Gradient Descent with Proximal Average for Nonconvex and Composite Regularization

Leon Wenliang Zhong **James T. Kwok**
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong
{wzhong, jamesk}@cse.ust.hk

Abstract

Sparse modeling has been highly successful in many real-world applications. While a lot of interests have been on convex regularization, recent studies show that nonconvex regularizers can outperform their convex counterparts in many situations. However, the resulting nonconvex optimization problems are often challenging, especially for composite regularizers such as the nonconvex overlapping group lasso. In this paper, by using a recent mathematical tool known as the proximal average, we propose a novel proximal gradient descent method for optimization with a wide class of nonconvex and composite regularizers. Instead of directly solving the proximal step associated with a composite regularizer, we average the solutions from the proximal problems of the constituent regularizers. This simple strategy has guaranteed convergence and low per-iteration complexity. Experimental results on a number of synthetic and real-world data sets demonstrate the effectiveness and efficiency of the proposed optimization algorithm, and also the improved classification performance resulting from the nonconvex regularizers.

Introduction

Risk minimization is a fundamental tool in machine learning. It admits a tradeoff between the empirical loss and regularization as:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \equiv \ell(\mathbf{x}) + r(\mathbf{x}), \tag{1}$$

where ℓ is the loss, and r is a regularizer on parameter $\mathbf{x}$. In particular, sparse modeling, which uses a sparsity-inducing regularizer for feature selection, has achieved great success in many real-world applications. A well-known sparsity-inducing regularizer is the ℓ_1-regularizer. As a surrogate of the ℓ_0-norm, it induces a sparse solution simultaneously with learning (Tibshirani 1996). When the features have some intrinsic structures, more sophisticated structured-sparsity-inducing regularizers (such as the group lasso regularizer (Yuan and Lin 2006)) can be used. More examples can be found in (Bach et al. 2011; Combettes and Pesquet 2011) and reference therein. Existing sparsity-inducing regularizers are often convex. Together with a convex loss, this leads to a convex optimization problem with globally optimal solution.

Despite such extensive popularity, convexity does not necessarily imply good prediction performance or feature selection. Indeed, it has been shown that lasso may lead to over-penalization and suboptimal feature selection (Zhang 2010b; Candes, Wakin, and Boyd 2008). To overcome this problem, several nonconvex variants have been recently proposed, such as the capped-ℓ_1 (Zhang 2010b), log-sum penalty (LSP) (Candes, Wakin, and Boyd 2008), smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) and minmax concave penalty (MCP) (Zhang 2010a). For more sophisticated scenarios, recent research efforts demonstrate that nonconvex regularizers, such as the nonconvex group lasso (Xiang, Shen, and Ye 2013; Chartrand and Wohlberg 2013), matrix MCP norm (Wang, Liu, and Zhang 2013), and grouping pursuit (Shen and Huang 2010), can outperform their convex counterparts.

However, these nonconvex models often yield challenging optimization problems. As most of them can be rewritten as $f_1 - f_2$, a difference of two convex functions f_1 and f_2 (Gong et al. 2013), a popular optimization solver is the multi-stage convex programming, which recursively approximates f_2 while leaving f_1 intact (Zhang 2010b; Zhang et al. 2013; Xiang, Shen, and Ye 2013). However, it involves nonlinear optimization in each iteration and thus expensive in general. The sequential convex program (SCP) (Lu 2012) further approximates the smooth part of f_1 so that the update can be more efficient for simple regularizers like the capped-ℓ_1. However, it is often trapped in poor local optimum (Gong et al. 2013). Recently, a general iterative shrinkage and thresholding (GIST) framework is proposed (Gong et al. 2013), which shows promising performance in a class of nonconvex penalties. However, for composite regularizers such as the nonconvex variants of overlapping group lasso (Zhao, Rocha, and Yu 2009), generalized lasso (Tibshirani, Hoefling, and Tibshirani 2011) and a combination of ℓ_1- and trace norms (Richard, Savalle, and Vayatis 2012), both SCP and GIST are inefficient as the underlying proximal steps for these composite regularizers are very difficult.

In this paper, we propose a simple algorithm called Gradient Descent with Proximal Average of Nonconvex functions (GD-PAN) and its line-search-based variant GD-PAN-LS, which are suitable for a wide class of nonconvex and composite regularization problems. We first extend a recent optimization tool called "proximal average" (Yu 2013;

Bauschke et al. 2008) to nonconvex functions. Instead of directly solving the proximal step associated with a nonconvex composite regularizer, we average the solutions from the proximal problems of individual regularizers. This simple strategy has convergence guarantee as existing approaches like multi-stage convex programming and SCP, but its per-iteration complexity is much lower.

Problem Formulation

In this paper, we consider the optimization problem in (1). Moreover, the following assumptions are made on ℓ and r.

(A1) ℓ is differentiable but possibly nonconvex with L_ℓ-Lipschitz continuous gradient, i.e., $\|\nabla\ell(\mathbf{x}_1) - \nabla\ell(\mathbf{x}_2)\| \leq L_\ell\|\mathbf{x}_1 - \mathbf{x}_2\|, \forall \mathbf{x}_1, \mathbf{x}_2$.

(A2) r is nonconvex, nonsmooth, and can be written as a convex combination of K functions $\{r_1, r_2, \ldots, r_K\}$:

$$r(\mathbf{x}) = \sum_{k=1}^{K} w_k r_k(\mathbf{x}), \tag{2}$$

where $\{w_k \geq 0\}$ are coefficients satisfying $\sum_{k=1}^K w_k = 1$, and

$$r_k(\mathbf{x}) = \Omega_1(\omega_k(\mathbf{x})) - \Omega_2(\omega_k(\mathbf{x})) \tag{3}$$

for some convex functions ω_k, Ω_1 and Ω_2. Moreover, each r_k (for $k = 1, \ldots, K$) is assumed to be L_k-Lipschitz continuous and "*simple*", i.e., the associated proximal step

$$\min_{\mathbf{x}} \frac{1}{2\eta}\|\mathbf{x} - \mathbf{u}\|^2 + r_k(\mathbf{x}), \tag{4}$$

where $\mathbf{u}$ is a constant vector in $\mathbb{R}^d$ and $\eta > 0$, can be solved efficiently and exactly.

(A3) $\ell(\mathbf{x}) > -\infty, r_k(\mathbf{x}) \geq -\infty, \forall \mathbf{x}$, and $f(\mathbf{x}) = \infty$ iff $\|\mathbf{x}\| = \infty$.

Assumption **A1** has been popularly used in the literature (Nesterov 2007; Beck and Teboulle 2009), and is satisfied by many loss functions. Examples include (i) the square loss $\ell(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{S}\mathbf{x}\|^2$, where $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_n]^T$ is the data matrix and $\mathbf{y} = [y_1, \ldots, y_n]$ is the corresponding label vector; (ii) logistic loss $\ell(\mathbf{x}) = \sum_{i=1}^n \log(1 + \exp(-y_i\mathbf{s}_i^T\mathbf{x}))$; and (iii) smooth zero-one loss $\ell(\mathbf{x}) = \sum_{i=1}^n \frac{1}{1+\exp(cy_i\mathbf{s}_i^T\mathbf{x})}$, where $c > 0$ is a constant (Shalev-Shwartz, Shamir, and Sridharan 2010).

For assumption **A2**, equation (3) is a core technique in the concave-convex procedure (Yuille and Rangarajan 2003) that decomposes a nonconvex function (in this case, $r_k(\mathbf{x})$) as a difference of convex functions (i.e., $\Omega_1(\omega_k(\mathbf{x}))$ and $\Omega_2(\omega_k(\mathbf{x}))$). Some concrete examples will be shown in the next section.

Assumption **A3** naturally holds for regularized risk minimization problems as both the parameter $\mathbf{x}$ and samples are often bounded.

Example Regularizers

The following introduces some examples of r in (2), which are nonconvex extensions of popular (convex) structured-sparsity-inducing regularizers. We will also show that the proximal step in (4) can be efficiently computed.

- Capped overlapping group-lasso regularizer: This is a hybrid of the (nonconvex) capped-ℓ_1 regularizer $\sum_{i=1}^d \min\{|x_i|, \theta\}$ (where $\theta > 0$ is a constant) (Zhang 2010b; Gong, Ye, and Zhang 2012) and the (convex) overlapping group-lasso regularizer $\sum_{k=1}^K w_k\|\mathbf{x}_{g_k}\|$ (where K is the number of feature groups, w_k is the weight on group k, and $\mathbf{x}_{g_k}$ is the subvector in $\mathbf{x}$ for the subset of indices $g_k \subseteq \{1, \ldots, d\}$) (Zhao, Rocha, and Yu 2009). Define

$$\omega_k(\mathbf{x}) = \|\mathbf{x}_{g_k}\|, \Omega_1(\cdot) = |\cdot|, \Omega_2(\cdot) = (|\cdot| - \theta)_+, \tag{5}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$. Plugging into (2) and (3), it can be shown that

$$r_k(\mathbf{x}) = \min\{\|\mathbf{x}_{g_k}\|, \theta\}.$$

To solve the proximal step (4), we first assume that $\|\mathbf{x}_{g_k}\|$ is known. From (5), $r_k(\mathbf{x})$ is then also fixed, and the optimal solution $\mathbf{x}^*$ of (4) can be obtained as[1]

$$x_j^* = \begin{cases} u_j & j \notin g_k \\ \frac{u_j\|\mathbf{x}_{g_k}^*\|}{\|\mathbf{u}_{g_k}\|} & j \in g_k. \end{cases} \tag{6}$$

In other words, $\mathbf{x}_{g_k}^*$ and $\mathbf{u}_{g_k}$ are in the same direction. From (6), we have $\|\mathbf{x}^* - \mathbf{u}\|^2 = (\|\mathbf{x}_{g_k}^*\| - \|\mathbf{u}_{g_k}\|)^2$. Let $y \equiv \|\mathbf{x}_{g_k}^*\|$, (4) leads to the following univariate problem

$$\min_y \frac{1}{2\eta}(y - \|\mathbf{u}_{g_k}\|)^2 + \min\{|y|, \theta\}. \tag{7}$$

Depending on the relative magnitudes of $|y|$ and θ, this can be split into two subproblems:

$$\begin{aligned} z_1 &= \arg\min_{z:z\geq\theta} h_1(z) \equiv \frac{1}{2\eta}(z - \|\mathbf{u}_{g_k}\|)^2 + \theta \\ &= \max\{\theta, \|\mathbf{u}_{g_k}\|\}, \end{aligned}$$

and

$$\begin{aligned} z_2 &= \arg\min_{z:0\leq z\leq\theta} h_2(z) \equiv \frac{1}{2\eta}(z - \|\mathbf{u}_{g_k}\|)^2 + z \\ &= \min\{\theta, \max\{0, \|\mathbf{u}_{g_k}\| - \eta\}\}. \end{aligned}$$

From these, we obtain

$$\|\mathbf{x}_{g_k}^*\| = \begin{cases} z_1 & h_1(z_1) \leq h_2(z_2) \\ z_2 & \text{otherwise} \end{cases},$$

and subsequently $\mathbf{x}^*$ from (6). Moreover, it is easy to check that r_k is 1-Lipschitz continuous.

With different combinations of Ω_1 and Ω_2, one can obtain other nonconvex regularizers such as the LSP, SCAD and MCP (Gong et al. 2013). For example, for the log-sum penalty (LSP) $\sum_{i=1}^d \log\left(1 + \frac{|x_i|}{\theta}\right)$ that will be used in the experiments, $\Omega_1(\cdot) = |\cdot|$ and $\Omega_2(\cdot) = |\cdot| - \log\left(1 + \frac{|\cdot|}{\theta}\right)$. The corresponding proximal problem can also be similarly solved as for the capped-ℓ_1 regularizer.

[1] Clearly, when $\|\mathbf{u}_{g_k}\| = 0$, we have $\mathbf{x}^* = \mathbf{u}$.

- Capped graph-guided fused lasso: This is a hybrid of the (non-convex) capped-ℓ_1 regularizer and the (convex) graph-guided fused lasso $\sum_{k=1}^{K} w_k|x_{k_1} - x_{k_2}|$ (where K is the number of edges in a graph of features, and $k = \{k_1, k_2\}$ is a pair of vertices connected by an edge with weight w_k) (Tibshirani and Taylor 2011; Ouyang et al. 2013). Define

$$\omega_k(\mathbf{x}) = |x_{k_1} - x_{k_2}|, \Omega_1(\cdot) = |\cdot|, \Omega_2(\cdot) = (|\cdot| - \theta)_+, \quad (8)$$

for some $\theta > 0$. It can be shown that

$$r_k(\mathbf{x}) = \min\{|x_{k_1} - x_{k_2}|, \theta\},$$

and is 1-Lipschitz continuous. This regularizer thus encourages coefficients of highly related features (which are connected by a graph edge) to stay close. Using a complete feature graph, Shen and Huang (2010) demonstrated that this regularizer outperforms its convex counterpart. In general, a sparse graph is preferred and can be induced by sparse inverse covariance matrix (Banerjee, El Ghaoui, and d'Aspremont 2008).

To solve the proximal step, we first assume that $|x_{k_1} - x_{k_2}|$ is known and equals y. From (3) and (8), $r_k(\mathbf{x})$ is then also fixed. Without loss of generality, assume that $u_{k_1} \geq u_{k_2}$. The optimal solution $\mathbf{x}^*$ of (4) can be obtained as

$$x_j^* = \begin{cases} u_j & j \notin \{k_1, k_2\} \\ u_{k_1} - \frac{1}{2}(|u_{k_1} - u_{k_2}| - y) & j = k_1 \\ u_{k_2} + \frac{1}{2}(|u_{k_1} - u_{k_2}| - y) & j = k_2 \end{cases}. \quad (9)$$

Problem (4) then leads to the problem

$$\min_y \frac{1}{4\eta}(|u_{k_1} - u_{k_2}| - y)^2 + \min\{|y|, \theta\},$$

which can be solved in a similar manner as (7). Finally, one can recover $\mathbf{x}$ from (9).

Proposed Algorithm

Given a stepsize $\eta > 0$ and a function h, let $M_h^\eta(\mathbf{u}) \equiv \min_{\mathbf{x}} \frac{1}{2\eta}\|\mathbf{x} - \mathbf{u}\|^2 + h(\mathbf{x})$ be the associated proximal problem at $\mathbf{u}$, and $P_h^\eta(\mathbf{u}) \equiv \arg M_h^\eta(\mathbf{u})$ the corresponding solution. The proximal gradient descent algorithm (Beck and Teboulle 2009; Gong et al. 2013) solves problem (1) by iteratively updating the parameter estimate as:

$$\begin{aligned} \mathbf{u}^{(t)} &\leftarrow \mathbf{x}^{(t)} - \eta\nabla\ell(\mathbf{x}^{(t)}), \\ \mathbf{x}^{(t+1)} &\leftarrow P_r^\eta(\mathbf{u}^{(t)}), \end{aligned} \quad (10)$$

where the superscript $^{(t)}$ denotes the iterate at iteration t, and $\nabla\ell(\mathbf{x}^{(t)})$ is the gradient of ℓ at $\mathbf{x}^{(t)}$. The proximal step (10) has been extensively studied for simple (convex and nonconvex) regularizers (Combettes and Pesquet 2011; Gong et al. 2013). However, when r is a combination of regularizers as in (2), efficient solutions are often not available. Very recently, for convex r_k's (i.e., $\Omega_2 = 0$ in (3)), Yu (2013) utilized the proximal average (Bauschke et al. 2008), and replaced (10) with

$$\mathbf{x}^{(t+1)} \leftarrow \sum_{k=1}^{K} w_k P_{r_k}^\eta(\mathbf{u}^{(t)}). \quad (11)$$

Obviously, this can be much easier than (10) when all r_k's are simple. Interestingly, it is shown that this trick implicitly uses another convex function to approximate r.

In this section, we propose a novel procedure called Gradient Descent with Proximal Average of Non-convex functions (GD-PAN) for the general case where r_k's are nonconvex. Inspired by (Yu 2013; Gong et al. 2013), it adopts the same update rule (11), and the (constant) stepsize in (11) is chosen as $\eta = \frac{1}{L_\ell + \mathcal{L}}$ for some $\mathcal{L} > 0$. Our analysis is related to that in (Yu 2013), though his proof relies heavily on tools in convex analysis (in particular, the Moreau envelope) and cannot be applied to our nonconvex setting.

Similar to the handling of convex r_k's in (Yu 2013), the following Proposition shows that GD-PAN also implicitly optimizes a surrogate of problem (1).

Proposition 1 *There exists a function $\hat{r}$ such that*

$$M_{\hat{r}}^\eta(\mathbf{u}) = \sum_{k=1}^{K} w_k M_{r_k}^\eta(\mathbf{u}), \text{ and } P_{\hat{r}}^\eta(\mathbf{u}) = \sum_{k=1}^{K} w_k P_{r_k}^\eta(\mathbf{u}).$$

Specifically, for a given $\mathbf{x}$, $\hat{r}(\mathbf{x})$ is the optimal value of the following problem

$$\begin{aligned} \min_{\{\mathbf{x}_k\}_{k=1}^K} \quad & \sum_{k=1}^{K} w_k \left[\frac{1}{2\eta}\|\mathbf{x}_k\|^2 + r_k(\mathbf{x}_k)\right] - \frac{\|\mathbf{x}\|^2}{2\eta} \\ \textit{s.t.} \quad & \sum_{k=1}^{K} w_k \mathbf{x}_k = \mathbf{x}. \end{aligned} \quad (12)$$

Using this Proposition, (11) becomes: $\mathbf{x}^{(t+1)} \leftarrow P_{\hat{r}}^\eta(\mathbf{u}^{(t)})$, and the surrogate of problem (1) is

$$\min_{\mathbf{x}} \ \hat{f}(\mathbf{x}) \equiv \ell(\mathbf{x}) + \hat{r}(\mathbf{x}). \quad (13)$$

The equality constraint (12) can be dropped by replacing $\mathbf{x}_K$ with $\tilde{\mathbf{x}}_K \equiv \frac{1}{w_K}(\mathbf{x} - \sum_{k=1}^{K-1} w_k \mathbf{x}_k)$. We can then rewrite (13) as

$$\min_{\mathbf{x}, \{\mathbf{x}_k\}_{k=1}^{K-1}} \underbrace{\ell(\mathbf{x}) + \hat{r}_1\left(\mathbf{x}, \{\mathbf{x}_k\}_{k=1}^{K-1}\right)}_{\equiv f_1\left(\mathbf{x}, \{\mathbf{x}_k\}_{k=1}^{K-1}\right)} - \underbrace{\hat{r}_2\left(\mathbf{x}, \{\mathbf{x}_k\}_{k=1}^{K-1}\right)}_{\equiv f_2\left(\mathbf{x}, \{\mathbf{x}_k\}_{k=1}^{K-1}\right)}, \quad (14)$$

where

$$\begin{aligned} \hat{r}_1\left(\mathbf{x}, \{\mathbf{x}_k\}_{k=1}^{K-1}\right) &= \sum_{k=1}^{K-1} w_k \left[\frac{1}{2\eta}\|\mathbf{x}_k\|^2 + \Omega_1(\omega_k(\mathbf{x}_k))\right] \\ &\quad + w_K \left[\frac{1}{2\eta}\|\tilde{\mathbf{x}}_K\|^2 + \Omega_1\left(\omega_K\left(\tilde{\mathbf{x}}_K\right)\right)\right], \end{aligned}$$

and

$$\begin{aligned} \hat{r}_2\left(\mathbf{x}, \{\mathbf{x}_k\}_{k=1}^{K-1}\right) &= \sum_{k=1}^{K-1} w_k \Omega_2(\omega_k(\mathbf{x}_k)) \\ &\quad + w_K \Omega_2(\omega_K\left(\tilde{\mathbf{x}}_K\right)) + \frac{\|\mathbf{x}\|^2}{2\eta}. \end{aligned}$$

Next, we bound the difference between r and $\hat{r}$. It shows that $\hat{r}$ can be made arbitrarily close to r with a suitable η.

Proposition 2 $0 \leq r(\mathbf{x}) - \hat{r}(\mathbf{x}) \leq \frac{\eta\overline{L^2}}{2}$, *where* $\overline{L^2} = \sum_{k=1}^{K} w_k L_{r_k}^2$.

The following Proposition assures the monotone property of GD-PAN when $\eta < \frac{1}{L_\ell}$.

Proposition 3 *With* $\eta < \frac{1}{L_\ell}$,

$$\hat{f}(\mathbf{x}^{(t+1)}) \leq \hat{f}(\mathbf{x}^{(t)}) - \frac{\mathcal{L}}{2}\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2, \tag{15}$$

where $\mathcal{L} = \frac{1}{\eta} - L_\ell > 0$.

Finally, we show that GD-PAN, with a proper η, converges to a critical point of $\hat{f}$.

Definition 1 (Critical point (Toland 1979)) *Consider the problem* $\min g_1(\mathbf{x}) - g_2(\mathbf{x})$, *where* g_1, g_2 *are convex.* $\mathbf{x}^*$ *is a critical point if* $\mathbf{0} \in \partial g_1(\mathbf{x}^*) - \partial g_2(\mathbf{x}^*)$, *where* $\partial g_1(\mathbf{x}^*), \partial g_2(\mathbf{x}^*)$ *are the subdifferentials of* g_1 *and* g_2 *at* $\mathbf{x}^*$, *respectively.*

Theorem 1 *With* $\eta < \frac{1}{L_\ell}$, *the sequence* $\{\mathbf{x}^{(t)}\}$ *generated by GD-PAN converges, i.e.,* $\lim_{t\to\infty} \|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\| = 0$. *Let* $\mathbf{x}^* = \lim_{t\to\infty} \mathbf{x}^{(t)}$. *Then,*

$$\mathbf{0} \in \partial_{\mathbf{x}} f_1\left(\mathbf{x}^*, \{\mathbf{x}_k^*\}_{k=1}^{K-1}\right) - \partial_{\mathbf{x}} f_2\left(\mathbf{x}^*, \{\mathbf{x}_k^*\}_{k=1}^{K-1}\right),$$

where f_1, f_2 *are as defined in (14), and* $\{\mathbf{x}_k^*\}_{k=1}^{K-1}$ *are the corresponding solutions at* $\mathbf{x}^*$.

Together with Proposition 2, GD-PAN converges to a critical point of the surrogate problem (13), whose objective is arbitrarily close to the original objective in (1) with a small enough η.

Instead of using a constant stepsize, one can update $\mathbf{x}^{(t+1)}$ with line search. For given stepsizes $\eta_{\max} > \eta_{\min}$ satisfying

$$\mathcal{L} < \frac{1}{\eta_{\min}} - L_\ell, \tag{16}$$

define $\hat{r}_{\eta_{\min}(\mathbf{x})}$ as for $\hat{r}(\mathbf{x})$ in Proposition 1 but with $\eta = \eta_{\min}$. Instead of (13), consider the surrogate problem

$$\min_{\mathbf{x}} \ \hat{f}_{\eta_{\min}}(\mathbf{x}) \equiv \ell(\mathbf{x}) + \hat{r}_{\eta_{\min}}(\mathbf{x}), \tag{17}$$

and the update step is changed accordingly from (11) to

$$\mathbf{x}^{(t+1)} \leftarrow \sum_{k=1}^{K} w_k P_{r_k}^{\eta_t}(\mathbf{u}^{(t)}), \tag{18}$$

with $\eta_t = \eta_{\max}$. We then check if (15) holds for $\hat{f}_{\eta_{\min}}$. If it does not, set $\eta_t \leftarrow \eta_t/2$, repeat the update and check again. From (16) and Proposition 3, we see that $\eta_t = \eta_{\min}$ satisfies (15) and thus $\{\eta_t\}$ is bounded. Moreover, the following convergence property can be guaranteed.

Theorem 2 *The sequence* $\{\mathbf{x}^{(t)}\}$ *generated by (18) converges. Moreover, let* $\mathbf{x}^* = \lim_{t\to\infty} \mathbf{x}^{(t)}$. *Then,* $\mathbf{0} \in \partial_{\mathbf{x}} f_1\left(\mathbf{x}^*, \{\mathbf{x}_k^*\}_{k=1}^{K-1}\right) - \partial_{\mathbf{x}} f_2\left(\mathbf{x}^*, \{\mathbf{x}_k^*\}_{k=1}^{K-1}\right)$, *where* f_1, f_2 *are as defined in (14) with* $\eta = \eta_{\min}$, *and* $\{\mathbf{x}_k^*\}_{k=1}^{K-1}$ *are the corresponding solutions at* $\mathbf{x}^*$.

In other words, $\mathbf{x}^*$ is a critical point of problem (17).

There are two potential advantages of using (18) over (11): First, it may employ a more aggressive stepsize and thus converges in fewer iterations. Second, an aggressive stepsize may help to jump out of a poor local optimum, as is observed in the experiments.

In general, f is easier to compute than $\hat{f}_{\eta_{\min}}$, and the difference between them is small (from Proposition 2 and the fact that $\eta_{\min}$ is very small). Thus, in the implementation, we will use f instead of $\hat{f}_{\eta_{\min}}$ to check condition (15).

Discussion

The concave-convex procedure (CCCP) (Yuille and Rangarajan 2003) is a popular optimization tool for problems whose objective can be expressed as a difference of convex functions. For (1), CCCP first rewrites it as: $\min_{\mathbf{x}} g_1(\mathbf{x}) - g_2(\mathbf{x})$, where

$$\begin{aligned} g_1(\mathbf{x}) &= \ell(\mathbf{x}) + \sum_{k=1}^{K} w_k \Omega_1(\omega_k(\mathbf{x})), \\ g_2(\mathbf{x}) &= \sum_{k=1}^{K} w_k \Omega_2(\omega_k(\mathbf{x})), \end{aligned} \tag{19}$$

and then iteratively updates the $\mathbf{x}$ solution as

$$\mathbf{x}^{(t+1)} \leftarrow \arg\min_{\mathbf{x}} g_1(\mathbf{x}) - \nabla g_2(\mathbf{x}^{(t)})^T(\mathbf{x} - \mathbf{x}^{(t)}). \tag{20}$$

However, though (20) is convex, it can still be challenging due to the loss function and/or the composite regularizer. Typically, this requires solvers such as the linearized ADMM (Ouyang et al. 2013; Suzuki 2013), accelerated gradient descent with proximal average (Yu 2013), or Nesterov's smoothing technique (Nesterov 2005). All these converge at the rate of $\mathcal{O}\left(\frac{1}{t}\right)$, and empirically can take dozens or even hundreds of iterations. Moreover, the per-iteration complexity is also high. On the other hand, the most expensive step of GD-PAN is on computing $P_{r_k}^{\eta}(\mathbf{u}^{(t)})$ in (4), which is often efficient as r_k's are simple.

Another related algorithm based on sequential convex programming (SCP) is proposed by Lu (2012). It approximates $\ell(\mathbf{x})$ by an upper bound, and employs the update rule:

$$\begin{aligned} \mathbf{x}^{(t+1)} \leftarrow \arg\min_{\mathbf{x}} & \nabla \ell(\mathbf{x}^{(t)})^T(\mathbf{x} - \mathbf{x}^{(t)}) + \frac{\|\mathbf{x} - \mathbf{x}^{(t)}\|^2}{2\eta_t} \\ & + \sum_{k=1}^{K} w_k \Omega_1(\omega_k(\mathbf{x})) - \nabla g_2(\mathbf{x}^{(t)})^T(\mathbf{x} - \mathbf{x}^{(t)}), \end{aligned} \tag{21}$$

where η_t is a constant, and g_2 is as defined in (19). Though (21) can be easily solved when $K = 1$ and $\Omega_1(\omega_k(\mathbf{x}))$ is simple, it is difficult for $K > 1$ in general. Existing works (Barbero and Sra 2011; Mairal et al. 2010; Liu, Yuan, and Ye 2010) often convert this proximal step to its dual form, which is then solved with nonlinear optimization (such as the network flow algorithm, (accelerated) gradient descent or Newton's method). However, this approach is difficult to generalize as the dual is highly problem-dependent and also requires many iterations.

Recently, Gong et al. (2013) proposed the GIST algorithm, which updates $\mathbf{x}$ as:

$$\mathbf{x}^{(t+1)} \leftarrow \arg\min_{\mathbf{x}} \nabla \ell(\mathbf{x}^{(t)})^T(\mathbf{x}-\mathbf{x}^{(t)}) + \frac{\|\mathbf{x}-\mathbf{x}^{(t)}\|^2}{2\eta_t} + r(\mathbf{x}).$$

This is appropriate for regularizers whose proximal step is simple, such as the capped-ℓ_1, LSP, SCAD and MCP, but not for our composite regularizer (2) here.

Experiments

In this section, we demonstrate the efficiency of the proposed algorithms on a number of structured-sparsity-inducing models with nonconvex, composite regularizers. The superiority of nonconvex regularizers over their convex counterparts is also empirically shown on some real-world data sets.

Nonconvex Overlapping Group Lasso

In this section, we apply the capped-ℓ_1 penalty (Zhang 2010b) and log-sum penalty (LSP) (Candes, Wakin, and Boyd 2008) as regularizers to the overlapping group lasso (Zhao, Rocha, and Yu 2009). This leads to the nonconvex optimization problems:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{y}-\mathbf{S}\mathbf{x}\|^2 + \lambda\sum_{k=1}^{K}\min\{\|\mathbf{x}_{\mathbf{g}_k}\|,\theta\}, \qquad (22)$$

and

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{y}-\mathbf{S}\mathbf{x}\|^2 + \lambda\sum_{k=1}^{K}\log\left(1+\frac{\|\mathbf{x}_{\mathbf{g}_k}\|}{\theta}\right), \qquad (23)$$

where $\mathbf{S}\in\mathbb{R}^{n\times d}$ is the input sample matrix, and $\mathbf{y}\in\mathbb{R}^n$ is the output vector. Similar to (Yu 2013), the ground truth parameter $\mathbf{x}^*$ is constructed as $x_j^* = (-1)^j\exp(-\frac{j-1}{100})$, and the overlapping groups are defined as

$$\underbrace{\{1,\ldots,100\},\{91,\ldots,190\},\ldots,\{d-99,\ldots,d\}}_{K \text{ groups}},$$

where $d = 90K+10$. Each element of the input sample $\mathbf{s}_i\in\mathbb{R}^d$ is generated i.i.d. from the normal distribution $\mathcal{N}(0,1)$, and $y_i = \mathbf{x}^{*T}\mathbf{s}_i + \vartheta_i$, where $\vartheta_i \sim 10\times\mathcal{N}(0,1)$ is the random noise. Moreover, for (22), we vary (K,n) in $\{(5,500),(10,1000),(20,2000),(30,3000)\}$, and set $\lambda = K/10, \theta = 0.1$. For (23), we set $K=10, n=1000$, and vary (λ,θ) in $\{(0.1,0.1),(1,10),(10,10),(100,100)\}$.

The following algorithms will be compared in the experiments:

1. GD-PAN: The proposed method with update rule (11). The individual $P^{\eta}_{r_k}(\mathbf{u}^{(t)})$'s can be computed efficiently as discussed in the "Problem Formulation" section. The stepsize η is set to $\frac{1}{2L_\ell}$, where L_ℓ is the largest eigenvalue of $\frac{1}{n}\mathbf{S}^T\mathbf{S}$.
2. GD-PAN-LS: The proposed method using line search, with update rule (18). We set $\eta_{\max} = \frac{100}{L_\ell}$ and $\eta_{\min} = \frac{0.01}{L_\ell}$. As discussed before, we check condition (15) with f (rather than $\hat{f}$) and $\mathcal{L} = 10^{-5}$. While this deviates slightly from the theoretical analysis, it works well in practice.
3. CCCP: Multi-stage convex programming (Zhang 2010b) which is based on CCCP. From update rule (20), the resultant problem is a standard overlapping group lasso, which is solved by accelerated gradient descent with proximal average (Yu 2013). We use 50 iterations and warm start.
4. SCP: Sequential convex programming (Lu 2012). It can be shown that (21) is the proximal step of overlapping group lasso, which does not admit a closed-form solution. Consequently, we solve its dual by accelerated gradient descent as in (Yuan, Liu, and Ye 2011). Moreover, we use line search as in (Gong et al. 2013).

We do not compare with GIST, as it needs to solve a proximal step associated with a composite, nonconvex regularizer. Again, to the best of our knowledge, this does not have an efficient solver.

All the algorithms are implemented in MATLAB, except for the proximal step in SCP which is based on the C++ code in the SLEP package (Liu, Ji, and Ye 2009). Experiments are performed on a PC with Intel i7-2600K CPU and 32GB memory. To reduce statistical variability, all initializations start from zero, and results are averaged over 10 repetitions.

Results are shown in Figures 1 and 2. Overall, CCCP is the slowest as it has to solve a standard overlapping group lasso problem in each iteration. SCP is sometimes as fast as GD-PAN(-LS). However, it is often trapped in poor local optimum (Figures 1(c), 1(d), 2(c) and 2(d)), as has also been observed in (Gong et al. 2013). As the core of SCP is implemented in C++ while the other methods are in pure MATLAB, SCP is likely to be slower than GD-PAN(-LS) if both are implemented in the same language. Overall, GD-PAN-LS is the best and converges well in all the experiments. This is then followed by GD-PAN, which may sometimes be trapped in poor local optimum (Figure 1(d)).

Nonconvex Graph-Guided Logistic Regression

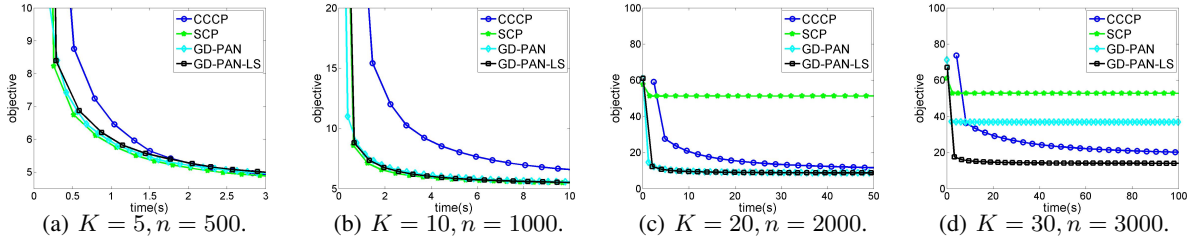
In this section, we apply the capped-ℓ_1 penalty to graph-guided logistic regression (Ouyang et al. 2013). The optimization problem becomes

$$\min_{\mathbf{x}\in\mathbb{R}^d} \ell(\mathbf{x}) + \frac{\lambda_1}{2}\|\mathbf{x}\|^2 + \lambda_2\sum_{\{k_1,k_2\}\in E}\min\{|x_{k_1}-x_{k_2}|,\theta\},$$

where $\ell(\mathbf{x}) = \sum_{i=1}^n \log(1+\exp(-y_i\mathbf{x}^T\mathbf{s}_i))$, and E contains edges for the graph defined on the d variates of $\mathbf{x}$. Following (Ouyang et al. 2013), we construct this graph by sparse inverse covariance selection on the training data (Banerjee, El Ghaoui, and d'Aspremont 2008). A similar setting is considered in (Tibshirani and Taylor 2011; Ouyang et al. 2013), though with a different loss.

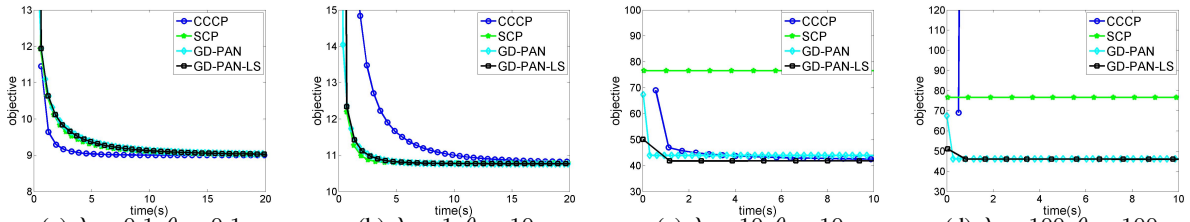
Experiments are performed on the 20newsgroup data set[2], which contains 16,242 samples with 100 binary features (words). There are 4 classes (*computer*, *recreation*, *science*, and *talks*), and we cast this as 4 one-vs-rest binary classification problems. We use 1% of the data for training, 80% for testing, and the rest for validation. Note that our main purpose here is to demonstrate the advantage of the nonconvex

[2] http://www.cs.nyu.edu/~roweis/data.html



(a) $K = 5, n = 500$. (b) $K = 10, n = 1000$. (c) $K = 20, n = 2000$. (d) $K = 30, n = 3000$.

Figure 1: Objective value versus time for the overlapping group lasso model with capped-ℓ_1 penalty.



(a) $\lambda = 0.1, \theta = 0.1$. (b) $\lambda = 1, \theta = 10$. (c) $\lambda = 10, \theta = 10$. (d) $\lambda = 100, \theta = 100$.

Figure 2: Objective value versus time for the overlapping group lasso model with log-sum penalty.

composite regularizer, rather than obtaining the best classification performance on this data set. Hence, we use logistic regression and the (convex) graph-guided logistic regression as baselines.

Results are shown in Table 1. As can be seen, the graph-guided logistic regression model with nonconvex regularizer is always the best, which is then followed by its convex counterpart, and finally regression.

Table 1: Classification accuracies (%) with graph-guided logistic regression on the 20newsgroup subset. "gg-ncvx" denotes the proposed graph-guided logistic regression with nonconvex capped-ℓ_1 regularizer; "gg-cvx" is its convex counterpart; and "lr" is logistic regression.

data set	lr	gg-cvx	gg-ncvx
com. vs rest	81.1±1.26	83.2±2.00	**85.01±1.74**
rec. vs rest	87.22±1.88	87.50±1.33	**88.59±0.89**
sci. vs rest	71.45±5.05	79.91±2.49	**84.06±1.08**
talks vs rest	82.80±2.39	82.37±3.27	**84.49±1.78**

Table 2: Classification accuracies (%) with fused lasso on the 20newsgroup subset. "fl-ncvx" denotes the proposed fused lasso with nonconvex capped-ℓ_1 regularizer; "fl-cvx" is its convex counterpart, and "lasso" is the standard lasso.

data set	lasso	fl-cvx	fl-ncvx
com. vs rest	76.90±1.96	81.91±2.00	**83.63±1.71**
rec. vs rest	81.22±1.75	85.79±2.05	**88.05±1.14**
sci. vs rest	75.13±2.05	82.11±2.15	**84.66±0.66**
talks vs rest	78.25±1.58	82.69±1.27	**84.08±1.10**

Nonconvex Fused Lasso

Here, we apply the capped-ℓ_1 penalty to the fused lasso (Tibshirani et al. 2005). The optimization problem becomes

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{y} - \mathbf{S}\mathbf{x}\|^2 + \frac{\lambda_1}{2}\|\mathbf{x}\|_1 + \lambda_2 \sum_{i=1}^{d-1} \min\{|x_i - x_{i+1}|, \theta\}.$$

As suggested in (Tibshirani et al. 2005), the features are ordered via hierarchical clustering. We use the same data set and setup as in the previous section. Results are shown in Table 2. As can be seen, the nonconvex regularizer again outperforms the rest.

Finally, we perform experiment on a breast cancer data set. As in (Jacob, Obozinski, and Vert 2009), we only use the 300 genes that are most correlated to the output, and the positive samples are reproduced twice to reduce class imbalance. 40% of the data are randomly chosen for training, another 20% for validation, and the rest for testing. Again, nonconvex fused lasso achieves the best classification accuracy of 75.69±3.96%. This is followed by the convex fused lasso (70.99±5.0%) and finally lasso (64.06±4.86%). The improvements are statistically significant according the pairwise t-test with p-value less than 0.05.

Conclusion

In this paper, we propose an efficient and simple algorithm for the optimization with a wide class of nonconvex and composite regularizers. Experimental results on a number of nonconvex sparsity-inducing models demonstrate improved accuracies. We hope this algorithm can serve as a useful tool to further popularize the use of nonconvex regularization in challenging machine learning problems.

Acknowledgment

This research was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region (Grant 614311).

References

Bach, F.; Jenatton, R.; Mairal, J.; and Obozinski, G. 2011. Convex optimization with sparsity-inducing norms. In *Optimization for Machine Learning*. 19–53.

Banerjee, O.; El Ghaoui, L.; and d'Aspremont, A. 2008. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* 9:485–516.

Barbero, A., and Sra, S. 2011. Fast Newton-type methods for total variation regularization. In *Proceedings of the 28th International Conference on Machine Learning*, 313–320.

Bauschke, H. H.; Goebel, R.; Lucet, Y.; and Wang, X. 2008. The proximal average: basic theory. *SIAM Journal on Optimization* 19(2):766–785.

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* 2(1):183–202.

Candes, E. J.; Wakin, M. B.; and Boyd, S. P. 2008. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications* 14(5-6):877–905.

Chartrand, R., and Wohlberg, B. 2013. A nonconvex ADMM algorithm for group sparsity with sparse groups. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*.

Combettes, P. L., and Pesquet, J.-C. 2011. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer. 185–212.

Fan, J., and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360.

Gong, P.; Zhang, C.; Lu, Z.; Huang, J.; and Ye, J. 2013. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of the 30th International Conference on Machine Learning*.

Gong, P.; Ye, J.; and Zhang, C. 2012. Multi-stage multi-task feature learning. In *Advances in Neural Information Processing Systems 25*. 1997–2005.

Jacob, L.; Obozinski, G.; and Vert, J. 2009. Group lasso with overlap and graph lasso. In *Proceedings of the 26th International Conference on Machine Learning*, 433–440.

Liu, J.; Ji, S.; and Ye, J. 2009. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University.

Liu, J.; Yuan, L.; and Ye, J. 2010. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining*, 323–332.

Lu, Z. 2012. Sequential convex programming methods for a class of structured nonlinear programming. Technical Report arXiv:1210.3039v1.

Mairal, J.; Jenatton, R.; Obozinski, G.; and Bach, F. 2010. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems 24*. 1558–1566.

Nesterov, Y. 2005. Smooth minimization of non-smooth functions. *Mathematical Programming* 103(1):127–152.

Nesterov, Y. 2007. Gradient methods for minimizing composite objective function. Technical Report 76, Catholic University of Louvain.

Ouyang, H.; He, N.; Tran, L.; and Gray, A. 2013. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*.

Richard, E.; Savalle, P.-A.; and Vayatis, N. 2012. Estimation of simultaneously sparse and low rank matrices. In *Proceedings of the 29th International Conference on Machine Learning*, 1351–1358.

Shalev-Shwartz, S.; Shamir, O.; and Sridharan, K. 2010. Learning kernel-based halfspaces with the zero-one loss. In *Proceedings of the 23rd Conference on Learning Theory*, 441–450.

Shen, X., and Huang, H. 2010. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association* 105(490):727–739.

Suzuki, T. 2013. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference on Machine Learning*, 392–400.

Tibshirani, R. J., and Taylor, J. 2011. The solution path of the generalized lasso. *Annals of Statistics* 39(3):1335–1371.

Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; and Knight, K. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B* 67(1):91–108.

Tibshirani, R.; Hoefling, H.; and Tibshirani, R. 2011. Nearly-isotonic regression. *Technometrics* 53(1):54–61.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58(1):267–288.

Toland, J. 1979. A duality principle for non-convex optimisation and the calculus of variations. *Archive for Rational Mechanics and Analysis* 71(1):41–61.

Wang, S.; Liu, D.; and Zhang, Z. 2013. Nonconvex relaxation approaches to robust matrix recovery. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*.

Xiang, S.; Shen, X.; and Ye, J. 2013. Efficient sparse group feature selection via nonconvex optimization. In *Proceedings of the 30th International Conference on Machine Learning*.

Yu, Y. 2013. Better approximation and faster algorithm using the proximal average. In *Advances in Neural Information Processing Systems 26*.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 68(1):49–67.

Yuan, L.; Liu, J.; and Ye, J. 2011. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems 24*, 352–360.

Yuille, A., and Rangarajan, A. 2003. The concave-convex procedure. *Neural Computation* 15(4):915–936.

Zhang, S.; Qian, H.; Chen, W.; and Zhang, Z. 2013. A concave conjugate approach for nonconvex penalized regression with the MCP penalty. In *Proceedings of the 27th National Conference on Artificial Intelligence*.

Zhang, C. 2010a. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38(2):894–942.

Zhang, T. 2010b. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research* 11:1081–1107.

Zhao, P.; Rocha, G.; and Yu, B. 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* 37(6A):3468?497.