Proximal Methods for Optimization with Sparsity-inducing Norms
Group Learning Presentation

Xiaowei Zhou

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology

2011-10-21
Outline

1 Background

2 Proximal Methods
   - Overview
   - Convergence
   - Computation of Proximal Operator

3 ProxFlow algorithm for group-sparsity

4 Discussion
Problem Setting

\[
\min_{w \in \mathbb{R}^p} f(w) + \lambda \Omega(w) \quad \text{.}
\]

data fitting \quad \text{sparsity inducing}

\( f : \mathbb{R}^p \to \mathbb{R} \) is convex differentiable
\( \Omega : \mathbb{R}^p \to \mathbb{R} \) is convex but nonsmooth

\( \ell_1 \)-norm : \quad \|w\|_1 \triangleq \sum_{j=1}^{p} |w_j| \\
\( \ell_1/\ell_q \)-norm : \quad \|w\|_{\ell_1/\ell_q} \triangleq \sum_{g \in G} \eta_g \|w|_g \|_q \\
nuclear norm : \quad \|w\|_* \triangleq \sum_{j=1}^{r} \sigma_j \)
Generic Methods

Generic convex optimization methods:

- Subgradient descent

\[ w_{t+1} = w_t - \alpha(s + \lambda s'), \text{ where } s \in \partial f(w_t), \ s' \in \partial \Omega(w_t) \]

- Reformulation as standard solvable convex programs (LP, QP, SOCP, SDP), e.g. Lasso can be formulated as:

\[
\min_{w_+,w_- \in \mathbb{R}_+^p} \frac{1}{2} \|y - Xw_+ - Xw_-\|_2^2 + \lambda(1^T w_+ + 1^T w_-) \tag{2}
\]

- General-purpose toolboxes can be used (e.g. CVX).

- Blind to problem structure, tend to be slow and memory-consuming.
Outline

1 Background

2 Proximal Methods
   - Overview
   - Convergence
   - Computation of Proximal Operator

3 ProxFlow algorithm for group-sparsity

4 Discussion
Proximal methods

In proximal methods, (1) is optimized by iteratively solving the following problem:

\[
\min_{w \in \mathbb{R}^p} \quad f(w_0) + (w - w_0)^\top \nabla f(w_0) + \lambda \Omega(w) + \frac{L}{2} \|w - w_0\|^2_2.
\] (3)

- The first two terms linearly expand \(f\) at \(w_0\).
- The last term keeps the solution in a neighborhood where the current linear approximation holds.
- \(L > 0\) is an upper bound on the Lipschitz constant of \(\nabla f\).
Definition: Proximal Operator

The **proximal operator** associated with the regularization $\lambda \Omega$ is defined as the function that maps a vector $u$ in $\mathbb{R}^p$ onto the unique solution of:

$$
\min_{v \in \mathbb{R}^p} \frac{1}{2} \| u - v \|_2^2 + \lambda \Omega(v),
$$

(4)

The operator is usually denoted as $\text{Prox}_{\lambda \Omega}(u)$

Example:

- When $\Omega$ is the $\ell_1$-norm, the proximal operator is the well-known elementwise soft-thresholding operator:

$$
\forall j \in [1, \cdots, p], \quad u_j \mapsto \text{sign}(u_j)(|u_j| - \lambda)_+ = \begin{cases} 
0 & \text{if } |u_j| \leq \lambda \\
\text{sign}(u_j)(|u_j| - \lambda) & \text{otherwise}.
\end{cases}
$$
Algorithm

Since (3) can be rewritten as:

\[
\min_{w \in \mathbb{R}^p} \frac{1}{2} \|w - (w_0 - \frac{1}{L} \nabla f(w_0))\|^2 + \frac{\lambda}{L} \Omega(w),
\]  

(5)

the proximal method can be written as:

Algorithm 1: proximal method to solve (1)

repeat
  \[ w \leftarrow \text{prox}_{\frac{\lambda}{L} \Omega}(w - \frac{1}{L} \nabla f(w)) \]
until Convergence

Remark. If \( \lambda = 0 \), it turns to be Gradient Descent.

Questions:

1. How to compute the proximal operator?
2. How about the convergence of the algorithm?
Outline

1. Background

2. Proximal Methods
   - Overview
   - Convergence
   - Computation of Proximal Operator

3. ProxFlow algorithm for group-sparsity

4. Discussion
Assumption

We assume that $f$ has Lipschitz continuous gradient:

$$\|\nabla f(w) - \nabla f(v)\| \leq L\|w - v\|. \quad (6)$$

Remark: It means the change of function gradient is upper-bounded.

Lemma

The above condition is equivalent to

$$f(w) \leq f(v) + \langle \nabla f(v), w - v \rangle + \frac{L}{2}\|w - v\|^2. \quad (7)$$
Two lemma

Following two are important to the proof of convergence:

**Lemma (Sandwich)**

If $\tilde{F}(w; v)$ is the linear approximation of $F(w) = f(w) + \lambda \Omega(w)$ in $v$, w.r.t. $f(w)$, that is $\tilde{F}(w; v) = f(v) + \langle \nabla f(v), w - v \rangle + \lambda \Omega(w)$, then:

$$F(w) \leq \tilde{F}(w; v) + \frac{L}{2} \|w - v\|^2 \leq F(w) + \frac{L}{2} \|w - v\|^2. \quad (8)$$

**Lemma (3-point property)**

If $\hat{w} = \arg\min_w \frac{1}{2} \|w - w_0\|^2 + \phi(w)$, then for any $w$:

$$\phi(\hat{w}) + \frac{1}{2} \|\hat{w} - w_0\|^2 \leq \phi(w) + \frac{1}{2} \|w - w_0\|^2 - \frac{1}{2} \|w - \hat{w}\|^2. \quad (9)$$
Proof of the convergence rate I

- $F(w_t)$ is monotone non-increasing for $t = 0, \cdots, T$:

  $$F(w_{t+1}) \leq \tilde{F}(w_{t+1}; w_t) + \frac{L}{2} \|w_{t+1} - w_t\|^2$$  
  Sandwich-left

  $$\leq \tilde{F}(w_t; w_t) + \frac{L}{2} \|w_t - w_t\|^2 = F(w_t)$$  
  Definition

- Also, we have

  $$F(w_{t+1}) \leq \tilde{F}(w_{t+1}; w_t) + \frac{L}{2} \|w_{t+1} - w_t\|^2$$  
  Sandwich-left

  $$\leq \tilde{F}(w^*; w_t) + \frac{L}{2} \|w^* - w_t\|^2 - \frac{L}{2} \|w^* - w_{t+1}\|^2$$  
  3-point property

  $$\leq F(w^*) + \frac{L}{2} \|w^* - w_t\|^2 - \frac{L}{2} \|w^* - w_{t+1}\|^2$$  
  Sandwich-right
Proof of the convergence rate II

Define \( \epsilon_t = F(w_t) - F(w^*) \), so that

\[
\epsilon_{t+1} \leq \frac{L}{2} \| w^* - w_t \|^2 - \frac{L}{2} \| w^* - w_{t+1} \|^2
\]

\[
T \epsilon_t \leq \sum_{t=0}^{T-1} \epsilon_{t+1} \leq \frac{L}{2} \| w^* - w_0 \|^2 - \frac{L}{2} \| w^* - w_t \|^2 \leq \frac{L}{2} \| w^* - w_0 \|^2
\]

\[
\epsilon_t \leq \frac{L \| w^* - w_0 \|^2}{2T}
\]

(10)

- At iteration \( T \), Algorithm 1 yields a solution \( w_T \) that satisfies:

\[
F(w_T) - F(w^*) \leq \frac{L \| w^* - w_0 \|^2}{2T}.
\]

- Algorithm 1 has a convergence rate of \( O(1/T) \).
Outline

1. Background

2. Proximal Methods
   - Overview
   - Convergence
   - Computation of Proximal Operator

3. ProxFlow algorithm for group-sparsity

4. Discussion
**Definition: Dual Norm**

The dual norm $\Omega^*$ of the norm $\Omega$ is defined by:

$$\Omega^*(z) = \max_{w \in \mathbb{R}^p} z^T w, \quad s.t. \quad \Omega(w) \leq 1$$

(11)

**Examples:**
- $\ell_2$-norm is self-dual
- $\ell_1$-norm is dual to $\ell_\infty$-norm
**Dual proximal operator**: In the case where $\Omega$ is a norm, the following problem is dual to the proximal problem in (4):

$$
\max_{v \in \mathbb{R}^p} \quad -\frac{1}{2} \|v - u\|^2 + \|u\|^2, \\
\text{s.t.} \quad \Omega^*(v) \leq \lambda.
$$

(12)

Let $\text{Proj}_{\Omega^* \leq \lambda}$ be the projector on the ball of radius $\lambda$ associated with $\Omega^*$, then $\text{Proj}_{\Omega^* \leq \lambda}(u)$ is the unique solution to the problem (12) and it has the following relation with $\text{Prox}_{\lambda\Omega}$:

$$
\text{Prox}_{\lambda\Omega} = I - \text{Proj}_{\Omega^* \leq \lambda}
$$

(13)
Definition: Dual Cone
Assume $K$ is a convex cone, the dual cone $K^*$ is defined as:

$$K^* = \{ z \in \mathbb{R}^p : z^T x \geq 0, \ \forall x \in K \}$$

Primal:

$$\min_x f(x), \ \text{s.t.} \ Ax + b \preceq_K 0$$

Lagrangian:

$$L(x, \lambda) = f(x) + \lambda^T (Ax + b), \ \lambda \succeq_{K^*}$$

Dual:

$$\max_{\lambda} \inf_x L(x, \lambda), \ \text{s.t.} \ \lambda \in K^*$$
Typical examples

- When $\Omega$ is the $\ell_1$-norm, the solution is:

  \[ u \mapsto u - \text{Proj}_{\|\cdot\|_{\infty} \leq \lambda}(u), \]

  which gives the elementwise soft-thresholding:

  \[ \forall j \in [1, \cdots, p], u_j \mapsto \text{sign}(u_j)(|u_j| - \lambda)_+. \]

- When $\Omega$ is the $\ell_1/\ell_2$-norm with nonoverlapping groups, the proximal problem is *separable* in every group, and the solution is a group-thresholding operator:

  \[ \forall g \in G, u|_g \mapsto u|_g - \text{Proj}_{\|\cdot\|_2 \leq \lambda}(u|_g) = \begin{cases} 0 & \text{if } \|u|_g\|_2 \leq \lambda \\ \frac{\|u|_g\|_2 - \lambda}{\|u|_g\|_2}u|_g & \text{otherwise}, \end{cases} \]


Outline

1. Background

2. Proximal Methods
   - Overview
   - Convergence
   - Computation of Proximal Operator

3. ProxFlow algorithm for group-sparsity

4. Discussion
The following two problems are dual:

\[ P : \min_{v \in \mathbb{R}^p} \frac{1}{2} \| u - v \|_2^2 + \lambda \sum_{g \in G} \eta_g \| v|_g \| \]  

\[ D : \max_{\xi \in \mathbb{R}^{p \times |G|}} -\frac{1}{2} \| u - \sum_{g \in G} \xi^g \|_2^2 + \| u \|_2^2 \]  

s.t. \quad \forall g \in G, \| \xi^g \|_* \leq \lambda \eta_g \text{ and } \xi^g_j = 0 \text{ if } j \notin g,

- \| \cdot \| \text{ and } \| \cdot \|_* \text{ are two convex norms dual to each other.}
- \( \xi = (\xi^g)_{g \in G} \) where \( \xi^g \in \mathbb{R}^p \) is the dual variable associated with \( v|_g \).
Block Coordinate Descent

BCD algorithm to solve the proximal problem in (14)

Initialization: $\xi = 0$.
repeat
\[ \forall g \in \mathcal{G}, \quad \xi^g \leftarrow \text{Proj}_{\|\cdot\| \leq \lambda \eta_g} \left( \left[ u - \sum_{h \neq g} \xi^h \right]_g \right). \]
until Convergence
\[ v \leftarrow u - \sum_{g \in \mathcal{G}} \xi^g. \]

- Generally, the above algorithm is not guaranteed to converge in finite number of iterations.
- When $\|\cdot\|$ is $\ell_1$– or $\ell_\infty$– norm and the groups in $\mathcal{G}$ are in a hierarchical structure, only one iteration is needed to reach convergence.

Jenatton, R. and Mairal, J. and Obozinski, G. and Bach, F.
Proximal methods for hierarchical sparse coding
*Journal of Machine Learning Research* 12 (2011) 2297-2334
How to solve the group-sparsity problem efficiently in general cases?

- If $\ell_1/\ell_\infty$-norm is used, the dual problem (15) becomes:

$$
\min_{\xi \in \mathbb{R}^{p \times |G|}} \frac{1}{2} \| u - \sum_{g \in G} \xi^g \|^2_2 \\
\text{s.t. } \forall g \in G, \| \xi^g \|_1 \leq \lambda \eta_g \text{ and } \xi^g_j = 0 \text{ if } j \notin g,
$$

The above problem can be formulated on a graph as a Quadratic Min-Cost Flow problem.

Mairal, J. and Jenatton, R. and Obozinski, G. and Bach, F. Convex and Network Flow Optimization for Structured Sparsity

*Journal of Machine Learning Research* 12 (2011) 2681-2720
(a) $\mathcal{G} = \{g = \{1, 2, 3\}\}$. 

(b) $\mathcal{G} = \{g = \{1, 2\}, h = \{2, 3\}\}$. 

(c) $\mathcal{G} = \{g = \{1, 2, 3\}, h = \{2, 3\}\}$. 

(d) $\mathcal{G} = \{g = \{1\} \cup h, h = \{2, 3\}\}$. 

\[ \xi_1^g + \xi_2^g + \xi_3^g \leq \lambda \eta_g \]

\[ \xi_1^g + \xi_2^g \leq \lambda \eta_g \]

\[ \xi_2^g + \xi_3^g \leq \lambda \eta_h \]

\[ \xi_1^g + \xi_2^g + \xi_3^g \leq \lambda \eta_g \]

\[ \xi_2^g + \xi_3^g \leq \lambda \eta_h \]

\[ \xi_1^h + \xi_2^h \leq \lambda \eta_g \]

\[ \xi_2^h + \xi_3^h \leq \lambda \eta_h \]

\[ \xi_1^h + \xi_2^h + \xi_3^h \leq \lambda \eta_g \]

\[ \xi_2^h + \xi_3^h \leq \lambda \eta_h \]
Speed comparison

Figure: Comparison: Proximal + Network Flow (ProxFlow), Quadratic Programming (QP), Conic Programming (CP) and Sub-gradient Descent (SG)
1. Background

2. Proximal Methods
   - Overview
   - Convergence
   - Computation of Proximal Operator

3. ProxFlow algorithm for group-sparsity

4. Discussion
Toolbox

SPAMS – SPArse Modeling Software

- Developer: CVML Lab @ INRI.
  http://www.di.ens.fr/willow/SPAMS/
- Addressing various machine learning and signal processing problems:
  - Dictionary learning and matrix factorization (NMF, sparse PCA, ...).
  - Sparse decomposition problems with LARS, coordinate descent, OMP, SOMP, proximal methods.
  - Structured sparse decomposition problems.
Convex formulation of DECOLOR?

Replace the MRFs penalty in DECOLOR with a group sparsity-inducing norm:

$$\min_{L, S} \frac{1}{2} \| M - L - S \|_F^2 + \mu \| L \|_* + \lambda \Omega(S),$$

where \( \Omega(S) := \sum_{g \in G} \eta_g \| s|_g \|_\infty \).

Stable PCP with Group Sparsity by BCD

**Initialize:** \( S_0 = 0 \).

repeat

- compute \( L_{k+1} = D_\mu(M - S_k) \);
- compute \( S_{k+1} = \text{Prox}_{\lambda \Omega}(M - L_{k+1}) \);

until convergence

**Output:** \( L, S \).
Key References


Luca Baldassarre
Proximal Methods (*Lecture Notes*)
www.cs.ucl.ac.uk/staff/L.Baldassarre/lectures