Supplementary Material for Sparse Partially Linear Additive Models

Yin Lou * Jacob Bien † Rich Caruana ‡ Johannes Gehrke §

August 23, 2015

1 Convergence of Algorithm 1 in the Main Paper

We show that Algorithm 1 fits the general BCGD framework (Tseng and Yun, 2009) and therefore the global convergence is guaranteed. We include this supplementary material for completeness although a similar convergence result for the group lasso is shown in Qin *et al.* (2010). We first briefly review the general BCGD algorithm.

Let $F(\beta) = L(\beta) + h(\beta)$, where $h(\beta) = \lambda\Omega^{SPLAM}(\beta)$. At each iteration k, for block j, choose a symmetric positive definite matrix H^k, and compute the search direction.

$$d^k = \arg\min_d \{\nabla L(\beta^k)^T d + \frac{1}{2} d^T H^k d + h(\beta^k + d)\} \tag{1}$$

where $\forall i \notin \mathcal{G}_j, d_i = 0$. Then a step size $\alpha^k > 0$ is chosen so that the following Armijo rule is

*Department of Computer Science, Cornell University
†Departments of BSCB and Statistical Science, Cornell University
‡Microsoft Research, Microsoft Corporation
§Department of Computer Science, Cornell University

satisfied,

$$F(\beta^k + \alpha^k d^k) \leq F(\beta^k) + \alpha^k \sigma \Delta^k \tag{2}$$

where $0 < \sigma < 1, 0 \leq \gamma < 1$, and

$$\Delta^k \stackrel{\text{def}}{=} \nabla L(\beta^k)^T d^k + \gamma {d^k}^T H^k d^d + h(\beta^k + d^k) - h(\beta^k), \tag{3}$$

Once the step size α^k is determined, update $\beta^{k+1} = \beta^k + \alpha^k d^k$.

Theorem 2 in Tseng and Yun (2009) guarantees the global convergence when $\overline{\theta} I \succeq H^k \succeq \underline{\theta} I$, $0 < \underline{\theta} \leq \overline{\theta}$.

Theorem 1. *Algorithm 1 fits the general BCGD framework of Tseng and Yun (2009). The global convergence is guaranteed and Algorithm 1 converges Q-linearly.*

Proof. First, for block j, setting $H^k = \frac{1}{t_j} I$, Equation (1) is equivalent to our proximal operator for block j after ignoring constants. Next, notice that when $\alpha^k = 1, \sigma = 1$, and $\gamma = \frac{1}{2}$, the Armijo rule becomes our backtracking line search step in Equation (10) in the main paper. That is, the effort of choosing step size is shifted to finding H^k. Besides, Lemma 1 in Tseng and Yun (2009) suggests $\nabla L(\beta^k)^T d^k + {d^k}^T H^k d^k + h(\beta^k + d^k) - h(\beta^k) \leq 0$. Since $H^k \succ 0$, with $\gamma = \frac{1}{2}$, we can easily see $\Delta^k \leq 0$ whenever $d^k \neq 0$, which means if the Armijo rule holds for $\sigma = 1$, it must also hold for $\sigma < 1$. Finally, we show that $\overline{\theta} I \succeq H^k \succeq \underline{\theta} I$. Assume the initial step size is t_j^0, this is true when $\overline{\theta} = \max\{C_j, 1/t_j^0\}$ and $\underline{\theta} = \min\{C_j, 1/t_j^0\}$. Thus, according to Theorem 2 in Tseng and Yun (2009), Algorithm 1 converges Q-linearly. □

Algorithm 1 Finding λ_{max}

1: $\lambda_h \leftarrow \max_j \frac{\|\nabla_j L(0)\|_2}{\alpha}$
2: $\lambda_l \leftarrow 0$
3: **while** $\lambda_h - \lambda_l \geq \epsilon$ **do**
4: $\quad \lambda \leftarrow \frac{\lambda_h + \lambda_l}{2}$
5: $\quad$ **if** $\forall j, P_t^j(0) = 0$ **then**
6: $\quad\quad \lambda_h \leftarrow \lambda$
7: $\quad$ **else**
8: $\quad\quad \lambda_l \leftarrow \lambda$
9: $\lambda_{max} = \lambda_h$

1.1 Practical Issues

1.1.1 Active Set Strategy

We employ the widely used active set strategy (Friedman *et al.*, 2010; Krishnapuram *et al.*, 2005; Meier *et al.*, 2008). After a complete cycle through all the variables, we iterate only on the active set till convergence. If another complete cycle does not change the active set, we are done, otherwise the process is repeated.

1.1.2 Regularization Path

Similar to `glmnet` (Friedman *et al.*, 2010), the optimization of SPLAM also uses two parameters, λ and α, which usually involves a grid search on values of (λ, α) pairs. As noted in the main paper, for each value of α, we start at the smallest value λ_{max} for which $\beta_j = 0$ for $j = 1, ..., p$. We then decrease λ from λ_{max} exponentially. To find λ_{max}, we note that for all $\lambda \geq \lambda_{init} = \max_j \frac{\|\nabla_{\beta_j} L(0)\|_2}{\alpha}$, the zero vector is the solution to our optimization problem. We perform a binary search to find λ_{max}. As described in Algorithm 1, we start with λ_{init} (Line 1) and effectively shrink the interval $[\lambda_l, \lambda_h]$ (Line 3 - 8) to locate λ_{max}.

2 Proof of Theorem 1 and Corollary 1 in the Main Paper

Proof of Theorem 1. By definition of $\hat{\beta}$,

$$\frac{1}{2N}\|y - X\hat{\beta}\|_2^2 + \lambda\Omega^{SPLAM}(\hat{\beta}) \leq \frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\Omega^{SPLAM}(\beta)$$

holds for any $\beta \in \mathbb{R}^p$. Some algebra (recalling that $y = f^0 + \epsilon$ and writing $\hat{\Delta} = \hat{\beta} - \beta$) leads to

$$\frac{1}{2N}\|X\hat{\beta} - f^0\|_2^2 + \lambda\Omega^{SPLAM}(\hat{\beta}) \leq \frac{1}{2N}\|X\beta - f^0\|_2^2 + \frac{1}{N}\epsilon^T X\hat{\Delta} + \lambda\Omega^{SPLAM}(\beta) \tag{4}$$

Define the empirical process as,

$$V_N(\hat{\Delta}) = \frac{1}{N}\epsilon^T X\hat{\Delta} = \frac{1}{\sqrt{N}}\sum_{j=1}^{p} V_j^T\hat{\Delta}_j \tag{5}$$

where $V_j = \frac{1}{\sqrt{N}}X_j^T\epsilon \in \mathbb{R}^M$.

Now we bound the empirical process. First we notice that,

$$|V_j^T\hat{\Delta}_j| \leq \frac{1}{2}\left[|V_j^T\hat{\Delta}_j| + |V_{j1}\hat{\Delta}_{j1}| + |V_{j,-1}^T\hat{\Delta}_{j,-1}|\right] \tag{6}$$

$$\leq \frac{1}{2}\left[\|V_j\|_2\|\hat{\Delta}_j\|_2 + |V_{j1}||\hat{\Delta}_{j1}| + \|V_{j,-1}\|_2\|\hat{\Delta}_{j,-1}\|_2\right] \tag{7}$$

Thus $|V_N(\hat{\Delta})|$ can be bounded as follows.

$$
\begin{aligned}
|V_N(\hat{\Delta})| \leq & \frac{1}{\sqrt{N}} \sum_{j=1}^{p} |V_j^T \hat{\Delta}_j| && (8)\\
\leq & \frac{1}{2\sqrt{N}} \left(\max_j \|V_j\|_2 \|\hat{\Delta}\|_{2,1} + \max_j |V_{j1}| \sum_j |\hat{\Delta}_{j1}| + \max_j \|V_{j,-1}\|_2 \|\hat{\Delta}_{\cdot,-1}\|_{2,1} \right) && (9)\\
\leq & \frac{1}{2\sqrt{N}} \left[(\max_j \|V_j\|_2 + \max_j |V_{j1}|) \|\hat{\Delta}\|_{2,1} + \max_j \|V_{j,-1}\|_2 \|\hat{\Delta}_{\cdot,-1}\|_{2,1} \right] && (10)
\end{aligned}
$$

Observing that $V_j \sim N(0, \sigma^2 I_M)$, we have $\|V_j\|_2^2 \sim \sigma^2 \chi_M^2$. Thus, by Lemma 6.2 and 8.1 of Bühlmann and van de Geer (2011), we have

$$
\begin{aligned}
P\left(\frac{\max_j |V_{j1}|}{2\sqrt{N}} > \nu_1 \right) \leq & 2e^{-x} && (11)\\
P\left(\frac{\max_j \|V_j\|_2}{2\sqrt{N}} > \nu_2 \right) \leq & e^{-x} && (12)\\
& && (13)
\end{aligned}
$$

where,

$$
\begin{aligned}
\nu_1^2 = & \frac{\sigma^2}{2N}(x + \log p) && (14)\\
\nu_2^2 = & \frac{\sigma^2}{4N} \left[M + \sqrt{4M(x + \log p)} + 4(x + \log p) \right] && (15)
\end{aligned}
$$

Thus, we have

$$P\left(\frac{\max_j \|V_j\|_2 + \max_j |V_{j1}|}{2\sqrt{N}} > \nu_1 + \nu_2\right) \leq 3e^{-x} \tag{16}$$

$$P\left(\frac{\max_j \|V_{j,-1}\|_2}{2\sqrt{N}} > \nu_2\right) \leq e^{-x} \tag{17}$$

Therefore (with union bound),

$$P\left(|V_N(\hat{\Delta})| \leq \left[(\nu_1 + \nu_2)\|\hat{\Delta}\|_{2,1} + \nu_2\|\hat{\Delta}_{\cdot,-1}\|_{2,1}\right]\right) \geq 1 - (e^{-x} + 3e^{-x}) \tag{18}$$

$$= 1 - 4e^{-x} \tag{19}$$

Thus, by (4) we have with probability at least $1 - 4e^{-x}$ that

$$\frac{1}{2N}\|X\hat{\beta} - f^0\|_2^2 + \lambda\Omega^{SPLAM}(\hat{\beta}) \leq \frac{1}{2N}\|X\beta - f^0\|_2^2 + (\nu_1 + \nu_2)\|\hat{\Delta}\|_{2,1} + \nu_2\|\hat{\Delta}_{\cdot,-1}\|_{2,1} + \lambda\Omega^{SPLAM}(\beta) \tag{20}$$

Let $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)$, we can take $\lambda_1 = 2(\nu_1 + \nu_2)$ and $\lambda_2 = 2\nu_2$. Thus, (20) implies

$$\frac{1}{2N}\|X\hat{\beta} - f^0\|_2^2 - \frac{1}{2N}\|X\beta - f^0\|_2^2 \leq (\lambda/2)\Omega^{SPLAM}(\hat{\Delta}) - \lambda\Omega^{SPLAM}(\hat{\beta}) + \lambda\Omega^{SPLAM}(\beta) \tag{21}$$

$$\leq (\lambda/2)\left[\Omega^{SPLAM}(\hat{\beta}) + \Omega^{SPLAM}(\beta)\right] - \lambda\Omega^{SPLAM}(\hat{\beta}) + \lambda\Omega^{SPLAM}(\beta) \tag{22}$$

$$= (3\lambda/2)\Omega^{SPLAM}(\beta) - (\lambda/2)\Omega^{SPLAM}(\hat{\beta}) \tag{23}$$

by the triangle inequality. Thus,

$$\frac{1}{2N}\|X\hat{\beta} - f^0\|_2^2 \leq \frac{1}{2N}\|X\beta - f^0\|_2^2 + 3\lambda\Omega^{SPLAM}(\beta). \tag{24}$$

By choosing $x = \log p$, we can ensure our inequality holds with probability at least $1 - 4/p$. This means,

$$\nu_1^2 = \frac{\sigma^2}{N}\log p \tag{25}$$

$$\nu_2^2 = \frac{\sigma^2}{4N}\left[M + \sqrt{8M\log p} + 8\log p\right]. \tag{26}$$

Define $\tilde{\nu}_1^2 \stackrel{\text{def}}{=} \frac{\sigma^2}{N}\log p$ and notice that $\nu_2^2 \leq 6\sigma^2 \log p/N \stackrel{\text{def}}{=} \tilde{\nu}_2^2$ if $\log p \geq M/8$. Now, as long as $\log p \geq M/8$, we can take $\lambda \geq 2(\tilde{\nu}_1 + 2\tilde{\nu}_2) = 2(1 + 2\sqrt{6})\sigma\sqrt{\log p/N}$ and

$$\alpha = \frac{\tilde{\nu}_1 + \tilde{\nu}_2}{\tilde{\nu}_1 + 2\tilde{\nu}_2} = \frac{1 + \sqrt{6}}{1 + 2\sqrt{6}}, \tag{27}$$

with probability at least $1 - 4/p$, we have

$$\frac{1}{2N}\|X\hat{\beta} - f^0\|_2^2 \leq \frac{1}{2N}\|X\beta - f^0\|_2^2 + 3\lambda\Omega^{SPLAM}(\beta). \tag{28}$$

This holds simultaneously for all β; this may be succinctly expressed by adding $\min_\beta$ to the right hand side.

□

Proof of Corollary 1. We plug β^0 into the right-hand side of Theorem 1 and observe that

$$\Omega^{SPLAM}(\beta^0) = \alpha \sum_{j \in S^0} \|\beta_j^0\|_2 + (1-\alpha) \sum_{j \in N^0} \|\beta_{j,-1}^0\|_2 \tag{29}$$

$$\leq \alpha \sum_{j \in L^0} \|\beta_j^0\|_2 + \sum_{j \in N^0} \|\beta_j^0\|_2 \tag{30}$$

$$= \alpha \sum_{j \in L^0} |\beta_{j1}^0| + \sum_{j \in N^0} \|\beta_j^0\|_2 \tag{31}$$

□

3 Proof of Lower Bound on SpAM's Prediction Error

We assume that all p features are linear with equal coefficients, i.e., $\beta_j^0 = be_1 \in \mathbb{R}^M$ and consider an asymptotic regime in which p is fixed and $N = pM$, with $M, N \to \infty$. We assume that all features are orthogonal, i.e., $\frac{1}{N} X^T X = I_{pM}$. In the main paper, we note that SpAM in this case is given by the expression:

$$\hat{\beta}_j^{SpAM} = \gamma_j(\lambda) \frac{1}{N} X_j^T y \quad \text{where} \quad \gamma_j(\lambda) = \left(1 - \frac{\lambda}{\|\frac{1}{N} X_j^T y\|_2}\right)_+ .$$

Now, $\frac{1}{N} X_j^T y = be_1 + U_j$ where $U_j = \frac{1}{N} X_j^T \epsilon \sim N(0, \frac{\sigma^2}{N} I_M)$. Since $\|\frac{1}{N} X_j^T y\|_2^2 \to b^2 + \sigma^2/p$, asymptotically, the shrinkage factor $\gamma_j(\lambda) = \gamma$ is a nonrandom value, not depending on j, and the prediction

error is

$$\begin{aligned}\frac{1}{N}\|X\hat{\beta}^{SpAM} - X\beta^0\|^2 &= \sum_{j=1}^{p} \|\gamma(be_1 + U_j) - be_1\|^2 \\ &= \gamma^2(b^2 p + \sum_{j=1}^{p} \left[\|U_j\|^2 + 2bU_{j1}\right]) + pb^2 - 2\gamma \sum_{j=1}^{p} b(b + U_{j1}) \\ &\to \gamma^2(b^2 p + \sigma^2) + pb^2 - 2\gamma pb^2.\end{aligned}$$

For the best possible asymptotic error, we can choose $\gamma = pb^2/(pb^2 + \sigma^2)$ (equivalent to choosing the best λ). At this value,

$$\lim_{N\to\infty} \frac{1}{N}\|X\hat{\beta}^{SpAM} - X\beta^0\|^2 \geq \frac{b^2}{1/b^2 + p/\sigma^2} > 0.$$

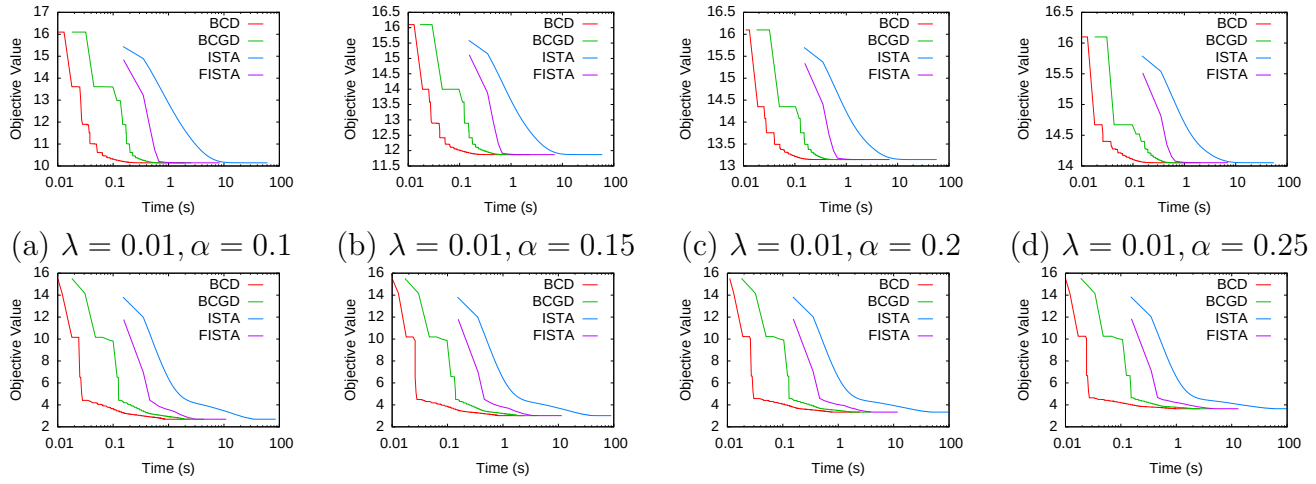
Thus, SpAM is not consistent in terms of prediction error in this asymptotic regime.

To see that SPLAM with $\lambda\alpha = 0$ and $\lambda(1-\alpha) = \infty$ is consistent in terms of prediction error, observe that $\hat{\beta}^{SPLAM} = (X_{j1}^T y)e_1 = (b + U_{j1})e_1$ and

$$\frac{1}{N}\|X\hat{\beta}^{SPLAM} - X\beta^0\|^2 = \sum_{j=1}^{p} \|(b + U_{j1})e_1 - be_1\|^2 = \sum_{j=1}^{p} U_{j1}^2 \sim \frac{\sigma^2}{N}\chi_p^2 \to 0.$$

4 Experiments

In this section, we compare our BCGD algorithm and BCD algorithm with ISTA and FISTA (Beck and Teboulle, 2009) using the synthetic function in Section 5.1. We report running time of all the methods on a single core. For BCGD, ISTA, and FISTA, we start with a same initial step size. For fair comparison, we turn off the active set strategy in BCGD and BCD, and we directly use the



(a) $\lambda = 0.01, \alpha = 0.1$ (b) $\lambda = 0.01, \alpha = 0.15$ (c) $\lambda = 0.01, \alpha = 0.2$ (d) $\lambda = 0.01, \alpha = 0.25$

(e) $\lambda = 0.001, \alpha = 0.1$ (f) $\lambda = 0.001, \alpha = 0.15$ (g) $\lambda = 0.001, \alpha = 0.2$ (h) $\lambda = 0.001, \alpha = 0.25$

Figure 1: *Objective value vs. running time for synthetic dataset in Section 5.1.*

design matrix after QR decomposition so that all methods are applied to the same optimization problem.

Figure 1 illustrates the running time for all methods using the same synthetic dataset in Section 5.1 for different combinations of λ and α. As expected, FISTA converges much faster than ISTA. However, the BCGD algorithm is faster than both of the these methods. This is because BCGD uses more information in the sense of more frequent updates. In addition, we can see that BCD further speeds up the optimization since there is no step size in BCD; this not only solves exactly the subproblem but also avoids the possibility of dampening the step size and repeating the computation on the same block.

References

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.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1.

Krishnapuram, B., Carin, L., Figueiredo, M., and Hartemink, A. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *Transactions on Pattern Analysis and Machine Intelligence*, **27**(6), 957–968.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 53–71.

Qin, Z., Scheinberg, K., and Goldfarb, D. (2010). Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation*, pages 1–27.

Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, **117**(1-2), 387–423.