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

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 Lipschitz continuous gradient

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

$\mu$ 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

minimize $g^*(z) + f^*(-A^Tz)$

- proximal gradient update:

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

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

  $$\hat{x} = \arg\min_x (f(x) + z^TAx)$$

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

- 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 11
Dual proximal gradient update

\[ \hat{x} = \arg\min_x (f(x) + z^T A x) \]

\[ z^+ = \prox_{t g^*}(z + t A \hat{x}) \]

- Moreau decomposition gives alternate expression for \( z \)-update:

\[ z^+ = z + t A \hat{x} - t \prox_{t^{-1} g}(t^{-1}z + A \hat{x}) \]

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

\[ \hat{y} = \prox_{t^{-1} g}(t^{-1}z + A \hat{x}) \]

\[ = \arg\min_y (g(y) + \frac{t}{2} \| A \hat{x} - t^{-1}z - y \|_2^2) \]

\[ = \arg\min_y (g(y) + z^T (A \hat{x} - y) + \frac{t}{2} \| A \hat{x} - y \|_2^2) \]
Alternating minimization interpretation

\[ \hat{x} = \arg\min_x (f(x) + z^T Ax) \]

\[ \hat{y} = \arg\min_y (g(y) - z^T y + \frac{t}{2} \| A\hat{x} - y \|^2_2) \]

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

- first minimize Lagrangian over \( x \), then augmented Lagrangian over \( y \)

- compare with augmented Lagrangian method:

\[ (\hat{x}, \hat{y}) = \arg\min_{x,y} (f(x) + g(y) + z^T (Ax - y) + \frac{t}{2} \| Ax - y \|^2_2) \]

- 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

primal: minimize $f(x) + \|Ax - b\|

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

subject to $\|z\|_* \leq 1$

(see page 7-20)

• we assume $f$ is strongly convex with constant $\mu$, 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\|:

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

primal: minimize \( f(x) + \|Ax - b\| \)

dual: maximize \(-b^T z - f^*(-A^T z)\)
subject to \( \|z\|_* \leq 1 \)

• dual gradient projection update:

\[
z^+ = P_C (z + t(A \nabla f^*(-A^T z) - b))
\]

• gradient of \( f^* \) can be computed by minimizing the partial Lagrangian:

\[
\hat{x} = \arg \min_x (f(x) + z^T Ax)
\]

\[
z^+ = P_C (z + t(A \hat{x} - b))
\]
Example

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

dual: maximize \(-f^*(-B_1^T z_1 - \cdots - B_p^T z_p)\)

subject to \( \|z_i\|_2 \leq 1, \ i = 1, \ldots, p \)

**Dual gradient projection update** (for strongly convex \( f \)):

\[
\hat{x} = \arg\min_x \left( f(x) + \sum_{i=1}^{p} B_i^T z_i^T x \right)
\]

\[
z_i^+ = P_{C_i} (z_i + tB_i \hat{x}), \ i = 1, \ldots, p
\]

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

Dual proximal gradient method 12-9
Example

- we take $f(x) = (1/2)\|Cx - d\|_2^2$
- each iteration requires solution of linear equation with coefficient $C^TC$
- randomly generated $C \in \mathbb{R}^{2000 \times 1000}, B_i \in \mathbb{R}^{10 \times 1000}, p = 500$
Minimization over intersection of convex sets

\[ \begin{align*} 
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in C_1 \cap \cdots \cap C_p 
\end{align*} \]

- \( f \) is strongly convex with constant \( \mu \)
- 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

\[ g(y_1, \ldots, y_p) = \delta_{C_1}(y_1) + \cdots + \delta_{C_p}(y_p) \]

\[ A = \begin{bmatrix} I & I & \cdots & I \end{bmatrix}^T \]

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),

\[
\text{prox}_{t\delta^*_{C_i}}(u) = u - tP_{C_i}(u/t)
\]

- 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

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

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

\[
\begin{align*}
  \hat{x} &= \arg\min_x \left( f(x) + (z_1 + \cdots + z_p)^T x \right) \\
  z_i^+ &= z_i + t \hat{x} - t P_{C_i} \left( z_i / t + \hat{x} \right), \quad i = 1, \ldots, p
\end{align*}
\]

- stepsize is fixed (\( t \leq \mu / p \)) or adjusted by backtracking
Euclidean projection on intersection of convex sets

minimize \( \frac{1}{2}\|x - a\|_2^2 \)
subject to \( x \in C_1 \cap \cdots \cap C_p \)

- 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 \( w_i = pz_i \)):

\[
\hat{x} = a - \frac{1}{p}(w_1 + \cdots + w_p)
\]

\[
w_i^+ = w_i + \hat{x} - P_{C_i}(w_i + \hat{x}), \quad i = 1, \ldots, p
\]

- the \( p \) projections in the second step can be computed in parallel
Nearest positive semidefinite unit-diagonal Z-matrix projection in Frobenius norm of $A \in S^{100}$ on the intersection of two sets:

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

\[
\text{minimize} \quad \sum_{j=1}^{n} f_j(x_j)
\]

subject to \( \sum_{j=1}^{n} A_{1j}x_j \in C_1 \)

\[ \cdots \]

\( \sum_{j=1}^{n} A_{mj}x_j \in C_m \)

- we assume each \( f_i \) is strongly convex; each \( g_i \) has inexpensive prox-operator
Decomposition of primal-dual separable problems

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

dual: maximize \( -\sum_{i=1}^{m} g^*_i(z_i) - \sum_{j=1}^{n} f^*_j(-A_{1j}^Tz_1 - \cdots - A_{mj}^Tz_j) \)

Dual proximal gradient update

\[ \hat{x}_j = \arg\min_{x_j} (f_j(x_j) + \sum_{i=1}^{m} z_i^T A_{ij}x_j), \quad j = 1, \ldots, n \]

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

• proximal gradient method applied to the dual

• examples

• alternating minimization method
Separable structure with one strongly convex term

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 12-4 applies

$$
\hat{x}_1 = \arg\min_{x_1} (f_1(x_1) + z^T A_1 x_1) \\
\hat{x}_2 = \arg\min_{x_2} (f_2(x_2) + z^T A_2 x_2) \\
z^+ = \text{prox}_{t g^*}(z + t(A_1 \hat{x}_1 + A_2 \hat{x}_2))
$$

• we now assume that one function ($f_2$) is not strongly convex
Separable structure with one strongly convex term

primal: \( \text{minimize } f_1(x_1) + f_2(x_2) + g(A_1 x_1 + A_2 x_2) \)

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

- 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 10-7:

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

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

\[
(\hat{x}_2, \hat{y}) = \arg\min_{x_2, y} (f_2(x_2) + g(y) + \frac{t}{2} \| A_2 x_2 - y + w/t \|_2^2)
\]
**Dual proximal gradient method**

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

- evaluate $\nabla f_1^*$ by minimizing partial Lagrangian:

$$\hat{x}_1 = \arg\min_{x_1} (f_1(x_1) + z^T A_1 x_1)$$

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

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

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

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

starting at some initial \( z \), repeat the following iteration

1. minimize the Lagrangian over \( x_1 \):

\[
\hat{x}_1 = \arg\min_{x_1} (f_1(x_1) + z^T A_1 x_1)
\]

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

\[
(\hat{x}_2, \hat{y}) = \arg\min_{x_2, y} \left( f_2(x_2) + g(y) + \frac{t}{2} \| A_1 \hat{x}_1 + A_2 x_2 - y + z/t \|_2^2 \right)
\]

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 12-19)

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

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

2. update dual variable:

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

Differences with alternating minimization (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

\[
\text{minimize } \frac{1}{2} x_1^T P x_1 + q_1^T x_1 + q_2^T x_2 \\
\text{subject to } B_1 x_1 \preceq d_1, \quad B_2 x_2 \preceq d_2, \\
A_1 x_1 + A_2 x_2 = b
\]

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

Formulation for dual decomposition

\[
\text{minimize } f_1(x_1) + f_2(x_2) \\
\text{subject to } A_1 x_1 + A_2 x_2 = b
\]

- first function is strongly convex

\[
f_1(x) = \frac{1}{2} x_1^T P x_1 + q_1^T x_1, \quad \text{dom } f_1 = \{ x_1 | B_1 x_1 \preceq d_1 \}
\]

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

Alternating minimization algorithm

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

\[
\text{minimize } \frac{1}{2} x_1^TP_1x_1 + (q_1 + A_1^Tz)^T x_1 \\
\text{subject to } B_1x_1 \preceq d_1
\]

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

\[
\text{minimize } (q_2 + A_2^Tz)^T x_2 + (t/2)\|A_1\hat{x}_1 + A_2x_2 - b\|^2_2 \\
\text{subject to } B_2x_2 \preceq d_2
\]

3. dual update:

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