

10. 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

primal: minimize $f(x) + g(Ax)$

dual: maximize $-g^*(z) - f^*(-A^T z)$

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

- f is strongly convex: this implies $f^*(-A^T z)$ has a 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 5.19)

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

Dual proximal gradient update

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

- proximal gradient update:

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

- ∇f^* can be computed by minimizing partial Lagrangian (from p. 5.15, p. 5.19):

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

$$z^+ = \text{prox}_{tg^*}(z + tA\hat{x})$$

- 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 7

Dual proximal gradient update

$$\hat{x} = \underset{x}{\operatorname{argmin}} (f(x) + z^T Ax)$$
$$z^+ = \operatorname{prox}_{tg^*}(z + tA\hat{x})$$

- Moreau decomposition gives alternate expression for z -update:

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

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

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

Alternating minimization interpretation

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

$$\hat{y} = \operatorname{argmin}_y (g(y) - z^T y + \frac{t}{2} \|A\hat{x} - y\|_2^2)$$

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

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

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

- 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

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

(see page 5.23)

- 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 10.3 with $g(y) = \|y - b\|$:

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

Dual gradient projection

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

- dual gradient projection update ($C = \{z \mid \|z\|_* \leq 1\}$):

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

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

$$\begin{aligned} \hat{x} &= \underset{x}{\operatorname{argmin}} (f(x) + z^T Ax) \\ z^+ &= P_C(z + t(A\hat{x} - b)) \end{aligned}$$

Example

$$\text{primal:} \quad \text{minimize} \quad f(x) + \sum_{i=1}^p \|B_i x\|_2$$

$$\begin{aligned} \text{dual:} \quad & \text{maximize} \quad -f^*(-B_1^T z_1 - \dots - B_p^T z_p) \\ & \text{subject to} \quad \|z_i\|_2 \leq 1, \quad i = 1, \dots, p \end{aligned}$$

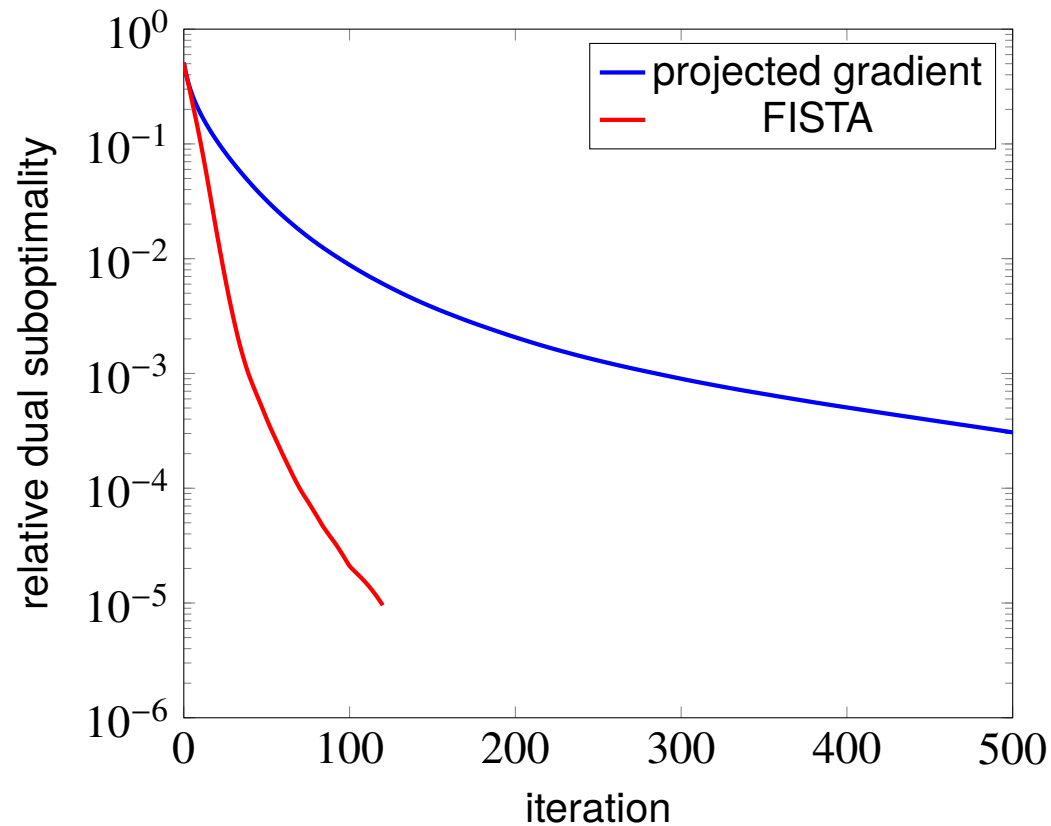
Dual gradient projection update (for strongly convex f):

$$\begin{aligned} \hat{x} &= \underset{x}{\operatorname{argmin}} (f(x) + (\sum_{i=1}^p B_i^T z_i)^T x) \\ z_i^+ &= P_{C_i}(z_i + t B_i \hat{x}), \quad i = 1, \dots, p \end{aligned}$$

- 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 $C^T C$
- randomly generated $C \in \mathbf{R}^{2000 \times 1000}$, $B_i \in \mathbf{R}^{10 \times 1000}$, $p = 500$



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 10.3 with

$$\begin{aligned} g(y_1, \dots, y_p) &= \delta_{C_1}(y_1) + \cdots + \delta_{C_p}(y_p) \\ A &= \begin{bmatrix} I & I & \cdots & I \end{bmatrix}^T \end{aligned}$$

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 6.18),

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

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

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

- $\nabla h(z)$ 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

$$s = -z_1 - \cdots - z_p$$
$$z_i^+ = z_i + t \nabla f^*(s) - t P_{C_i}(t^{-1} z_i + \nabla f^*(s)), \quad i = 1, \dots, p$$

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

$$\hat{x} = \operatorname{argmin}_x (f(x) + (z_1 + \cdots + z_p)^T x)$$
$$z_i^+ = z_i + t \hat{x} - t P_{C_i}(z_i/t + \hat{x}), \quad i = 1, \dots, p$$

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

Euclidean projection on intersection of convex sets

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

- special case of previous problem with

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

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

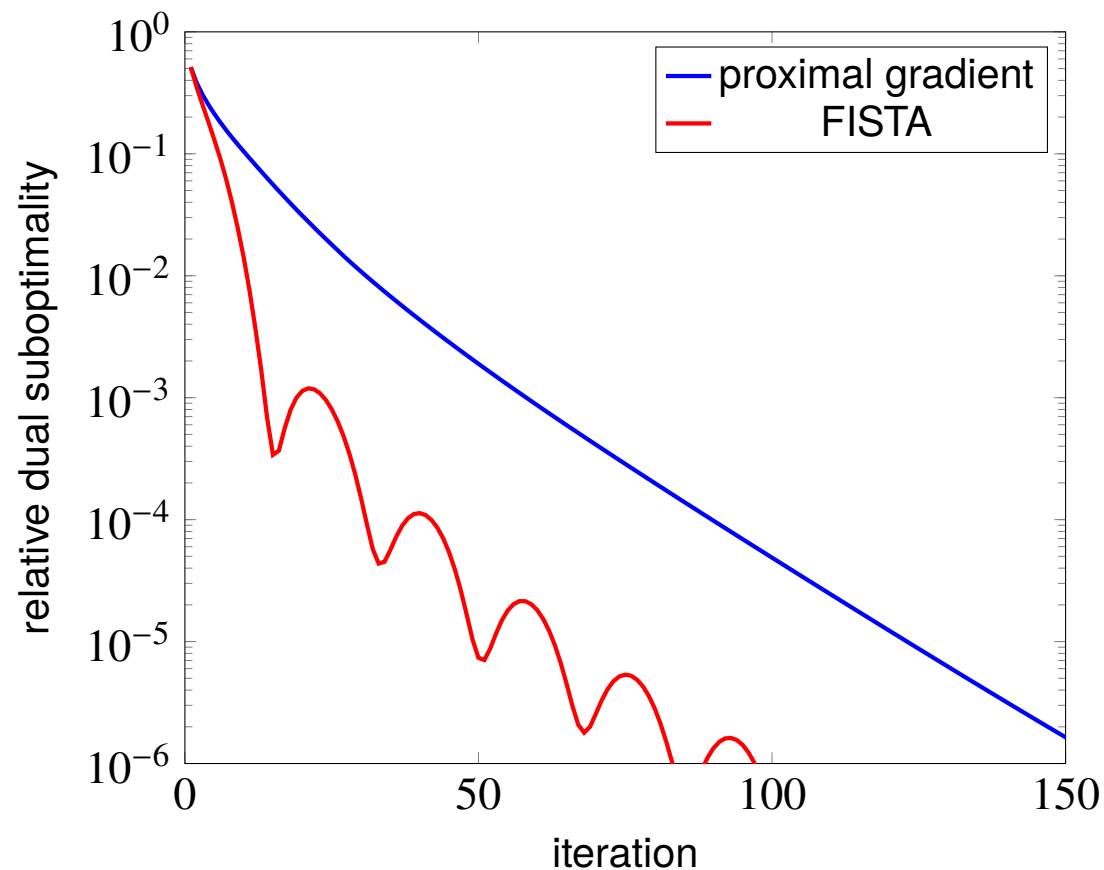
$$\begin{aligned} \hat{x} &= a - \frac{1}{p}(w_1 + \cdots + w_p) \\ w_i^+ &= w_i + \hat{x} - P_{C_i}(w_i + \hat{x}), \quad i = 1, \dots, p \end{aligned}$$

- 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:

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

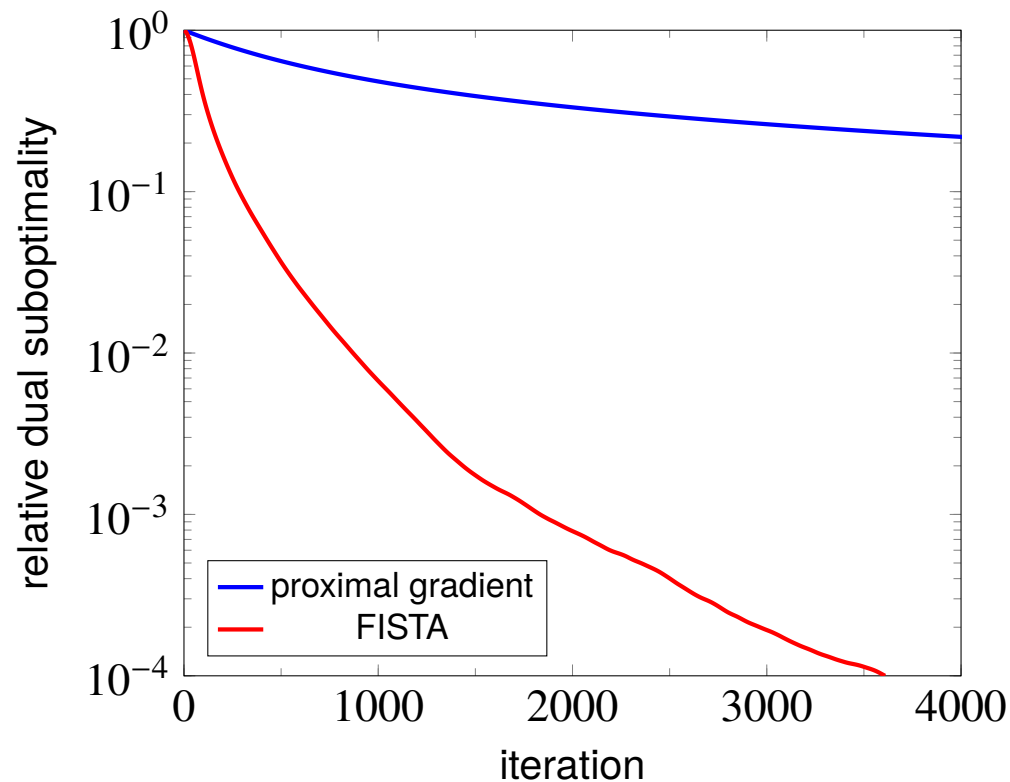


Euclidean projection on polyhedron

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

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

- 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,

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

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

Decomposition of primal-dual separable problems

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

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

Dual proximal gradient update

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

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

Outline

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

Separable structure with one strongly convex term

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

- composite problem with separable f (two terms, for simplicity)
- if f_1 and f_2 are strongly convex, dual method of page 10.4 applies

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

$$\hat{x}_2 = \underset{x_2}{\operatorname{argmin}} (f_2(x_2) + z^T A_2 x_2)$$

$$z^+ = \operatorname{prox}_{tg^*}(z + t(A_1 \hat{x}_1 + A_2 \hat{x}_2))$$

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

Separable structure with one strongly convex term

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

$$\text{dual:} \quad \text{maximize} \quad -g^*(z) - f_1^*(-A_1^T z) - f_2^*(-A_2^T z)$$

- we split dual objective in components $-f_1^*(-A_1^T z)$ and $-g^*(z) - f_2^*(-A_2^T z)$
- component $f_1^*(-A_1^T z)$ is differentiable with Lipschitz continuous gradient
- proximal mapping of $h(z) = g^*(z) + f_2^*(-A_2^T z)$ was discussed on page 8.7:

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

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

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

Dual proximal gradient method

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

- evaluate ∇f_1^* by minimizing partial Lagrangian:

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

$$z^+ = \text{prox}_{th}(z + tA_1 \hat{x}_1)$$

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

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

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

Alternating minimization method

starting at some initial z , repeat the following iteration

1. minimize the Lagrangian over x_1 :

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

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

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

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 10.19)

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

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

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

Formulation for dual decomposition

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

- first function is strongly convex

$$f_1(x) = \frac{1}{2}x_1^T P x_1 + q_1^T x_1, \quad \text{dom } f_1 = \{x_1 \mid B_1 x_1 \leq d_1\}$$

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

Example

Alternating minimization algorithm

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

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

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

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

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).