Prediction-time Efficient Classification Using Computational Dependencies in Feature Generation (Supplementary Materials)

This supplementary material is to provide the proof process for Theorem 5.1 of the paper: **Liang Zhao, Amir Alipour-Fanid, Martin Slawski, Kai Zeng. Prediction-time Efficient Classification Using Computational Dependencies in Feature Generation. KDD 2018, submitted.** Theorem 5.1 and its proof process are as follows:

Theorem 5.1 [Convergence Analysis] *Given that the loss function $\mathcal{L}(\cdot)$ is a Lipschitz-differentiable function (e.g., logistic loss), and the function $\mathcal{L}(\Omega_1 \cdot B) + \lambda \cdot \mathcal{R}_c(H\Omega_2 B, D)$ is a coercive function [1], then the proposed ADMM-based method will converge when optimizing Equation (10) when the following conditions are satisfied: 1) the matrix $H \cdot \Omega_2$ has full rank, 2) the number of FCCs is not larger than the number of features.*

Proof. The Equation (10) can be written as the follows:

$$\min_{M,B\geq 0} \phi(M,B) = f(B) + \lambda \cdot h(M) \quad s.t.\ P \cdot M + Q \cdot B = 0 \tag{1}$$

where $f(B) = \mathcal{L}(\Omega_1 \cdot B)$ is convex and Lipschitz-differentable. And $h(M) = \mathcal{R}_c(M, D)$ is nonsmooth and nonconvex. In addition, $Q = -H \cdot \Omega_2 \in \mathbb{R}^{|E|\times 2|V|}$ and $P \in \mathbb{R}^{|E|\times|E|}$ is identity matrix. The above problem amounts to a nonconvex-regularized optimization objective with linear equality constraint on two decision variables. For this type of problem, Wang et al. [6] provided the sufficient condition for proving its convergence when using Alternating Direction Methods of Multipliers (ADMM) [2], which amounts to the following five requirements:

(1) (Coercivity) Define the feasible set $\mathcal{F} := \{(M, B) \in \mathbb{R}^{n+q} : PM + QB = 0\}$. The objective function $\phi(M, B)$ is coercive over this set, namely, $\phi(M, B) \to \infty$ if $(M, B) \in \mathcal{F}$ and $\|(M, B)\| \to \infty$.
(2) (Feasibility) $Im(P) \subseteq Im(Q)$, where $Im(\cdot)$ is defined as the image of a matrix.
(3) (Lipschitz sub-minimization paths) Three sub-requirements need to be satisfied: (i) $\Phi : Im(Q) \to \mathbb{R}^{|E|}$ defined by $\Phi(u) \triangleq \arg\min_B\{\phi(M, B) : QB = u\}$ is a Lipschitz-continuous map; (ii) $\Psi : Im(P) \to \mathbb{R}^{|E|}$ defined by $\Psi(u) \triangleq \arg\min_M\{\phi(M, B) : PM = u\}$ is a Lipschitz-continuous map; and (iii) Ψ and Φ have a positive universal Lipschitz constant.
(4) (Regularity of the regularization term) $h(M_i)$ is either "continuous and piecewise linear" or "*restricted prox-regular*". Here the *restricted prox-regular* [6] is defined as follows. For a lower semi-continuous function $\mathcal{F}$, let $m \in \mathbb{R}_+$, $\mathcal{F} : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$, and define the exclusion set $S_m := \{x \in \text{dom}(\mathcal{F}) : \|d\| > m \text{ for all } d \in \frac{\partial \mathcal{F}(x)}{\partial x}\}$. And thus $\mathcal{F}$ is called *restricted prox-regular* if, for any $m > 0$ and bounded set $T \subseteq \text{dom} f$, there exists $\theta > 0$ such that $\mathcal{F}(y)+\theta/2\|x-y\|^2 \geq \mathcal{F}(x)+\langle d, y-x\rangle$, $\forall x \in T - S_m$, $y \in T$, $d \in \frac{\partial \mathcal{F}(x)}{\partial x}$, $\|d\| \leq m$.
(5) (Regularity of the loss function term) $f(M)$ is Lipschitz-differentiable.

Since the first and fifth requirements have been satisfied in our assumptions, in the following, we prove that our problem satisfies the second, third, and fourth requirements.

Because the matrix P is identity matrix and thus it has full rank. Therefore, the Requirement (3)(i) is satisfied. Similarly, because $H \cdot \Omega_2 \in \mathbb{R}^{|E|\times 2|V|}$ has full rank and $2|V| \geq |E|$, the Requirement (3)(ii) is satisfied, too. And hence the third requirement is satisfied, too. Also because the matrix Q has full rank and $2|V| \geq |E|$, we have $\text{Im}(A_{|E|}) \subseteq \text{Im}(Q)$, where $A_{|E|} \in \mathbb{R}^{|E|\times|E|}$ is the identity matrix. This means the second requirement is also satisfied. In the following, we prove that the Requirement (4) is satisfied when using different types of regularization terms $\mathcal{R}_c(M, D)$.

In our algorithm, the re-weighted nonconvex regularization term $\mathcal{R}_c(M, D)$ has been used which can be based on commonly used nonconvex regularization terms such as MCP, SCAD, capped ℓ_1 norm, and ℓ_p quasi-norm $(0 < p < 1)$ [3].

And when $\mathcal{R}_c(M, D) = \sum_i R'_c(M_i, D_i)$ is re-weighted capped ℓ_1 norm, we have:

$$\mathcal{R}'_c(M_i, D_i) = D_i \min(|M_i|, \gamma)\ (\gamma > 0) \tag{2}$$

where it is clearly to see that $\mathcal{R}_c(M_i, D_i)$ is continuous and piece-wise linear. Thus it satisfies Requirement (4).

When $\mathcal{R}_c(M, D)$ is the re-weighted MCP term, then $\mathcal{R}_c(M, D) = \sum_i R'_c(M_i, D_i)$ is defined as follows:

$$\mathcal{R}'_c(M_i, D_i) = \begin{cases} |D_i \cdot M_i| - M_i^2/(2\gamma), & |M_i| \leq \gamma D_i \\ \frac{1}{2}\gamma D_i^2, & |M_i| \geq \gamma D_i \end{cases} \tag{3}$$

which is the maximum of a set of quadratic functions and thus can be proved to be proximal regular, as also shown in Example 2.9 in [4].

In addition, when $\mathcal{R}_c(M, D) = \mathcal{R}'_c(M_i, D_i)$ is re-weighted SCAD, we define:

$$\mathcal{R}'_c(M_i, D_i) = \begin{cases} |D_i \cdot M_i|, & |M_i| \leq D_i \\ \frac{2\gamma|D_i \cdot M_i| - M_i^2 - D_i^2}{2\gamma - 2}, & D_i < |M_i| \leq \gamma D_i \\ \frac{1}{2}(\gamma + 1) \cdot D_i^2, & |M_i| > \gamma D_i \end{cases} \tag{4}$$

which is again the maximum of a set of quadratic functions and thus its property of proximal-regular is readily to be proved by Example 2.9 in [4]. As proximal regularity is a sufficient condition of restricted proximal regularity (see [5] and Definition 1.1 in [4]), the re-weighted MCP and SCAD are proved to be restricted proximal regularity (i.e., prox-regularity).

Now, we prove when $\mathcal{R}_c(M, D)$ is re-weighted ℓ_p quasi-norm $(0 < p < 1)$ such that $\mathcal{R}_c(M, D) = \sum_i D_i |M_i|^p$, Requirement (4) still holds. The proof is extended from that in [6].

First, given any two positive constants $\alpha > 0,\ \beta > 1$, we define $c \equiv \frac{1}{3}(\frac{p}{\beta})^{\frac{1}{1-p}}$. Define a set $C_\beta = \{x | \min_{x_i \neq 0} |x_i| \leq 3c\}$. Then for any vectors $z, y \in \mathbb{R}^{|E|}$ such that $z \in \{x | \|x\| < \alpha\} - C_\beta$ and $y \in \{x | \|x\| \in \alpha\}$, if $\|z - y\| \leq c$, then $\text{supp}(z) \subset \text{supp}(y)$, where $\text{supp}(z)$ denotes the index set of all non-zero elements of z. Also define y' whose element is defined as: $y'_i = y_i \cdot I'(i)$, where $I'(i)$ is the indicator function such that $I'(i \in \text{supp}(z_i)) = 1$ while $I'(i \notin \text{supp}(z_i)) = 0$.

Define $d(z) = \frac{\partial \mathcal{R}_c(z,D)}{\partial z}$ and $d'(z) = \frac{\partial \mathcal{R}(z)}{\partial z}$, where $\mathcal{R}(z) = \sum_i |z_i|^p$ and hence for each element $d(z_i) = D_i \cdot d'(z_i)$. Therefore, when $\|z - y\| \leq c$, we have:

$$\mathcal{R}_c(y, D) - \mathcal{R}_c(z, D) - \langle d, y - z \rangle \tag{5}$$

$$= \sum_i D_i |y_i|^p - \sum_i D_i |z_i|^p - \sum_i d_i \cdot (y_i - z_i) \tag{6}$$

$$\geq \sum_i D_i |y'_i|^p - \sum_i D_i |z_i|^p - \sum_i D_i \cdot d'_i \cdot (y'_i - z_i) \qquad \text{(by the definition of } y' \text{and } d') \tag{7}$$

$$\geq -\sum_i \frac{p(1-p)}{2} c^{p-2} D_i \cdot (z_i - y'_i)^2 \qquad \Big(\text{the second order derivative of } \mathcal{R}(z) \text{ is no bigger than } p(1-p)c^{p-2}\Big) \tag{8}$$

$$\geq -\sum_i \frac{p(1-p)}{2} c^{p-2} D_i \cdot (z_i - y_i)^2 \tag{9}$$

$$\geq -|E| \cdot D_{\max} \cdot \frac{p(1-p)}{2} c^{p-2} \cdot \|z - y\|^2 \qquad \Big(D_{\max} = \max_i D_i\Big) \tag{10}$$

When $\|z - y\| > c$, we have:

$$\mathcal{R}_c(y, D) - \mathcal{R}_c(z, D) - \langle d, y - z \rangle \tag{11}$$

$$= \sum_i D_i |y_i|^p - \sum_i D_i |z_i|^p - \sum_i d_i \cdot (y_i - z_i) \tag{12}$$

$$= \sum_i D_i \left(|y_i|^p - |z_i|^p - d'_i \cdot (y_i - z_i) \right) \qquad \text{(by the definition of } y' \text{and } d') \tag{13}$$

$$\geq -\sum_i D_i (2\alpha^p + 2\alpha\beta) \tag{14}$$

$$\geq -\left(\frac{2\alpha^p + 2\alpha\beta}{c^2} \sum_i D_i \right) \|y - z\|^2 \qquad (\|z - y\| > c) \tag{15}$$

Therefore, according to the definition of restricted proximal regularity given in the Requirement (4), Requirement (4) is proved to be satisfied when $\mathcal{R}_c$ is re-weighted ℓ_p quasi-norm $(0 < p < 1)$.

In all, when $\mathcal{R}_c(M, D)$ is the re-weighted version of commonly used nonconvex regularization terms that satisfy Requirement (4), the proposed ADMM-based algorithm will converge based on the aforementioned conditions.

□

REFERENCES

[1] Aleksandr Y Aravkin, James V Burke, and Gianluigi Pillonetto. 2013. Sparse/robust estimation and kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory. *The Journal of Machine Learning Research* 14, 1 (2013), 2689–2728.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3, 1 (2011), 1–122.

[3] Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. 2013. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning*. 37–45.

[4] René Poliquin and R Rockafellar. 1996. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.* 348, 5 (1996), 1805–1838.

[5] Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. 2015. Efficient learning by directed acyclic graph for resource constrained prediction. In *Advances in Neural Information Processing Systems*. 2152–2160.

[6] Yu Wang, Wotao Yin, and Jinshan Zeng. 2015. Global convergence of ADMM in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324* (2015).