14. Primal-dual proximal methods

- primal-dual optimality conditions
- monotone operators
- proximal point algorithm
- Chambolle-Pock algorithm
- Douglas-Rachford operator splitting
Primal and dual problem

primal: minimize $f(x) + g(Ax)$
dual: maximize $-g^*(z) - f^*(-A^T z)$

- $f$ and $g$ are closed convex functions
- dual problem is Lagrange dual of reformulated problem

$$\text{minimize } f(x) + g(y)$$
$$\text{subject to } Ax = y$$

Optimality (KKT) conditions

- primal feasibility: $x \in \text{dom } f$ and $Ax = y \in \text{dom } g$
- $(x, y)$ is a minimizer of the Lagrangian $f(x) + g(y) + z^T(Ax - y)$:
  $$-A^T z \in \partial f(x), \quad z \in \partial g(y) \quad \text{(equivalently, } y \in \partial g^*(z))$$
Primal-dual optimality conditions

- the optimality conditions can be written symmetrically as

\[
0 \in \begin{bmatrix}
0 & A^T \\
-A & 0
\end{bmatrix} \begin{bmatrix}
x \\
z
\end{bmatrix} + \begin{bmatrix}
\partial f(x) \\
\partial g^*(z)
\end{bmatrix}
\]

- second term on right-hand side denotes the product set

\[
\partial f(x) \times \partial g^*(z) = \{(u, v) \mid u \in \partial f(x), v \in \partial g^*(z)\}
\]

- solutions are saddle points of convex-concave function

\[
f(x) - g^*(z) + z^T Ax
\]

in this lecture we assume that the optimality conditions are solvable
(a sufficient condition is that primal is solvable and Slater’s condition holds)
Outline

• primal-dual optimality conditions

• monotone operators

• resolvent

• proximal point algorithm

• Chambolle-Pock algorithm

• Douglas-Rachford operator splitting
Multivalued (set-valued) operator

**Definition:** operator $F$ maps vectors $x \in \mathbb{R}^n$ to sets $F(x) \subseteq \mathbb{R}^n$

- the domain and graph of $F$ are defined as
  
  \[
  \text{dom } F = \{ x \in \mathbb{R}^n \mid F(x) \neq \emptyset \} \\
  \text{gr}(F) = \{ (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid x \in \text{dom } F, y \in F(x) \} 
  \]

- if $F(x)$ is a singleton, we write $F(x) = y$ instead of $F(x) = \{y\}$

**Example:** sign operator

\[
F(x) = \begin{cases} 
  -1 & x < 0 \\
  [-1, 1] & x = 0 \\
  1 & x > 0 
\end{cases}
\]
Transformations as operations on graph

**Inverse:** $F^{-1}(x) = \{y \mid x \in F(y)\}$

$$\text{gr}(F^{-1}) = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \text{gr}(F)$$

**Composition with scaling:** $(\lambda F)(x) = \lambda F(x)$ and $(F \lambda)(x) = F(\lambda x)$

$$\text{gr}(\lambda F) = \begin{bmatrix} I & 0 \\ 0 & \lambda I \end{bmatrix} \text{gr}(F), \quad \text{gr}(F \lambda) = \begin{bmatrix} (1/\lambda)I & 0 \\ 0 & I \end{bmatrix} \text{gr}(F)$$

**Addition to identity:** $(I + \lambda F')(x) = \{x + \lambda y \mid y \in F(x)\}$

$$\text{gr}(I + \lambda F') = \begin{bmatrix} I & 0 \\ I & \lambda I \end{bmatrix} \text{gr}(F')$$

note that these are all *linear* operations on the graph
**Monotone operator**

**Definition:** $F$ is a monotone operator if

$$(y - \hat{y})^T(x - \hat{x}) \geq 0 \quad \forall x, \hat{x} \in \text{dom } F, \ y \in F(x), \ \hat{y} \in F(\hat{x})$$

in terms of the graph,

$$\begin{bmatrix} x - \hat{x} \\ y - \hat{y} \end{bmatrix}^T \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} x - \hat{x} \\ y - \hat{y} \end{bmatrix} \geq 0 \quad \forall (x, y), \ (\hat{x}, \hat{y}) \in \text{gr}(F)$$

**Monotone inclusion problem:** find $x \in F^{-1}(0)$, i.e., solve

$$0 \in F(x)$$

includes many equilibrium/optimality conditions as special cases
Examples

we will encounter the following three types of monotone operators

- subdifferentials $\partial f(x)$ of convex functions $f$

- affine monotone operators: $F(x) = Cx + d$ is monotone if

$$C + C^T \succeq 0$$

- sums of the above; in particular,

$$F(x, z) = \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix}$$
Maximal monotone operator

graph is not properly contained in the graph of another monotone operator

$$F(x)$$  

maximal monotone

$$F(x)$$  

monotone, but not maximal monotone
Conditions for maximal monotonicity

- the subdifferential of a closed convex function is maximal monotone
- affine monotone operators are maximal monotone
- (Minty) a monotone operator $F$ is maximal monotone if and only if

$$\text{im}(I + F) = \bigcup_{x \in \text{dom } F} (x + F(x)) = \mathbb{R}^n$$

i.e., for every $y \in \mathbb{R}^n$, there exists an $x$ such that $y \in x + F(x)$

- sums $F + G$ of maximal monotone operators are not necessarily maximal
  (sufficient condition: $\text{int dom } F \cap \text{dom } G \neq \emptyset$)
Coercivity (strong monotonicity)

$F$ is **coercive** with parameter $\mu > 0$ if

$$(y - \hat{y})^T(x - \hat{x}) \geq \mu \|x - \hat{x}\|^2 \quad \forall x, \hat{x} \in \text{dom } F, \ y \in F(x), \ \hat{y} \in F(\hat{x})$$

- $F - \mu I$ is a monotone operator
- equivalently,

$$\begin{bmatrix} x - \hat{x} \\ y - \hat{y} \end{bmatrix}^T \begin{bmatrix} -2\mu I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} x - \hat{x} \\ y - \hat{y} \end{bmatrix} \geq 0 \quad \forall (x, y), (\hat{x}, \hat{y}) \in \text{gr}(F)$$

**Examples**

- subdifferential of strongly convex function
- affine operator $F(x) = Ax + b$ if $A + A^T > 0$ (with $\mu = \lambda_{\min}(A + A^T)/2$)
Co-coercivity

$F$ is **co-coercive** with parameter $\gamma > 0$ if $F^{-1}$ is coercive:

$$(F(x) - F(\hat{x}))^T (x - \hat{x}) \geq \gamma \|F(x) - F(\hat{x})\|_2^2 \quad \forall x, \hat{x} \in \text{dom } F$$

- equivalently,

$$\begin{bmatrix} x - \hat{x} \\ y - \hat{y} \end{bmatrix}^T \begin{bmatrix} 0 & I \\ I & -2\gamma I \end{bmatrix} \begin{bmatrix} x - \hat{x} \\ y - \hat{y} \end{bmatrix} \geq 0 \quad \forall (x, y), (\hat{x}, \hat{y}) \in \text{gr}(F')$$

- $F$ is **firmly nonexpansive** if it is co-coercive with $\gamma = 1$

**Example:** affine operator $F(x) = Ax + b$ with

$$A + A^T \succeq 2\gamma A^T A \iff \|2\gamma A - I\|_2 \leq 1$$

for symmetric positive definite $A$, this means $\lambda_{\text{max}}(A) \leq 1/\gamma$
Lipschitz continuity

- \( F \) is **Lipschitz continuous** with parameter \( L \) if

\[
\|F(x) - F(\hat{x})\|_2 \leq L\|x - \hat{x}\|_2 \quad \forall x, \hat{x} \in \text{dom } F
\]

- \( F \) is **nonexpansive** if it is Lipschitz continuous with \( L = 1 \)

**Example:** any affine \( F(x) = Ax + b \) (parameter \( L = \|A\|_2 \))

**Relation to co-coercivity**

- co-coercivity implies Lipschitz continuity (with \( L = 1/\gamma \))
- Lipschitz continuity does not imply co-coercivity (see homework 1)
- properties are equivalent for gradients of closed convex functions (page 1-15)
Outline

• primal-dual optimality conditions

• monotone operators

• **resolvent**

• proximal point algorithm

• Chambolle-Pock algorithm

• Douglas-Rachford operator splitting
Resolvent

the **resolvent** of an operator $F$ is the operator

$$(I + \lambda F)^{-1} \quad (\text{with } \lambda > 0)$$

- inverse denotes the operator inverse:

  $$y \in (I + \lambda F)^{-1}(x) \iff x - y \in \lambda F(y)$$

- graph of resolvent is a linear transformation of graph of $F$:

  $$\text{gr}((I + \lambda F)^{-1}) = \begin{bmatrix} I & \lambda I \\ I & 0 \end{bmatrix} \text{gr}(F)$$
Examples

**Subdifferential:** resolvent is proximal mapping

\[
(I + \lambda \partial f)^{-1}(x) = \text{prox}_{\lambda f}(x)
\]

follows from subgradient characterization of \(\text{prox}_{\lambda f}\) (page 6-7)

\[
y = \text{prox}_{\lambda f}(x) \iff x - y \in \lambda \partial f(y)
\]

**Monotone affine mapping:** resolvent of \(F(x) = Ax + b\) is

\[
(I + \lambda F)^{-1}(x) = (I + \lambda A)^{-1}(x - \lambda b)
\]

- inverse on right-hand side is standard matrix inverse
- \(I + \lambda A\) is nonsingular for all \(\lambda \geq 0\) because \(A + A^T \succeq 0\)
Monotonicity properties

• an operator is monotone if and only if its resolvent is firmly nonexpansive:
  this follows from the matrix identity

\[
\lambda \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} = \begin{bmatrix} I & I \\ \lambda I & 0 \end{bmatrix} \begin{bmatrix} 0 & I \\ I & -2I \end{bmatrix} \begin{bmatrix} I & \lambda I \\ I & 0 \end{bmatrix}
\]

and the expression of the graph of the resolvent on page 14-13

• a monotone operator $F$ is maximal monotone if and only

\[
\text{dom}(I + \lambda F)^{-1} = \mathbb{R}^n
\]

follows from Minty’s theorem on page 14-9
Outline

• primal-dual optimality conditions
• monotone operators
• proximal point algorithm
• Chambolle-Pock algorithm
• Douglas-Rachford operator splitting
Proximal point algorithm

Monotone inclusion problem: given maximal monotone $F$, find $x$ such that

$$0 \in F(x)$$

this is equivalent to finding a fixed point of the resolvent $R_t = (I + tF)^{-1}$ of $F$:

$$x = R_t(x) \iff x \in (I + tF)(x) \iff 0 \in F(x)$$

Proximal-point algorithm: fixed point iteration

$$x^+ = R_t(x)$$

Proximal-point algorithm with relaxation (relaxation parameter $\rho \in (0, 2)$):

$$x^+ = x + \rho(R_t(x) - x)$$
Convergence

if $F^{-1}(0) \neq \emptyset$, proximal point algorithm converges

• with constant $t > 0$ and $\rho \in (0, 2)$

• with $t_k, \rho_k$ varying and bounded away from their limits, i.e.,

$$t_k \geq t_{\min} > 0, \quad 0 < \rho_{\min} \leq \rho_k \leq \rho_{\max} < 2 \quad \text{for all } k$$

proof relies on firm nonexpansiveness of resolvent
Linear change of variables

make a change of variables \( x = Ay \), with \( A \) nonsingular:

\[
G(y) = A^T F(Ay)
\]

- graph of \( G \) is

\[
gr(G) = \begin{bmatrix} A^{-1} & 0 \\ 0 & A^T \end{bmatrix} gr(F)
\]

- (maximal) monotonicity of \( G \) follows from (maximal) monotonicity of \( F \) and

\[
\begin{bmatrix} A^{-1} & 0 \\ 0 & A^T \end{bmatrix}^T \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & A^T \end{bmatrix} = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}
\]
‘Preconditioned’ proximal point algorithm

\[ y^{(k)} = (I + tG)^{-1}(y^{(k-1)}) \]

- \( y^{(k)} \) is the solution of the inclusion problem

\[ \frac{1}{t}(y^{(k-1)} - y) \in A^T F(Ay) \]

- in the original coordinates \( x = Ay \), this can be written as

\[ \frac{1}{t}H(x^{(k-1)} - x) \in F(x) \]

where \( H = A^{-T}A^{-1} \) and \( x^{(k-1)} = Ay^{(k-1)} \)

- we obtain a generalized proximal point update, with \( H \succ 0 \) substituted for \( I \):

\[ x^{(k)} = (H + tF)^{-1}(Hx^{(k-1)}) \]

the purpose is often to make the resolvents cheaper, not preconditioning
Proximal method of multipliers

the proximal point algorithm applied to

\[ F(x, z) = \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix} \]

is known as the proximal method of multipliers

• basic iteration (without relaxation) is

\[(x^{(k)}, z^{(k)}) = (I + tF)^{-1}(x^{(k-1)}, z^{(k-1)})\]

• \((x^{(k)}, z^{(k)})\) is the solution of the monotone inclusion

\[0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix} + \frac{1}{t} \begin{bmatrix} x - x^{(k-1)} \\ z - z^{(k-1)} \end{bmatrix}\]
Evaluation of the resolvent

- equivalent inclusion problem

\[
0 \in \begin{bmatrix}
0 & 0 & A^T \\
0 & 0 & -I \\
-A & I & 0
\end{bmatrix} \begin{bmatrix}
x \\
y \\
z
\end{bmatrix} + \begin{bmatrix}
\partial f(x) \\
\partial g(y)
\end{bmatrix} + \frac{1}{t} \begin{bmatrix}
x - x^{(k-1)} \\
0 \\
z - z^{(k-1)}
\end{bmatrix}
\]

- this is the optimality condition of the optimization problem (variables \(x, y\))

\[
\text{minimize} \quad f(x) + g(y) + \frac{t}{2} \| Ax - y + \frac{1}{t} z^{(k-1)} \|^2 + \frac{1}{2} \left( \frac{1}{2t} \right) \| x - x^{(k-1)} \|^2
\]

(the augmented Lagrangian with an extra penalty term on \(x\))

- from the minimizer \((\hat{x}, \hat{y})\), we make the update

\[
x^{(k)} = \hat{x}, \quad z^{(k)} = z^{(k-1)} + t( A\hat{x} - \hat{y})
\]
Outline

• primal-dual optimality conditions

• monotone operators

• proximal point algorithm

• Chambolle-Pock algorithm

• Douglas-Rachford operator splitting
Chambolle-Pock algorithm

\[ 0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix} \]

Algorithm

\[ x^{(k)} = \text{prox}_{t f}(x^{(k-1)} - t A^T z^{(k-1)}) \]
\[ z^{(k)} = \text{prox}_{s g^*}(z^{(k-1)} + s A (2 x^{(k)} - x^{(k-1)})) \]

- primal and dual step sizes \( t, s \) are positive and satisfy \( st \|A\|_2^2 \leq 1 \)
- each iteration requires evaluations of proximal mappings of \( f \) and \( g^* \)
- also requires multiplications with \( A \), \( A^T \), but no solutions of linear equations
- for \( A = I \), \( s = t = 1 \) this is the Douglas-Rachford algorithm (page 13-8)
Relation to proximal point algorithm

apply ‘preconditioned’ proximal point algorithm of page 14-19 with

$$H = \begin{bmatrix} I & -tA^T \\ -tA & (t/s)I \end{bmatrix}$$

- $H$ is positive definite for $st\|A\|^2 < 1$

- $x^{(k)}$ and $z^{(k)}$ are the solution of

$$\frac{1}{t} \begin{bmatrix} I & -tA^T \\ -tA & (t/s)I \end{bmatrix} \begin{bmatrix} x^{(k-1)} - x \\ z^{(k-1)} - z \end{bmatrix} \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix}$$

- this simplifies to

$$0 \in \partial f(x) + \frac{1}{t}(x - x^{(k-1)} + tA^Tz^{(k-1)})$$

$$0 \in \partial g^*(z) + \frac{1}{s}(z - z^{(k-1)} - sA(2x - x^{(k-1)})),$$

and writing the solution in terms of prox-operators gives the CP algorithm
Outline

- primal-dual optimality conditions
- monotone operators
- proximal point algorithm
- Chambolle-Pock algorithm
- Douglas-Rachford operator splitting
Operator splitting

given maximal monotone operators $F$ and $G$, solve

$$0 \in F(x) + G(x)$$

**Algorithm:** start at any $y^{(0)}$ and repeat for $k = 1, 2, \ldots$

$$x^{(k)} = (I + tF)^{-1}(y^{(k-1)})$$
$$y^{(k)} = y^{(k-1)} + (I + tG)^{-1}(2x^{(k)} - y^{(k-1)}) - x^{(k)}$$

- for $F = \partial f$ and $G = \partial g$, this is the algorithm of page 13-2
- useful when resolvents of $F$ and $G$ are inexpensive, but not resolvent of sum
- converges under weak conditions (existence of solution)
- can add relaxation to $y$-update
Primal-dual optimality conditions

$$0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix}$$

Simplest splitting

$$F(x, z) = \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix}, \quad G(x, z) = \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix}$$

- resolvent of $F$: reduces to linear equation with coefficient $I + t^2 A^T A$
- resolvent of $G$: apply prox-operators of $f$ and $g$
- complexity per iteration is similar to primal or dual DR (p. 13-11 and p. 13-19)

Other splittings: exploit additive structure in $A, f, g^*$ (see references)
References

Monotone operators and the proximal point algorithm

  The convergence result on page 14-17 is Theorem 3 of this paper.

Chambolle-Pock algorithm and extensions

  Includes a proof of convergence for $s t \| A \|_2^2 = 1$. Also includes an extension to cost functions $f(x) + g(Ax) + h(x)$, with differentiable $h$. 
Douglas-Rachford operator splitting

  
  Includes examples of other ways to split the primal-dual optimality conditions on page 14-25.