Smoothing Proximal Gradient Method for General Structured Sparse Regression

Xi Chen, Qihang Lin, Seyoung Kim, Jaime G. Carbonell, Eric P. Xing (Annals of Applied Statistics, 2012)

Zoltán Szabó

Gatsby Unit, Tea Talk

October 25, 2013

Outline

- Motivation: (structured) sparse coding.
- Proximal operators, FISTA.
- Solution: dual norm + smooth approximation.

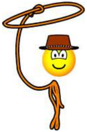
Motivation: least squares, sparse coding

- Given: $\mathbf{x} \in \mathbb{R}^{d_x}$, $\mathbf{D} \in \mathbb{R}^{d_x \times d_\alpha}$.
- Least squares problem:

$$J(\boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \to \min_{\boldsymbol{\alpha} \in \mathbb{R}^{d_\alpha}}. \tag{1}$$

- Sparse coding (JPEG; convex relaxation, Lasso, $w > 0$):

$$J(\boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + w \|\boldsymbol{\alpha}\|_1 \to \min_{\boldsymbol{\alpha} \in \mathbb{R}^{d_\alpha}}. \tag{2}$$

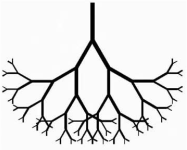


Motivation: structured sparse coding

- Group Lasso ($\mathcal{G}$ partition = non-overlapping, blocks):

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + w\sum_{G\in\mathcal{G}}\|\boldsymbol{\alpha}_G\|_2 \to \min_{\boldsymbol{\alpha}\in\mathbb{R}^{d_\alpha}}. \qquad (3)$$

- Overlapping $\mathcal{G}$:
    - hierarchy, grid, total variation, graphs.
    - many successful application: gene analysis, face expression recognition, ...



Non-overlapping group Lasso

- FISTA objective:

$$J(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha}) + g(\boldsymbol{\alpha}) \to \min_{\boldsymbol{\alpha} \in \mathbb{R}^{d_\alpha}}. \tag{4}$$

- Assumptions:
  - f, g: convex,
  - f is 'smooth' (Lipschitz continuous gradient, L).
- Fast convergence:

$$J(\boldsymbol{\alpha}_t) - J(\boldsymbol{\alpha}^*) = O\left(\frac{1}{t^2}\right). \tag{5}$$

FISTA

- Ingredients:
  - Gradient of the smooth term: ∇f.
  - Lipschitz constant of ∇f: L.
  - Proximal operator of the non-smooth term ($p > 0$):

$$prox_{pg}(\mathbf{v}) = \arg\min_{\mathbf{y}} \left[g(\mathbf{y}) + \frac{1}{2p} \|\mathbf{y} - \mathbf{v}\|_2^2 \right]. \tag{6}$$

- Example: $f(\boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2$, $g(\boldsymbol{\alpha}) = w \sum_{G \in \mathcal{G}} \|\boldsymbol{\alpha}_G\|_2$,

$$\nabla f(\boldsymbol{\alpha}) = \mathbf{D}^T(\mathbf{D}\boldsymbol{\alpha} - \mathbf{x}), \qquad L = \lambda_{\max}\left(\mathbf{D}^T\mathbf{D}\right), \tag{7}$$

$$prox_g\text{: analytical (for } \textit{partition } \mathcal{G}\text{).} \tag{8}$$

Goal

- Objective ($\lambda > 0$; $w_G > 0$, $\forall G \in \mathcal{G}$):

$$J(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha}) + \Omega(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|_1 \to \min_{\boldsymbol{\alpha} \in \mathbb{R}^{d_\alpha}}, \tag{9}$$

$$\Omega(\boldsymbol{\alpha}) = \sum_{G \in \mathcal{G}} w_G \|\boldsymbol{\alpha}_G\|_2 . \tag{10}$$

- Assumption:
  - f: convex (FISTA assumptions).
  - $\mathcal{G}$: *non-overlapping* $\Rightarrow$ no analytical formula for $prox_{pg}$.

Solution

- The ℓ_2-norm is self-dual:

$$\|\mathbf{a}\|_2 = \max_{\mathbf{b}:\|\mathbf{b}\|_2 \leq 1} \mathbf{b}^T \mathbf{a}. \tag{11}$$

Solution

- The ℓ_2-norm is self-dual:

$$\|\mathbf{a}\|_2 = \max_{\mathbf{b}:\|\mathbf{b}\|_2 \le 1} \mathbf{b}^T \mathbf{a}. \tag{11}$$

- We rewrite Ω ($\boldsymbol{\alpha}_G \mapsto \boldsymbol{\beta}_G \in \mathbb{R}^{|G|}$: auxiliary variable):

$$\boldsymbol{\beta} = \left[(\boldsymbol{\beta}_G)_{G \in \mathcal{G}}\right] \in \mathbb{R}^{\sum_{G \in \mathcal{G}} |G|}, \tag{12}$$

$$\Omega(\boldsymbol{\alpha}) = \sum_{G \in \mathcal{G}} w_G \|\boldsymbol{\alpha}_G\|_2 = \sum_{G \in \mathcal{G}} w_G \max_{\boldsymbol{\beta}_G:\|\boldsymbol{\beta}_G\|_2 \le 1} \boldsymbol{\beta}_G^T \boldsymbol{\alpha}_G \tag{13}$$

$$= \max_{\boldsymbol{\beta} \in \mathcal{Q}} \sum_{G \in \mathcal{G}} w_G \boldsymbol{\beta}_G^T \boldsymbol{\alpha}_G =: \max_{\boldsymbol{\beta} \in \mathcal{Q}} \boldsymbol{\beta}^T \mathbf{C} \boldsymbol{\alpha}, \tag{14}$$

$$\mathcal{Q} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}_G\|_2 \le 1, \forall G \in \mathcal{G}\} \text{ (product of unit balls).}$$

Solution - continued

- Smooth approximation to $\Omega(\boldsymbol{\alpha})$ ($\mu \geq 0$):

$$\Omega(\boldsymbol{\alpha}) = \max_{\boldsymbol{\beta} \in \mathcal{Q}} \boldsymbol{\beta}^T \mathbf{C} \boldsymbol{\alpha} \approx \max_{\boldsymbol{\beta} \in \mathcal{Q}} \left(\boldsymbol{\beta}^T \mathbf{C} \boldsymbol{\alpha} - \mu s(\boldsymbol{\beta}) \right) =: \Omega_\mu(\boldsymbol{\alpha}),$$

$$s(\boldsymbol{\beta}) = \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 \geq 0. \tag{15}$$

- Maximum gap is μM:

$$M = \max_{\boldsymbol{\beta} \in \mathcal{Q}} s(\boldsymbol{\beta}) = \frac{|\mathcal{G}|}{2}, \tag{16}$$

$$\Omega(\boldsymbol{\alpha}) - \mu M \leq \Omega_\mu(\boldsymbol{\alpha}) \leq \Omega(\boldsymbol{\alpha}). \tag{17}$$

Solution: FISTA on the smooth approximation

- Original objective ($\lambda > 0$):

$$J(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha}) + \Omega(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|_1 \to \min_{\boldsymbol{\alpha} \in \mathbb{R}^{d_\alpha}}. \tag{18}$$

- Smooth approximation ($\mu > 0$, $\lambda > 0$):

$$J_\mu(\boldsymbol{\alpha}) = \underbrace{f(\boldsymbol{\alpha}) + \Omega_\mu(\boldsymbol{\alpha})}_{\text{FISTA: } f} + \underbrace{\lambda \|\boldsymbol{\alpha}\|_1}_{g} \to \min_{\boldsymbol{\alpha} \in \mathbb{R}^{d_\alpha}}. \tag{19}$$

Result (=FISTA can be applied)

- $\Omega_\mu(\boldsymbol{\alpha})$: convex with Lipschitz continuous gradient

$$\nabla\Omega_\mu(\boldsymbol{\alpha}) = \mathbf{C}^T\boldsymbol{\beta}^*, \tag{20}$$

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}\in\mathcal{Q}}\left(\boldsymbol{\beta}^T\mathbf{C}\boldsymbol{\alpha} - \mu s(\boldsymbol{\beta})\right) \tag{21}$$

$$= \left[\left(\Pi_2\left(\frac{w_G\boldsymbol{\alpha}_G}{\mu}\right)\right)_{G\in\mathcal{G}}\right]. \tag{22}$$

- Lipschitz constant: $L_\mu = \frac{1}{\mu}\|\mathbf{C}\|_2^2$.

Proof (intuition)

- Convexity, smoothness of Ω_μ:

$$\Omega_\mu(\boldsymbol{\alpha}) = \max_{\boldsymbol{\beta}\in\mathcal{Q}} \left(\boldsymbol{\beta}^T \mathbf{C}\boldsymbol{\alpha} - \mu s(\boldsymbol{\beta})\right) = \mu \max_{\boldsymbol{\beta}\in\mathcal{Q}} \left(\boldsymbol{\beta}^T \frac{\mathbf{C}\boldsymbol{\alpha}}{\mu} - s(\boldsymbol{\beta})\right)$$
$$= \mu d^* \left(\frac{\mathbf{C}\boldsymbol{\alpha}}{\mu}\right). \quad (23)$$

- Gradient $\nabla\Omega_\mu$: Danskin's theorem with

$$h(\boldsymbol{\alpha}) = \max_{\boldsymbol{\beta}\in K:compact} \varphi(\boldsymbol{\beta}, \boldsymbol{\alpha}), \quad (24)$$
$$\nabla h(\boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}} \varphi(\boldsymbol{\beta}^*, \boldsymbol{\alpha}). \quad (25)$$

- Lipschitz constant L_μ: Nesterov '05.

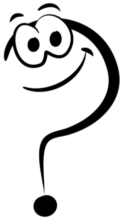
Convergence rate: $O\left(\frac{1}{\epsilon}\right)$

- Given: ϵ (precision).
- We want
$$J(\boldsymbol{\alpha}_t) - J(\boldsymbol{\alpha}^*) \leq \epsilon. \tag{26}$$
- Set $\mu = \frac{\epsilon}{2M}$, where $M = \frac{|\mathcal{G}|}{2}$.
- Sufficient number of iterations:
$$O\left(\frac{1}{\epsilon}\right) = \sqrt{\frac{4\left\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}_0\right\|_2^2}{\epsilon}\left[\lambda_{\max}\left(\mathbf{D}^T\mathbf{D}\right) + \frac{2M\left\|\mathbf{C}\right\|_2^2}{\epsilon}\right]}.$$
- Note (subgradient descent is much slower): $O\left(\frac{1}{\epsilon^2}\right)$.

Summary

- Task: non-overlapping group Lasso.
- Difficulty: non-overlapping $\Rightarrow$ non-separability.
- Proposed solution:
  - $\|\cdot\|_2 = \|\cdot\|_2^*$. Smooth approximation.
  - $|\mathcal{G}|$ independent subproblems, analytical expressions to FISTA.
  - convergence rate: $O\left(\frac{1}{\epsilon}\right)$.

Thank you for the attention!



Analytical solution for β^*

$$\beta^* = \arg\max_{\beta \in \mathcal{Q}} \left(\beta^T \mathbf{C} \boldsymbol{\alpha} - \frac{\mu}{2} \|\beta\|_2^2 \right) \tag{27}$$

$$= \arg\max_{\beta \in \mathcal{Q}} \sum_{G \in \mathcal{G}} \left(w_G \beta_G^T \boldsymbol{\alpha}_G - \frac{\mu}{2} \|\beta_G\|_2^2 \right) \tag{28}$$

$$= \arg\min_{\beta \in \mathcal{Q}} \sum_{G \in \mathcal{G}} \left\| \beta_G - \frac{w_G \boldsymbol{\alpha}_G}{\mu} \right\|_2^2. \tag{29}$$

Thus

$$(\beta^*)_G = \Pi_2 \left(\frac{w_G \boldsymbol{\alpha}_G}{\mu} \right). \tag{30}$$

Combination of Lipschitz constants

- Let L_f (L_g) be a Lipschitz constant of ∇f (∇g).
- Then $L_{f+g} \leq L_f + L_g$, since

$$\|(\nabla f + \nabla g)(\mathbf{x}) - (\nabla f + \nabla g)(\mathbf{y})\|_2 \tag{31}$$
$$\leq \|[\nabla f(\mathbf{x}) + \nabla g(\mathbf{x})] - [\nabla f(\mathbf{y}) + \nabla g(\mathbf{y})]\|_2 \tag{32}$$
$$\leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 + \|\nabla g(\mathbf{y}) - \nabla g(\mathbf{y})\|_2 \tag{33}$$
$$= L_f \|\mathbf{x} - \mathbf{y}\|_2 + L_g \|\mathbf{x} - \mathbf{y}\|_2 \tag{34}$$
$$\leq (L_f + L_g) \|\mathbf{x} - \mathbf{y}\|_2 . \tag{35}$$

Rate of convergence for SPG

$$J(\boldsymbol{\alpha}_t) - J(\boldsymbol{\alpha}^*) \tag{36}$$
$$= [J(\boldsymbol{\alpha}_t) - J_\mu(\boldsymbol{\alpha}_t)] + [J_\mu(\boldsymbol{\alpha}_t) - J_\mu(\boldsymbol{\alpha}^*)] + [J_\mu(\boldsymbol{\alpha}^*) - J(\boldsymbol{\alpha}^*)]$$
$$\leq \mu M + \frac{2L_\mu \left\|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\right\|_2^2}{t^2} + 0 \tag{37}$$
$$\leq \mu M + \frac{2\left\|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\right\|_2^2}{t^2}\left(\lambda_{\max}\left(\mathbf{D}^T\mathbf{D}\right) + \frac{\|\mathbf{C}\|_2^2}{\mu}\right). \tag{38}$$

Plug-in $\mu = \frac{\epsilon}{2M}$, and solve for t:

$$J(\boldsymbol{\alpha}_t) - J(\boldsymbol{\alpha}^*) \leq \frac{\epsilon}{2} + \frac{2\left\|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\right\|_2^2}{t^2}\left(\lambda_{\max}\left(\mathbf{D}^T\mathbf{D}\right) + \frac{2M\|\mathbf{C}\|_2^2}{\epsilon}\right) \leq \epsilon.$$

Proximal operator

- $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$: closed proper convex function, i.e.,
$$epi(f) = \{(\mathbf{y}, t) \in \mathbb{R}^d \times \mathbb{R} : f(\mathbf{y}) \leq t\} \tag{39}$$
is nonempty closed convex.
- Proximal operator of f:
$$prox_f(\mathbf{v}) = \arg\min_{\mathbf{y}} \left[f(\mathbf{y}) + \frac{1}{2} \|\mathbf{y} - \mathbf{v}\|_2^2 \right]. \tag{40}$$
- Strictly convex r.h.s. of (40) $\Rightarrow$ $prox_f$: exists, unique.

Proximal operator = generalization of projection

- $\mathfrak{C}$: closed convex set.
- $f = I_C$: indicator function of $\mathfrak{C}$

$$I_C(\mathbf{y}) = \begin{cases} 0 & \mathbf{y} \in \mathfrak{C}, \\ \infty & \mathbf{y} \notin \mathfrak{C}. \end{cases} \quad (41)$$

- Then, $prox_f$ = Euclidean projection onto $\mathfrak{C}$:

$$prox_{I_{\mathfrak{C}}}(\mathbf{v}) = \Pi_{\mathfrak{C}}(\mathbf{v}) = \arg\min_{\mathbf{y}} \|\mathbf{v} - \mathbf{y}\|_2 \,. \quad (42)$$

Conjugate function

- $f: \mathbb{R}^d \to \mathbb{R}$, not necessarily convex.
- Conjugate of f:

$$f^*(\mathbf{v}) = \sup_{\mathbf{y}} \left[\mathbf{v}^T\mathbf{y} - f(\mathbf{y})\right]. \tag{43}$$

- Notes:
  - f^*: convex $\Leftarrow$ pointwise sup of convex functions.
  - if f is convex, closed: $(f^*)^* = f$.
  - if f is differentiable: f^* = Legendre transform of f.

Conjugate function: properties

- If f = indicator function of a unit ball, i.e.,

$$f = I_C, \qquad C = B_{\|\cdot\|} = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\| \leq 1\}, \tag{44}$$

then f^* is the dual norm

$$f^*(\mathbf{v}) = \|\mathbf{v}\|^* = \max_{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\| \leq 1} \mathbf{v}^T \mathbf{y}. \tag{45}$$

- Dual norm of $\|\cdot\|_p$ ($p \geq 1$) is $\|\cdot\|_{p'}$ with $\frac{1}{p} + \frac{1}{p'} = 1$.
- Similarly ($\mathcal{G}$: partition):

$$\|\mathbf{u}\| = \sum_{G \in \mathcal{G}} \|\mathbf{u}_G\|_p, \qquad \|\mathbf{u}\|^* = \max_{G \in \mathcal{G}} \|\mathbf{u}_G\|_{p'}. \tag{46}$$