Supplementary Materials to "*SparseNet*: Coordinate Descent with Non-Convex Penalties"

Rahul Mazumder Jerome Friedman Trevor Hastie

1 Supplementary Materials

This portion has three main parts. In Section 1.1, we present certain Lemmas and their proofs, complementing the technical results related to Convergence Analysis. Sections 1.2 and 1.3 describes properties and proofs related to the calibration of MC+ and properties of $\lambda_S, \gamma\lambda_S$. Section 1.4 takes a closer look at the LLA and MLLA via simple examples.

1.1 Some Lemmas related to Convergence Analysis

We will first address the boundedness of the sequence $\boldsymbol{\beta}^k$. For this purpose we present the following Lemma:

Lemma 1. *Suppose the data $(y, \mathbf{X})$ lies on a compact set. For fixed $P(\cdot;\lambda;\gamma)$ and $\boldsymbol{\beta}_0 \in \Re^p$ let $\mathcal{H}_\prime = \{\boldsymbol{\beta} | Q(\boldsymbol{\beta}) \leq Q(\boldsymbol{\beta}_0)\}$. There exists a bounded set $\tilde{\mathcal{H}}_0 \subset \Re^p$ such that*

$$\{Q(\boldsymbol{\beta}) | \boldsymbol{\beta} \in \mathcal{H}_\prime\} = \{Q(\boldsymbol{\beta}) | \boldsymbol{\beta} \in \tilde{\mathcal{H}}_0\} \tag{1}$$

Proof. This lemma is trivially true if $P(|t|;\lambda;\gamma)$ is unbounded in $|t|$.

The lemma requires a proof if $P(|t|;\lambda;\gamma)$ is uniformly bounded in $|t|$ (MC+ penalty, for example). We prove this part via contradiction.

Suppose there exists a sequence, $\{\boldsymbol{\beta}^k\}_k \subset \mathcal{H}_\prime$ such that $\|\boldsymbol{\beta}^k\|_2 \to \infty$. This necessarily implies that, there is a subset of indices $(i_1, \ldots, i_m) \subset \{1, \ldots, p\}$ such that $\{(\beta^k_{i_1}, \ldots, \beta^k_{i_m})\}_k$ is an unbounded sequence. This sequence satisfies (passing on to a subsequence if necessary),

$$\mathbf{X}[i_1, \ldots, i_m](\beta^k_{i_1}, \ldots, \beta^k_{i_m})' = \mathbf{X}[i_1, \ldots, i_m](\beta^0_{i_1}, \ldots, \beta^0_{i_m})'$$

for some $(\beta^0_{i_1}, \ldots, \beta^0_{i_m}) \in \Re^m$ independent of k and $\mathbf{X}[i_1, \ldots, i_m] = [\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m}]$.

The sequence $\{\boldsymbol{\beta}^k\}_k \subset \mathcal{H}_\prime$, considered above is arbitrary. Hence the range of values taken by the function $Q(\boldsymbol{\beta})$ remains the same if one restricts to bounded subsets of $\Re^p$. So there exists a bounded set $\tilde{\mathcal{H}}_0 \subset \Re^p$ such that (1) holds true. □

Lemma 1, shows that the boundedness assumption on the coordinate-wise updates is not by any means restrictive. If required, one can restrict $\|\boldsymbol{\beta}\|_\infty \leq M$, in the Algorithm for some $M > 0$, sufficiently large.

Lemmas 1 and 2 characterize the fate of *limit points* of the sequence produced by the coordinate-wise updates under milder conditions than those required by Theorem 4.

Proof of Lemma 1

Proof. For any $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\boldsymbol{\delta}_i = (0, \ldots, \delta_i, \ldots, 0) \in \Re^p$,

$$\liminf_{\alpha \downarrow 0+} \left\{ \frac{Q(\boldsymbol{\beta} + \alpha \boldsymbol{\delta}_i) - Q(\boldsymbol{\beta})}{\alpha} \right\} = \nabla_i f(\boldsymbol{\beta}) \delta_i + \liminf_{\alpha \downarrow 0+} \left\{ \frac{P(|\beta_i + \alpha \delta_i|) - P(|\beta_i|)}{\alpha} \right\} \quad (2)$$

where $f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, ∇_i is the partial derivative w.r.t. the i^{th} coordinate.

Observe that the second term in (2) simplifies to

$$\begin{aligned} \partial P(\beta_i; \delta_i) &:= \liminf_{\alpha \downarrow 0+} \left\{ \frac{P(|\beta_i + \alpha \delta_i|) - P(|\beta_i|)}{\alpha} \right\} \\ &= \begin{cases} P'(|\beta_i|) \operatorname{sgn}(\beta_i) \delta_i & \text{if } |\beta_i| > 0; \\ P'(0)|\delta_i| & \text{if } |\beta_i| = 0. \end{cases} \end{aligned} \quad (3)$$

where

$$\operatorname{sgn}(x) \in \begin{cases} \{1\} & \text{if } x > 0; \\ \{-1\} & \text{if } x < 0; \\ \text{[-1,1]} & \text{if } x = 0. \end{cases}$$

Assume $\boldsymbol{\beta}^{n_k} \to \boldsymbol{\beta}^\infty = (\beta_1^\infty, \ldots, \beta_p^\infty)$. Using (3) and Assumption 3, as $k \to \infty$

$$\boldsymbol{\beta}_i^{n_k - 1} := (\beta_1^{n_k}, \ldots, \beta_{i-1}^{n_k}, \beta_i^{n_k}, \beta_{i+1}^{n_k - 1}, \beta_p^{n_k - 1}) \;\to\; (\beta_1^\infty, \ldots, \beta_{i-1}^\infty, \beta_i^\infty, \beta_{i+1}^\infty, \beta_p^\infty) \quad (4)$$

If $\beta_i^\infty \neq 0, \partial P(\beta_i^{n_k}; \delta_i) \to \partial P(\beta_i^\infty; \delta_i)$; If $\beta_i^\infty = 0, \partial P(\beta_i^\infty; \delta_i) \geq \liminf_k \partial P(\beta_i^{n_k}; \delta_i)$

By definition of a coordinate-wise minimum, using (2,3) we have

$$\nabla_i f(\boldsymbol{\beta}_i^{n_k - 1}) \delta_i + \partial P(\beta_i^{n_k}; \delta_i) \geq 0 \; \forall k \quad (5)$$

The above imply $\forall i \in \{1, \ldots, p\}$

$$\nabla_i f(\boldsymbol{\beta}^\infty) \delta_i + \partial P(\beta_i^\infty; \delta_i) \geq \liminf_k \left\{ \nabla_i f(\boldsymbol{\beta}_i^{n_k - 1}) \delta_i + \partial P(\beta_i^{n_k}; \delta_i) \right\} \underbrace{\geq}_{\text{by (5)}} 0 \quad (6)$$

Using (2,3) the l.h.s. of (6) gives

$$\liminf_{\alpha \downarrow 0+} \left\{ \frac{Q(\boldsymbol{\beta}^\infty + \alpha \boldsymbol{\delta}_i^\infty) - Q(\boldsymbol{\beta}^\infty)}{\alpha} \right\} \geq 0 \quad (7)$$

For $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p) \in \Re^p$, due to differentiability of $f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$

$$\begin{aligned} \liminf_{\alpha \downarrow 0+} \left\{ \frac{Q(\boldsymbol{\beta}^\infty + \alpha \boldsymbol{\delta}) - Q(\boldsymbol{\beta}^\infty)}{\alpha} \right\} &= \sum_{j=1}^p \nabla_j f(\boldsymbol{\beta}^\infty) \delta_j + \sum_{j=1}^p \lim_{\alpha \downarrow 0+} \frac{P(|\beta_j^\infty + \alpha \delta_j|) - P(|\beta_j^\infty|)}{\alpha} \\ &= \sum_{j=1}^p \left\{ f(\boldsymbol{\beta}^\infty) \delta_j + \lim_{\alpha \downarrow 0+} \frac{P(|\beta_j^\infty + \alpha \delta_j|) - P(|\beta_j^\infty|)}{\alpha} \right\} \\ &= \sum_{j=1}^p \liminf_{\alpha \downarrow 0+} \left\{ \frac{Q(\boldsymbol{\beta}^\infty + \alpha \boldsymbol{\delta}_j^\infty) - Q(\boldsymbol{\beta}^\infty)}{\alpha} \right\} \underbrace{\geq}_{\text{By (7)}} 0 \end{aligned} \quad (8)$$

This completes the proof. □

Lemma 2 gives a sufficient condition under which Assumption 3 of Lemma 1 is true.

Lemma 2. *Suppose in addition to Assumption 1 of Lemma 1 the following assumptions hold:*

1. *For every $i = 1, \ldots, p$ and $(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_p) \in \Re^{p-1}$ the univariate function*

$$\chi^i_{(u_1,\ldots,u_{i-1},u_{i+1},\ldots,u_p)}(\cdot) : u \mapsto \quad Q(u_1, \ldots, u_{i-1}, u, u_{i+1}, \ldots, u_p) \tag{9}$$

   has a unique global minimum.

2. *If $u_G \equiv u_G(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_p)$ is the global minimum and $u_L \equiv u_L(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_p)$ any other local minimum of the function (9); there exists $\epsilon > 0$, independent of the choice of $(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_p)$ such that*

$$|u_G(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_p) - u_L(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_p)| \geq \epsilon$$

   — that is u_G and u_L are uniformly separated.

Then for every convergent subsequence $\{\boldsymbol{\beta}^{n_k}\}_k \subset \{\boldsymbol{\beta}^k\}_k$, the successive differences $\boldsymbol{\beta}^{n_k} - \boldsymbol{\beta}^{n_k-1} \to 0$; i.e. Assumption 3 of Lemma 1 is true.

Proof. Borrowing notation from (4), consider the sequence $\{\boldsymbol{\beta}^{n_k-1}_{p-1}\}$

$$\boldsymbol{\beta}^{n_k-1}_{p-1} := (\beta_1^{n_k}, \ldots, \beta_{p-2}^{n_k}, \beta_{p-1}^{n_k}, \beta_p^{n_k-1}) \tag{10}$$

where $\boldsymbol{\beta}^{n_k-1}_{p-1}$ differs from $\boldsymbol{\beta}^{n_k}$ at the p^{th} coordinate. We will show $\boldsymbol{\beta}^{n_k-1}_{p-1} - \boldsymbol{\beta}^{n_k} \to 0$ by the method of contradiction.

Writing $\boldsymbol{\beta}^{n_k} - \boldsymbol{\beta}^{n_k-1}_{p-1} = \boldsymbol{\Delta}^{n_k}_p := (0, \ldots, 0, \Delta^{n_k}_p)$ we have

$$Q(\boldsymbol{\beta}^{n_k} - \boldsymbol{\Delta}^{n_k}_p) - Q(\boldsymbol{\beta}^{n_k}) \downarrow 0 \text{ as } n_k \to \infty$$

As $k \to \infty$, let $\boldsymbol{\beta}^{n_k} \to \boldsymbol{\beta}^+$. Passing on to a subsequence (if necessary) we have $\Delta^{n_k}_p \to \Delta^+_p \neq 0$ (as $\Delta^{n_k}_p \nrightarrow \mathbf{0}$) and

$$\boldsymbol{\beta}^{n_k} - \boldsymbol{\Delta}^{n_k}_p \to \boldsymbol{\beta}^+ - \boldsymbol{\Delta}^+_p$$

Furthermore, the continuity of the function $Q(\cdot)$ implies

$$\big(Q(\boldsymbol{\beta}^{n_k} - \boldsymbol{\Delta}^{n_k}_p) - Q(\boldsymbol{\beta}^{n_k})\big) \downarrow \big(Q(\boldsymbol{\beta}^+ - \boldsymbol{\Delta}^+_p) - Q(\boldsymbol{\beta}^+)\big) = 0 \tag{11}$$

By the definition of $\boldsymbol{\beta}^{n_k} := (\beta_1^{n_k}, \ldots, \beta_p^{n_k})$ we have

$$\nabla_p f(\boldsymbol{\beta}^{n_k})\delta_p + \partial P(\beta_p^{n_k}; \delta_p) \geq 0, \quad \forall k$$

Passing on to the limit $k \to \infty$ we get (using arguments as in proof of Lemma 1)

$$\nabla_p f(\boldsymbol{\beta}^+)\delta_p + \partial P(\beta_p^+; \delta_p) \geq 0$$

where $\boldsymbol{\beta}^+ = (\beta_1^+, \ldots, \beta_p^+)$. This implies that $\chi^p_{(\beta_1^+,\ldots,\beta_{p-1}^+)}(u)$ has *a minimum* at $u = \beta_p^+$. By Assumption 2 of this Lemma the global minimum u_G is uniformly separated from other local minima u_L — hence β_p^+ is the *global* minimum. But (11) implies that $\chi^p_{(\beta_1^+,\ldots,\beta_{p-1}^+)}(u)$ has two distinct global minima at $u = \beta_p^+$ and $u = \beta_p^+ - \Delta_p^+$ — a contradiction to Assumption 1. Hence $\Delta_p^+ = 0$ and $\beta_p^{n_k-1} - \beta_p^{n_k} \to 0$.

In the above spirit consider sequence (12) (differing from $\boldsymbol{\beta}_{p-1}^{n_k-1}$ at $(p-1)^{\text{th}}$ coordinate)

$$\boldsymbol{\beta}_{p-2}^{n_k-1} := (\beta_1^{n_k}, \ldots, \beta_{p-3}^{n_k}, \beta_{p-2}^{n_k}, \beta_{p-1}^{n_k-1}, \beta_p^{n_k-1}) \tag{12}$$

We will show $\boldsymbol{\beta}_{p-2}^{n_k-1} - \boldsymbol{\beta}_{p-1}^{n_k-1} \to 0$, which along with $\boldsymbol{\beta}_{p-1}^{n_k-1} - \boldsymbol{\beta}^{n_k} \to 0$ (shown above) will imply $\boldsymbol{\beta}_{p-2}^{n_k-1} - \boldsymbol{\beta}^{n_k} \to 0$. The proof of this part is similar to the above.

Write $\boldsymbol{\beta}_{p-1}^{n_k-1} - \boldsymbol{\beta}_{p-2}^{n_k-1} = \boldsymbol{\Delta}_{p-1}^{n_k-1} := (0, \ldots, 0, \Delta_{p-1}^{n_k}, 0)$ and suppose $\Delta_{p-1}^{n_k} \nrightarrow 0$. We will arrive at a contradiction.

By hypothesis, there is a subsequence of $\Delta_{p-1}^{n_k}$ converging to $\Delta_{p-1}^+ \neq 0$. Passing onto a further subsequence, if necessary and using $\boldsymbol{\Delta}_{p-1}^{n_k-1} \to \boldsymbol{\Delta}_{p-1}^+$ we have

$$\big(Q(\boldsymbol{\beta}_{p-1}^{n_k-1} - \boldsymbol{\Delta}_{p-1}^{n_k-1}) - Q(\boldsymbol{\beta}_{p-1}^{n_k-1})\big) \downarrow \big(Q(\boldsymbol{\beta}^+ - \boldsymbol{\Delta}_{p-1}^+) - Q(\boldsymbol{\beta}^+)\big) = 0 \tag{13}$$

This essentially implies that the map $\chi^{p-1}_{(\beta_1^+,\ldots,\beta_{p-2}^+,\beta_p^+)}(u)$ has two distinct global minima at $u = \beta_{p-1}^+$ and $u = \beta_{p-1}^+ - \Delta_{p-1}^+$; which is a contradiction — hence $\Delta_{p-1}^{n_k} \to 0$.

We can follow the above arguments inductively for finitely many steps namely $i = p-3, p-4, \ldots, 1, p$ and show that $\Delta_i^{n_k} \to 0$ for every i. This shows $\boldsymbol{\beta}^{n_k-1} - \boldsymbol{\beta}^{n_k} \to 0$, thereby completing the proof. □

The uniform separation condition — Assumption 2 of Lemma 2 is a trivial consequence of strict convexity of the univariate functions (Theorem 4) and are more general.

1.2 Calibration of MC+ and Monotonicity Properties of $\lambda_S, \gamma\lambda_S$

Proof of Theorem 2 on page 21

Proof. For convenience we rewrite $\lambda_S(\lambda, \gamma)$ in terms of a function $f_\lambda(\gamma)$ as

$$\lambda_S(\lambda, \gamma) = \lambda f_\lambda(\gamma). \tag{14}$$

Using (28) and the equality of effective *df* across all γ for every λ, $f_\lambda(\gamma)$ can be defined as a solution to the following functional equation

$$\frac{\gamma}{\gamma - 1}\Pr(\lambda f_\lambda(\gamma) \leq |\tilde{\beta}| < \gamma\lambda f_\lambda(\gamma)) \quad + \quad \Pr(|\tilde{\beta}| > \gamma\lambda f_\lambda(\gamma)) = \Pr(|\tilde{\beta}| > \lambda) \tag{15}$$

$$\forall \gamma > 1,\ f_\lambda(\gamma) > 1. \tag{16}$$

Since we have assumed $\sigma^2/n = 1$ and a null generative model, we have

$$\frac{\gamma}{\gamma - 1}(\Phi(\gamma\lambda f_\lambda(\gamma)) - \Phi(\lambda f_\lambda(\gamma)) + (1 - \Phi(\gamma\lambda f_\lambda(\gamma)) = (1 - \Phi(\lambda)) \tag{17}$$

or equivalently

$$\Phi(\gamma\lambda f_\lambda(\gamma)) - \gamma\Phi(\lambda f_\lambda(\gamma)) = -(\gamma - 1)\Phi(\lambda) \tag{18}$$

where Φ is the standard normal cdf and ϕ the pdf of the same. Equation (29) is obtained from (18) and (14). □

Proof of Theorem 3 on page 21 Before we begin the proof, we need a lemma.

Lemma 3. *For $\delta \leq \lambda_H$, where $\lambda_H = \lambda_S(\lambda, \gamma = 1+)$ the function $h(\delta)$*

$$h(\delta) = \Phi(\delta) - \Phi(\lambda) - \delta\phi(\delta) \tag{19}$$

satisfies $h(\delta) < 0, \ \forall \lambda_H > \delta > \lambda$

Proof. Observe that for every fixed $\lambda > 0$, $h(\lambda) = -\delta\phi(\delta) < 0$. In addition, the derivative of $h(\cdot)$ satisfies

$$h'(\delta) = -\delta\phi'(\delta) > 0 \tag{20}$$

since $\phi(\delta)$ is the standard normal density and $\delta > 0$.

Also, from equation (17) it follows that (taking $\gamma \to 1+$),

$$\lambda_H\phi(\lambda_H) + 1 - \Phi(\lambda_H) = 1 - \Phi(\lambda)$$

which implies $\Phi(\lambda_H) - \Phi(\lambda) - \lambda_H\phi(\lambda_H) = 0$ that is $h(\lambda_H) = 0$. Since $h(\cdot)$ is strictly monotone increasing on $[\lambda, \lambda_H]$, with $h(\lambda) < 0$ and $h(\lambda_H) = 0$ it is necessarily negative on $[\lambda, \lambda_H)$. This completes the proof. □

Proof. Proof of part [a]. Let $\lambda'_{\max}$ denote the partial derivative of $\lambda_{\max} \equiv \gamma\lambda_S(\lambda, \gamma)$ w.r.t γ.

Differentiating both sides of equation (29) w.r.t γ, we get

$$\lambda'_{\max}(\phi(\gamma\lambda_S) - \phi(\lambda_S)) = \Phi(\lambda_S) - \Phi(\lambda) - \lambda_S\phi(\lambda_S) \tag{21}$$

Observe that $\phi(\gamma\lambda_S) < \phi(\lambda_S)$, since $\gamma\lambda_S > \lambda_S$.

In order to complete the proof it suffices to show that $\Phi(\lambda_S) - \Phi(\lambda) - \lambda_S\phi(\lambda_S) < 0$. This follows from Lemma 3, which gives $h(\lambda_S) < 0$ since $\lambda < \lambda_S < \lambda_H$.

Proof of part [b]. Let $\lambda'_{\min}$ denote the partial derivative of $\lambda_{\min} \equiv \lambda_S(\lambda, \gamma)$ w.r.t γ. Differentiating both sides of equation (29) w.r.t γ, we get

$$\lambda'_{\min}\gamma(\phi(\gamma\lambda_S) - \phi(\lambda_S)) = \Phi(\lambda_S) - \Phi(\lambda) - \lambda_S\phi(\gamma\lambda_S) \tag{22}$$

Since $\phi(\gamma\lambda_S) < \phi(\lambda_S)$, it suffices to show that r.h.s of (22) is greater than zero. Define for a fixed $\lambda > 0$ the function

$$g(\lambda_S, \gamma) = \Phi(\lambda_S) - \phi(\gamma\lambda_S) \tag{23}$$

From (29) it follows that,

$$\Phi(\lambda_S) = \frac{\gamma - 1}{\gamma}\Phi(\lambda) + \Phi(\lambda_S\gamma)/\gamma \tag{24}$$

Using (24), $g(\lambda_S, \gamma)$ becomes

$$g(\lambda_S, \gamma) = \Phi(\lambda) - \frac{\Phi(\lambda)}{\gamma} + \frac{\Phi(\lambda_S\gamma)}{\gamma} - \phi(\lambda_S\gamma)\lambda_S \tag{25}$$

Consider the function $g(\lambda_S, \gamma) - \Phi(\lambda)$, for fixed $\lambda > 0$.

By Theorem 3 (part [a]), we know that for fixed λ, the map $\gamma \mapsto \lambda_{\max} = \gamma\lambda_S$ is increasing in γ. Hence, the function $g(\lambda_S, \gamma) - \Phi(\lambda)$ is greater than zero iff the function $h(u)$ (with $u = \gamma\lambda_S$)

$$h(u) = \Phi(u) - u\phi(u) - \Phi(\lambda)$$

satisfies

$$h(u) > 0 \quad \text{on } u > \lambda_H$$

Observe that $h'(u) = -u\phi'(u) > 0$ on $u > \lambda_H$. In addition, from (17), taking limits over γ it follows that $h(\lambda_H) = 0$.

This proves that $h(u) > 0$ on $u > \lambda_H$. Consequently, $g(\lambda_S, \gamma) > \Phi(\lambda) \ \ \forall\gamma > 1$. This proves that the r.h.s of (22) is greater than zero. □

1.3 Parametric functional form for $\lambda_S(\lambda, \gamma)$

We detail out the steps leading to the expressions (30,31) given in Section 5.1.2.

For every fixed $\lambda > 0$ the following are the list of properties that the function $f_\lambda(\gamma)$ (see Appendix 1.2) should have

1. monotone increase of shrinkage thresholds λ_S from soft to hard thresholding, implies $\lambda f_\lambda(\gamma)$ should be increasing as γ decreases.

2. monotone decrease of bias thresholds $\gamma\lambda_S$ from soft to hard thresholding requires $\gamma\lambda f_\lambda(\gamma)$ to be increasing as γ increases.

3. The sandwiching property:

$$\lambda_S(\gamma) > \lambda \ \forall\gamma \in (1, \infty) \implies f_\lambda(\gamma) > 1 \quad \forall\gamma \in (1, \infty) \tag{26}$$

4. The *df* calibration at the two extremes of the family: $df(\hat{\boldsymbol{\mu}}_{\gamma=1+,\lambda}) = df(\hat{\boldsymbol{\mu}}_{\gamma=\infty,\lambda})$.

The above considerations suggest a nice parametric form for $f_\lambda(\gamma)$ given by

$$f_\lambda(\gamma) = 1 + \tau/\gamma \quad \text{for some} \quad \tau > 0 \tag{27}$$

Let us denote $\lambda_S(\lambda, 1+) = \lambda_H$

Equating the *df*'s for the soft and hard thresholding τ is to be chosen as:

$$(1 - \Phi(\lambda)) = \lambda_H\phi(\lambda_H) + (1 - \Phi(\lambda_H))$$

which implies

$$(1 + \tau)^{-1} = f_\lambda(\gamma = 1+) = \lambda_H^{-1}\Phi^{-1}(\Phi(\lambda_H) - \lambda_H\phi(\lambda_H))$$

Hence we get

$$\lambda_S = \lambda_H \left(\frac{1}{1+\tau} + \frac{\tau}{\gamma(1+\tau)} \right)$$

In addition, observe that as $\gamma \to \infty$, $\lambda_S(\lambda, \gamma) \to \frac{\lambda_H}{1+\tau} = \lambda$. This gives expression (31).

1.3.1 Tightening up the initial approximation of λ_S

In order to get a better approximation of the mapping $(\lambda, \gamma) \mapsto (\lambda_S, \gamma\lambda_S)$, we can define a sequence of recursive updates based on equation (29). Theorem 1 gives a proof of the convergence of these updates and the algorithm for obtaining the better approximations are given in Section (1.3.2).

Theorem 1. *With reference to (29), define a sequence of estimates* $\{z_k\}_{k\geq 0}$ *recursively as*

$$\frac{1}{\gamma-1}\Phi(\gamma z_k) - \frac{\gamma}{\gamma-1}\Phi(z_{k+1}) = -\Phi(\lambda) \tag{28}$$

where the variables $\{z_k\}_{k\geq 0}$ *are actually the iterative updates of* $\lambda_S(\gamma)$ *for a fixed* (λ, γ)*. Under the condition*

$$\sup_k \{\frac{|\phi(\xi_k^*)|}{|\phi(\xi_k)|}\} < 1 \tag{29}$$

where $\xi_k^* \in (\min\{\gamma z_k, \gamma z_{k+1}\}, \max\{\gamma z_k, \gamma z_{k+1}\})$ *and* $\xi_k \in (\min\{z_{k+1}, z_{k+2}\}, \max\{z_{k+1}, z_{k+2}\})$ *for all* $k \geq 0$*, the sequence* $\{z_k\}_{k\geq 0}$ *converges to a fixed point* z^* *of the following equation:*

$$\frac{1}{\gamma-1}\Phi(\gamma z^*) - \frac{\gamma}{\gamma-1}\Phi(z^*) = -\Phi(\lambda) \tag{30}$$

and hence provides a solution to $(\lambda, \gamma) \mapsto (\lambda_S(\gamma) = z^*, \gamma\lambda_S(\gamma) = \gamma z^*)$

Proof. By the definition of the updates z_k, in (28) we get for $k = k, k+1$

$$\frac{1}{\gamma-1}\Phi(\gamma z_k) - \frac{\gamma}{\gamma-1}\Phi(z_{k+1}) = -\Phi(\lambda) \quad k \in \{k, k+1\}$$

Using the above two we get:

$$\Phi(\gamma z_k) - \Phi(\gamma z_{k+1}) = \gamma(\Phi(z_{k+1}) - \Phi(z_{k+2})) \tag{31}$$

Applying the mean-value theorem on both sides of (31) we get:

$$(\gamma z_k - \gamma z_{k+1})\phi(\gamma\eta_k) = \gamma(z_{k+1} - z_{k+2})\phi(\eta_k^*) \tag{32}$$

where η_k lies in between z_k,z_{k+1} and η_k^* lies in between z_{k+1},z_{k+2}. Writing $R_k = \frac{\phi(\gamma\eta_k)}{\phi(\eta_k^*)}$, $\forall k \geq 0$, we observe that

$$|z_{m+1} - z_{m+2}| < C(\sup_k |R_k|)^m$$

for some C which does not depend upon m. Suppose $\sup_k |R_k| < 1$, then we observe that the telescopic sum $\sum_k (z_{k+1} - z_k)$ converges absolutely. Hence in particular the sequence z_k converges.

The condition (29) stated in this theorem, is a sufficient condition for $\sup_k |R_k| < 1$, hence the sequence of updates converge. Observe that, the sufficient condition is very plausible since $\gamma > 1$, and the density $\phi(\cdot)$ decreases on the positive reals.

This completes the proof. □

1.3.2 Algorithm for tightening up the approximation of *shrinkage thresholds*

For a given (λ, γ) the re-parametrized MC+ penalty is of the form given in (26) where $\lambda_S, \gamma\lambda_S$ defined through (29) are obtained by the following steps:

1. Obtain $f_\lambda(\gamma)$ based on the starting proposal function as in (28). This gives an estimate for $\hat{\lambda}_{\min}$ and $\hat{\lambda}_{\max} = \gamma\lambda_S$ using (14).

2. With $z_0 = \hat{\lambda}_{\min}$ obtain recursively the sequence of estimates $\{z_k\}_{k\geq 1}$ using Theorem (1) till convergence criterion is met, that is, $|df(\hat{\boldsymbol{\mu}}_{\gamma,\lambda}) - df(\hat{\mu}_{\gamma=1+}(y,\lambda))| <$ tol.

3. If the sequence $z_k \to z_*$ then assign $\lambda_S = z_*$ and $\gamma\lambda_S = \gamma z_*$

Theorem 1 gives a formal proof of the convergence of the above algorithm.

1.4 Understanding the sub-optimality of LLA through examples

In this section, as a continuation to Section 8 we explore the reasons behind the sub-optimality of LLA through two univariate problem instances.

1.4.1 Example 1: log-penalty and the phase-transition phenomenon

Consider the log-penalized least-squares criterion (10)

$$Q^{(1)}(\beta) = \tfrac{1}{2}(\beta - \tilde{\beta})^2 + \frac{\lambda}{\log(\gamma+1)}\log(\gamma|\beta|+1) \tag{33}$$

for some large value of γ such that $Q^{(1)}(\beta)$ has multiple local minima, but a unique global minimizer. The corresponding (multi-stage) LLA for this problem leads to a sequence $\{\beta^k\}_{k\geq 1}$ defined by

$$\begin{aligned} \beta^{k+1} &= \arg\min_{\beta} \quad \tfrac{1}{2}(\beta - \tilde{\beta})^2 + w^k|\beta| \\ &= \mathrm{S}(\tilde{\beta}, w^k), \quad \text{with } \; w^k = \frac{\lambda\gamma}{(\gamma|\beta^k|+1)\log(\gamma+1)} \end{aligned} \tag{34}$$

Convergence of the sequence of estimates β^k corresponds to fixed points of the function

$$M_{\mathrm{lla}} : \;\; \beta \mapsto \mathrm{sgn}(\tilde{\beta})(|\tilde{\beta}| - \frac{\lambda^*}{(\gamma|\beta|+1)})_+ \quad \text{with} \quad \lambda^* = \frac{\lambda\gamma}{\log(\gamma+1)}. \tag{35}$$

$M_{\mathrm{lla}}(\beta)$ on $\beta \geq 0$, has multiple fixed points for certain choices of $\lambda, \tilde{\beta}$.

For example, let $\gamma = 500$, $\lambda = 74.6$ and $\tilde{\beta} = 10$; here $Q(\beta)$, $\beta \geq 0$ has two minima, with the global minimum at zero. Figure 1[Right] shows the iteration map $M_{\mathrm{lla}}(\beta)$. It has two fixed points, corresponding to the two local minima $0 = \beta_1 < \beta_2$ of $Q(\beta)$, $\beta \geq 0$. There exists a critical value $r > 0$ such that

$$\beta^1 < r \Longrightarrow \beta^k \to 0 = \beta_1, \quad \beta^1 > r \Longrightarrow \beta^k \to \beta_2 \;\text{ as } k \to \infty \tag{36}$$

This *phase transition* phenomenon is responsible for the sub-optimality of the LLA in the one-dimensional case. Note that the coordinate-wise procedure applied in this context will give the global minimum.

The importance of the initial choice in obtaining "optimal" estimates via the re-weighted ℓ_1 minimization has also been pointed out in an *asymptotic* setting by **?**; our observations are non-asymptotic.

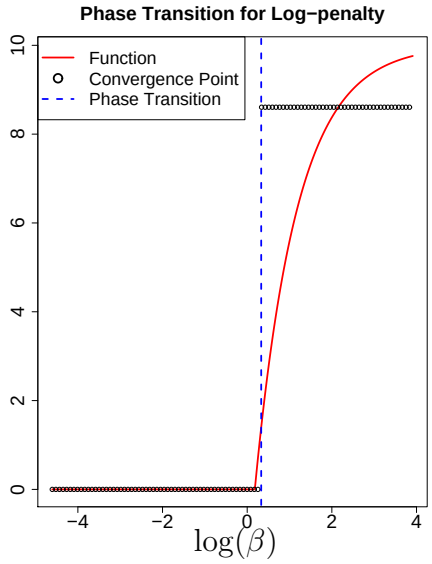

Figure 1: $M_{\text{lla}}(\beta)$ for the univariate log-penalized least squares problem. The starting point β^1 determines which stationary point the LLA sequence will converge to. If the starting point is on the right side of the phase transition line (vertical dotted line), then the LLA will converge to a strictly sub-optimal value of the objective function. In this example, the global minimum of the one-dimensional criterion is at zero.

1.4.2 Example 2: MC+ penalty and rate of convergence

Here we study the rate of convergence to the global minimum for the MC+ penalized least squares problem (16). The multi-stage LLA for this problem gives the following updates

$$\begin{aligned} \beta^{k+1} &= \arg\min_{\beta} \quad \tfrac{1}{2}(\beta - \tilde{\beta})^2 + w^k|\beta| \\ &= \mathrm{S}(\tilde{\beta}, w^k), \quad \text{with} \;\; w^k = \lambda(1 - \frac{|\beta^k|}{\gamma\lambda})_+ \end{aligned} \tag{37}$$

The function $\tilde{\beta} \mapsto \mathrm{S}(\tilde{\beta}, \lambda(1 - \frac{|\beta^k|}{\gamma\lambda})_+)$ need not be a contraction mapping, hence one cannot readily appeal to Banach's Fixed Point Theorem to obtain a proof of the convergence and a bound on its rate. However, in this example we can justify convergence (see Lemma 4)

Lemma 4. *Given $\tilde{\beta}, \gamma > 1, \lambda > 0$ the sequence $\{\beta^{k+1}\}_k$ defined in (37) converges to the minimum of $Q^{(1)}(\beta)$ for the MC+ penalized criterion at a (worst-case) geometric*

rate

$$|\beta^k - \arg\min Q^{(1)}(\beta)| = O(1/\gamma^k)$$

Sketch of proof. Assume, for the sake of simplicity $\tilde{\beta} \geq 0, \beta^k \geq 0$. The update β^{k+1} is given explicitly by

$$\beta^{k+1} = \begin{cases} \tilde{\beta} - \lambda + \frac{1}{\gamma}\beta^k & \text{if } \gamma(\lambda - \tilde{\beta}) \leq \beta^k \leq \lambda\gamma \\ 0 & \text{if } \gamma(\lambda - \tilde{\beta}) > \beta^k \leq \lambda\gamma \\ \tilde{\beta} & \text{if } \beta^k > \lambda\gamma \end{cases}$$

On iteratively expanding a particular case of the above expression, we obtain

$$\beta^{k+1} = (\tilde{\beta} - \lambda)(\sum_{i=0}^{k} \gamma^{-i}) + \gamma^{-k-1}\beta^1, \text{ if } \gamma(\lambda - \tilde{\beta}) \leq \beta^i \leq \lambda\gamma \ \forall, 1 \leq i \leq k \qquad (38)$$

The proof essentially considers different cases based on where β^k lies wrt $\lambda\gamma, \gamma(\lambda - \tilde{\beta})$ and derives expressions for the sequence β^k. The convergence of β^k follows in a fairly straightforward way analyzing these expressions. We omit the details for the sake of brevity.

The following corollary is a consequence of Lemma 4.

Corollary 1. *The number of iterations required for the multi-stage LLA algorithm of the univariate MC+ penalized least squares to converge within an ϵ tolerance of the minimizer of the objective $Q^{(1)}(\boldsymbol{\beta})$ is of the order of $-\frac{\log(\epsilon)}{\log(\gamma)}$. The coordinate-wise algorithm applied in this setting converges in exactly one iteration.*

Section 8.1, in the text describes empirical results and comparisons showing the performance of *SparseNet* as an optimization algorithm. In the spirit of Figure 7, attached below is a figure showing the objective value curves for all the 24 simulation examples.

[Please Include the figure file "perform24.pdf" about here]