_C(z + t(A\hat{x} - b)) \end{array}$$

Example

$$\begin{array}{lll} \text{primal:} & \text{minimize} & f(x) + \sum_{i=1}^{p} \|B_i x\|_2 \\ \text{dual:} & \text{maximize} & -f^*(-B_1^T z_1 - \cdots - B_p^T z_p) \\ & \text{subject to} & \|z_i\|_2 \leq 1, \quad i = 1, \ldots, p \end{array}$$

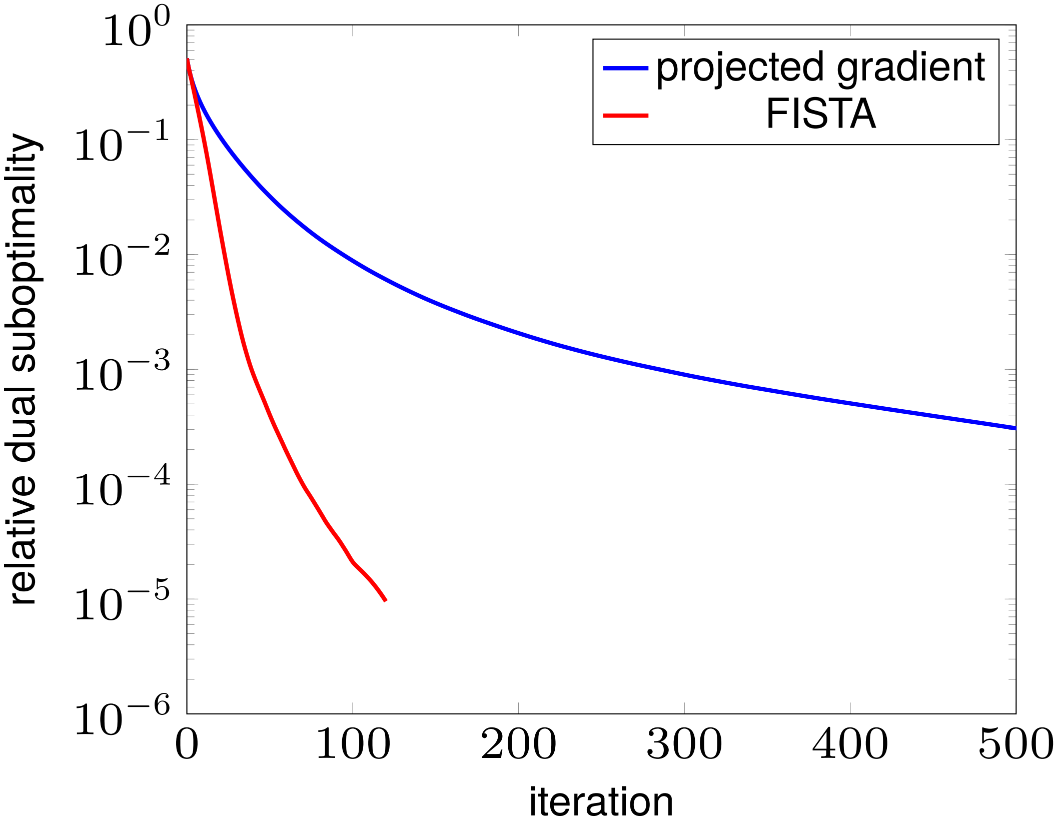
Dual gradient projection update (for strongly convex f):

$$\begin{array}{rcl} \hat{x} & = & \underset{x}{\operatorname{argmin}} \left(f(x) + (\sum_{i=1}^{p} B_i^T z_i)^T x \right) \\ z_i^+ & = & P_{C_i}\left(z_i + tB_i\hat{x}\right), \quad i = 1, \ldots, p \end{array}$$

- C_i is unit Euclidean norm ball in $\mathbf{R}^{m_i}$, if $B_i \in \mathbf{R}^{m_i \times n}$
- $\hat{x}$-calculation decomposes if f is separable

Example

- we take $f(x) = (1/2)\|Cx - d\|_2^2$
- each iteration requires solution of linear equation with coefficient C^TC
- randomly generated $C \in \mathbf{R}^{2000\times1000}$, $B_i \in \mathbf{R}^{10\times1000}$, $p = 500$



Minimization over intersection of convex sets

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C_1 \cap \cdots \cap C_p \end{array}$$

- f is strongly convex with constant μ
- we assume each set C_i is closed, convex, and easy to project onto
- this is a special case of the problem on page 12-3 with

$$\begin{aligned} g(y_1, \ldots, y_p) &= \delta_{C_1}(y_1) + \cdots + \delta_{C_p}(y_p) \\ A &= \left[\begin{array}{cccc} I & I & \cdots & I \end{array} \right]^T \end{aligned}$$

with this choice of g and A,

$$f(x) + g(Ax) = f(x) + \delta_{C_1}(x) + \cdots + \delta_{C_p}(x)$$

Dual problem

primal: minimize $f(x) + \delta_{C_1}(x) + \cdots + \delta_{C_p}(x)$

dual: maximize $-\delta^*_{C_1}(z_1) - \cdots - \delta^*_{C_p}(z_p) - f^*(-z_1 - \cdots - z_p)$

- proximal mapping of $\delta^*_{C_i}$: from Moreau decomposition (page 8-18),

$$\mathrm{prox}_{t\delta^*_{C_i}}(u) = u - tP_{C_i}(u/t)$$

- gradient of $h(z_1, \ldots, z_p) = f^*(-z_1 - \cdots - z_p)$:

$$\nabla h(z) = -A\nabla f(-A^T z) = -\begin{bmatrix} I \\ \vdots \\ I \end{bmatrix} \nabla f^*(-z_1 - \cdots - z_p)$$

- $\nabla h(z)$ is Lipschitz continuous with constant $\|A\|_2^2/\mu = p/\mu$

Dual proximal gradient method

primal: minimize $f(x) + \delta_{C_1}(x) + \cdots + \delta_{C_p}(x)$

dual: maximize $-\delta^*_{C_1}(z_1) - \cdots - \delta^*_{C_p}(z_p) - f^*(-z_1 - \cdots - z_p)$

- dual proximal gradient update

$$
\begin{aligned}
s &= -z_1 - \cdots - z_p \\
z_i^+ &= z_i + t\nabla f^*(s) - tP_{C_i}\left(t^{-1}z_i + \nabla f^*(s)\right), \quad i = 1, \ldots, p
\end{aligned}
$$

- gradient of f^* can be computed by minimizing the Lagrangian

$$
\begin{aligned}
\hat{x} &= \operatorname*{argmin}_x \left(f(x) + (z_1 + \cdots + z_p)^T x\right) \\
z_i^+ &= z_i + t\hat{x} - tP_{C_i}\left(z_i/t + \hat{x}\right), \quad i = 1, \ldots, p
\end{aligned}
$$

- stepsize is fixed ($t \leq \mu/p$) or adjusted by backtracking

Euclidean projection on intersection of convex sets

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|x-a\|_2^2 \\ \text{subject to} & x \in C_1 \cap \cdots \cap C_p \end{array}$$

- special case of previous problem with

$$f(x) = \frac{1}{2}\|x-a\|_2^2, \qquad f^*(u) = \frac{1}{2}\|u\|_2^2 + a^T u$$

- strong convexity constant $\mu = 1$; hence stepsize $t = 1/p$ works
- dual proximal gradient update (with change of variable $w_i = pz_i$):

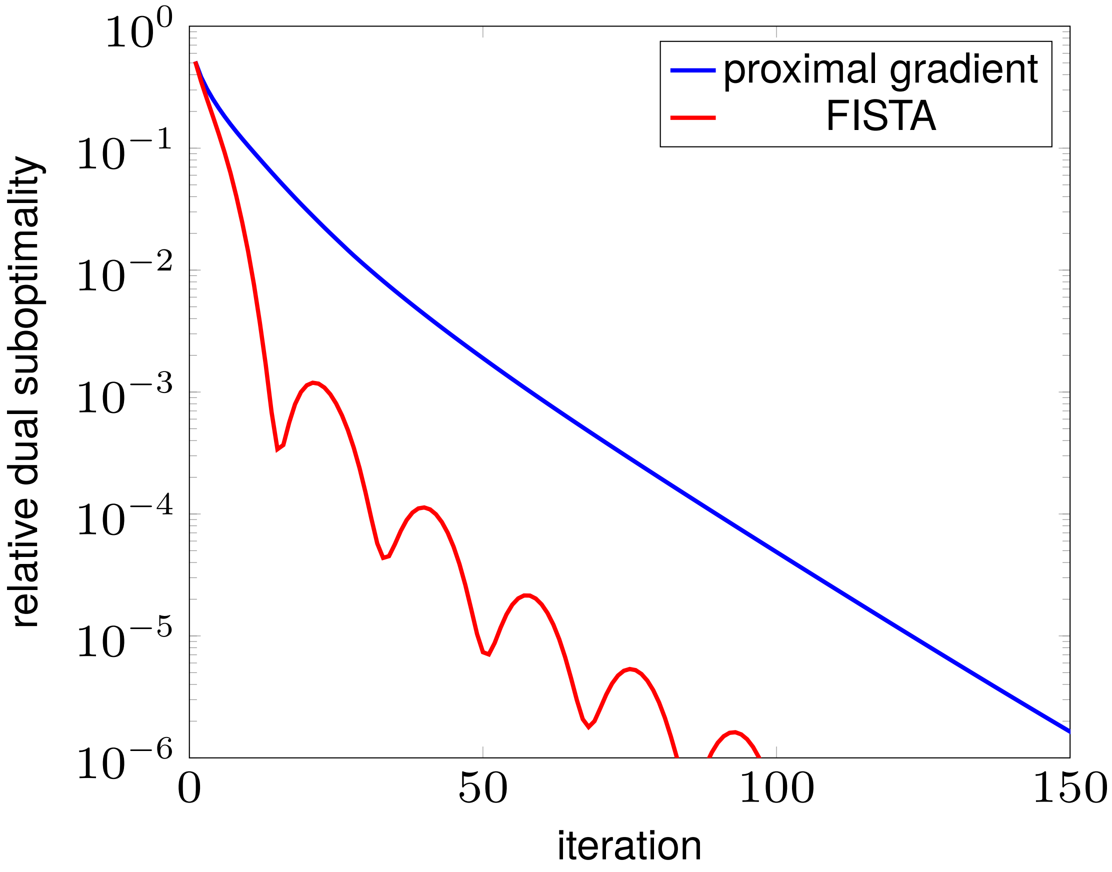
$$\begin{array}{rcl} \hat{x} & = & a - \frac{1}{p}(w_1 + \cdots + w_p) \\ w_i^+ & = & w_i + \hat{x} - P_{C_i}(w_i + \hat{x}), \quad i = 1, \ldots, p \end{array}$$

- the p projections in the second step can be computed in parallel

Nearest positive semidefinite unit-diagonal Z-matrix

projection in Frobenius norm of $A \in \mathbf{S}^{100}$ on the intersection of two sets:

$$C_1 = \mathbf{S}_+^{100}, \qquad C_2 = \{X \in \mathbf{S}^{100} \mid \mathbf{diag}(X) = \mathbf{1},\ X_{ij} \leq 0 \text{ for } i \neq j\}$$

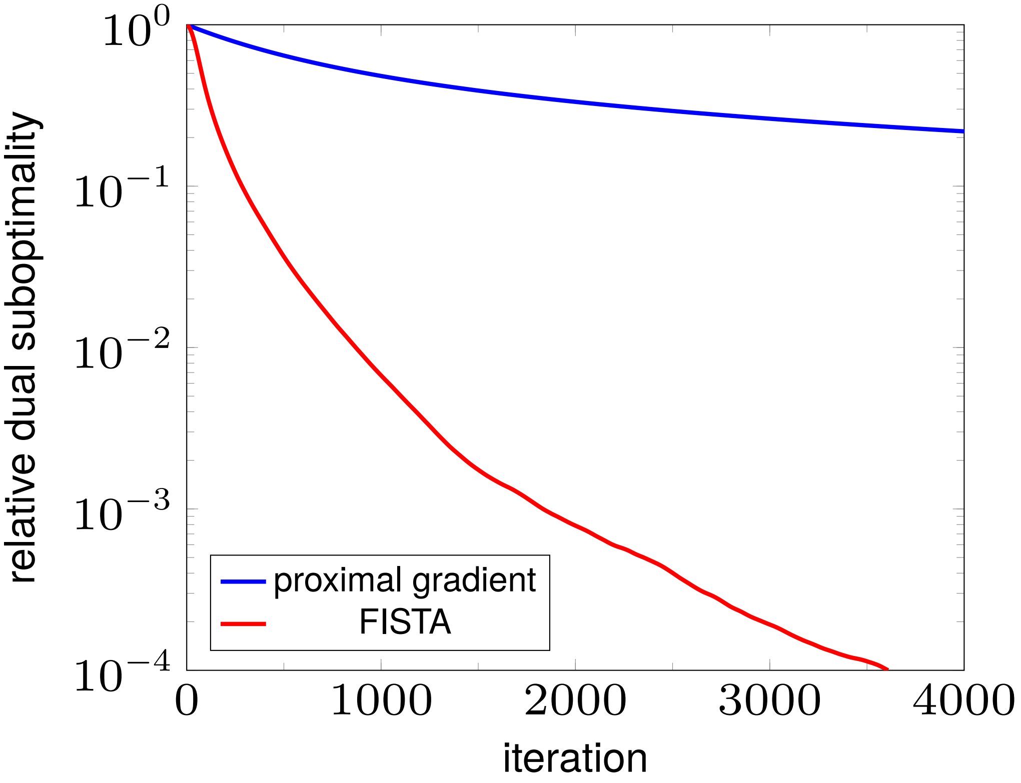


Euclidean projection on polyhedron

- intersection of p halfspaces $C_i = \{x \mid a_i^T x \le b_i\}$

$$P_{C_i}(x) = x - \frac{\max\{a_i^T x - b_i, 0\}}{\|a_i\|_2^2} a_i$$

- example with $p = 2000$ inequalities and $n = 1000$ variables



Decomposition of primal-dual separable problems

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

- special case of $f(x) + g(Ax)$ with (block-)separable f and g
- for example,

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{n} f_j(x_j) \\ \text{subject to} & \sum_{j=1}^{n} A_{1j}x_j \in C_1 \\ & \cdots \\ & \sum_{j=1}^{n} A_{mj}x_j \in C_m \end{array}$$

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

Decomposition of primal-dual separable problems

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

dual: maximize $\displaystyle -\sum_{i=1}^{m} g_i^*(z_i) - \sum_{j=1}^{n} f_j^*(-A_{1j}^T z_1 - \cdots - A_{mj}^T z_j)$

Dual proximal gradient update

$$\begin{aligned}
\hat{x}_j &= \operatorname*{argmin}_{x_j} \big(f_j(x_j) + \sum_{i=1}^{m} z_i^T A_{ij} x_j\big), \quad j = 1, \ldots, n \\
z_i^+ &= \operatorname{prox}_{tg_i^*}\big(z_i + t\sum_{j=1}^{n} A_{ij}\hat{x}_j\big), \quad i = 1, \ldots, m
\end{aligned}$$

Outline

- proximal gradient method applied to the dual
- examples
- **alternating minimization method**

Separable structure with one strongly convex term

$$\text{minimize} \quad f_1(x_1) + f_2(x_2) + g(A_1x_1 + A_2x_2)$$

- composite problem with separable f (two terms, for simplicity)
- if f_1 and f_2 are strongly convex, dual method of page 12-4 applies

$$\begin{array}{rcl}\hat{x}_1 & = & \underset{x_1}{\operatorname{argmin}} \left(f_1(x_1) + z^T A_1 x_1\right) \\ \hat{x}_2 & = & \underset{x_2}{\operatorname{argmin}} \left(f_2(x_2) + z^T A_2 x_2\right) \\ z^+ & = & \operatorname{prox}_{tg^*}(z + t(A_1\hat{x}_1 + A_2\hat{x}_2))\end{array}$$

- we now assume that one function (f_2) is not strongly convex

Separable structure with one strongly convex term

primal: minimize $f_1(x_1) + f_2(x_2) + g(A_1x_1 + A_2x_2)$

dual: maximize $-g^*(z) - f_1^*(-A_1^Tz) - f_2^*(-A_2^Tz)$

- we split dual objective in components $-f_1^*(-A_1^Tz)$ and $-g^*(z) - f_2^*(-A_2^Tz)$
- component $f_1^*(-A_1^Tz)$ is differentiable with Lipschitz continuous gradient
- proximal mapping of $h(z) = g^*(z) + f_2^*(-A_2^Tz)$ was discussed on page 10-7:

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

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

$$(\hat{x}_2, \hat{y}) = \underset{x_2,y}{\mathrm{argmin}}\,(f_2(x_2) + g(y) + \frac{t}{2}\|A_2x_2 - y + w/t\|_2^2)$$

Dual proximal gradient method

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

- evaluate ∇f_1^* by minimizing partial Lagrangian:

$$\begin{array}{rcl} \hat{x}_1 & = & \underset{x_1}{\mathrm{argmin}}\,(f_1(x_1) + z^T A_1 x_1) \\ z^+ & = & \mathrm{prox}_{th}(z + tA_1\hat{x}_1) \end{array}$$

- evaluate $\mathrm{prox}_{th}(z + tA_1\hat{x}_1)$ by minimizing augmented Lagrangian:

$$\begin{array}{rcl} (\hat{x}_2, \hat{y}) & = & \underset{x_2, y}{\mathrm{argmin}}\,(f_2(x_2) + g(y) + \frac{t}{2}\|A_2x_2 - y + z/t + A_1\hat{x}\|_2^2) \\ z^+ & = & z + t(A_1\hat{x}_1 + A_2\hat{x}_2 - \hat{y}) \end{array}$$

Alternating minimization method

starting at some initial z, repeat the following iteration

1. minimize the Lagrangian over x_1:

$$\hat{x}_1 = \underset{x_1}{\operatorname{argmin}} \left(f_1(x_1) + z^T A_1 x_1 \right)$$

2. minimize the augmented Lagrangian over $\hat{x}_2$, $\hat{y}$:

$$(\hat{x}_2, \hat{y}) = \underset{x_2, y}{\operatorname{argmin}} \left(f_2(x_2) + g(y) + \frac{t}{2} \|A_1 \hat{x}_1 + A_2 x_2 - y + z/t\|_2^2 \right)$$

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 (for problem on page 12-19)

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

$$f_1(x_1) + f_2(x_2) + g(y) + \frac{t}{2}\|A_1x_1 + A_2x_2 - y + z/t\|_2^2$$

2. update dual variable:

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

Differences with alternating minimization (dual proximal gradient method)

- augmented Lagrangian method does not require strong convexity of f_1
- there is no upper limit on the step size t in augmented Lagrangian method
- quadratic term in step 1 of AL method destroys separability of $f_1(x_1) + f_2(x_2)$

Example

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x_1^T P x_1 + q_1^T x_1 + q_2^T x_2 \\ \text{subject to} & B_1 x_1 \preceq d_1, \quad B_2 x_2 \preceq d_2 \\ & A_1 x_1 + A_2 x_2 = b \end{array}$$

- without equality constraint, problem would separate in independent QP and LP
- we assume $P \succ 0$

Formulation for dual decomposition

$$\begin{array}{ll} \text{minimize} & f_1(x_1) + f_2(x_2) \\ \text{subject to} & A_1 x_1 + A_2 x_2 = b \end{array}$$

- first function is strongly convex

$$f_1(x) = \frac{1}{2}x_1^T P x_1 + q_1^T x_1, \qquad \operatorname{dom} f_1 = \{x_1 \mid B_1 x_1 \preceq d_1\}$$

- second function is not: $f_2(x) = q_2^T x_2$ with domain $\{x_2 \mid B_2 x_2 \preceq d_2\}$

Example

Alternating minimization algorithm

1. compute the solution $\hat{x}_1$ of the QP

$$\begin{array}{ll} \text{minimize} & (1/2)x_1^T P_1 x_1 + (q_1 + A_1^T z)^T x_1 \\ \text{subject to} & B_1 x_1 \preceq d_1 \end{array}$$

2. compute the solution $\hat{x}_2$ of the QP

$$\begin{array}{ll} \text{minimize} & (q_2 + A_2^T z)^T x_2 + (t/2)\|A_1\hat{x}_1 + A_2 x_2 - b\|_2^2 \\ \text{subject to} & B_2 x_2 \preceq d_2 \end{array}$$

3. dual update:

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

References

- 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).