Submitted to the Annals of Statistics

SUPPLEMENTARY MATERIAL OF "A NEW PERSPECTIVE ON BOOSTING IN LINEAR REGRESSION VIA SUBGRADIENT OPTIMIZATION AND RELATIVES"

By Robert M. Freund, Paul Grigas and Rahul Mazumder

A.1. Additional Details for Section 1.

A.1.1. *Figure showing training error versus ℓ_1-shrinkage bounds.* Figure A.1 shows profiles of ℓ_1 norm of the regression coefficients versus training error for LS-Boost(ε), FS$_\varepsilon$ and Lasso.

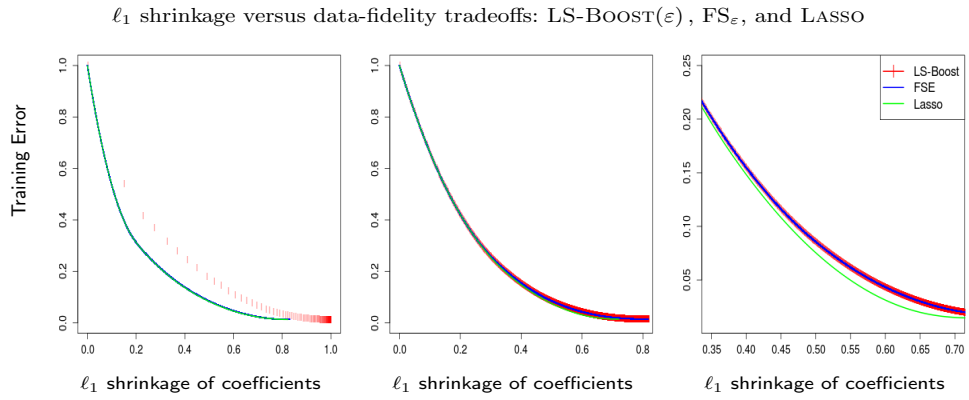

Fig A.1. *Figure showing profiles of ℓ_1 norm of the regression coefficients versus training error for* LS-Boost(ε), *FS$_\varepsilon$ and* Lasso. *[Left panel] Shows profiles for a synthetic dataset where the covariates are drawn from a Gaussian distribution with pairwise correlations $\rho = 0.5$. The true β has ten non-zeros with $\beta_i = 1$ for $i = 1, \ldots, 10$, and SNR $= 1$. Here we ran* LS-Boost(ε) *with $\varepsilon = 1$ and ran FS$_\varepsilon$ with $\varepsilon = 10^{-2}$. The middle (and right) panel profiles corresponds to the Prostate cancer dataset (described in Section 6). Here we ran* LS-Boost(ε) *with $\varepsilon = 0.01$ and we ran FS$_\varepsilon$ with $\varepsilon = 10^{-5}$. The right panel figure is a zoomed-in version of the middle panel in order to highlight the difference in profiles between* LS-Boost(ε), *FS$_\varepsilon$ and* Lasso. *The vertical axes have been normalized so that the training error at $k = 0$ is one, and the horizontal axes have been scaled to the unit interval (to express the ℓ_1-norm of $\hat{\beta}^k$ as a fraction of the maximum).*

A.2. Additional Details for Section 2.

A.2.1. *Properties of Convex Quadratic Functions.* Consider the following quadratic optimization problem (QP) defined as:

$$h^* := \min_{x \in \mathbb{R}^n} h(x) := \tfrac{1}{2} x^T Q x + q^T x + q^o \ ,$$

where Q is a symmetric positive semi-definite matrix, whereby $h(\cdot)$ is a convex function. We assume that $Q \neq 0$, and recall that $\lambda_{\rm pmin}(Q)$ denotes the smallest nonzero (and hence positive) eigenvalue of Q.

PROPOSITION A.2.1. *If $h^* > -\infty$, then for any given x, there exists an optimal solution x^* of (QP) for which*

$$\|x - x^*\|_2 \leq \sqrt{\frac{2(h(x) - h^*)}{\lambda_{\rm pmin}(Q)}} \ .$$

Also, it holds that

$$\|\nabla h(x)\|_2 \geq \sqrt{\frac{\lambda_{\rm pmin}(Q) \cdot (h(x) - h^*)}{2}} \ .$$

Proof: The result is simply manipulation of linear algebra. Let us assume without loss of generality that $q^o = 0$. If $h^* > -\infty$, then (QP) has an optimal solution x^*, and the set of optimal solutions are characterized by the gradient condition

$$0 = \nabla h(x) = Qx + q \ .$$

Now let us write the sparse eigendecomposition of Q as $Q = PDP^T$ where D is a diagonal matrix of non-zero eigenvalues of Q and the columns of P are orthonormal, namely $P^T P = I$. Because (QP) has an optimal solution, the system of equations $Qx = -q$ has a solution, and let $\tilde{x}$ denote any such solution. Direct manipulation establishes:

$$PP^T q = -PP^T Q\tilde{x} = -PP^T PDP^T \tilde{x} = -PDP^T \tilde{x} = -Q\tilde{x} = q \ .$$

Furthermore, let $\hat{x} := -PD^{-1}P^T q$. It is then straightforward to see that $\hat{x}$ is an optimal solution of (QP) since in particular:

$$Q\hat{x} = -PDP^T PD^{-1}P^T q = -PP^T q = -q \ ,$$

and hence

$$h^* = \tfrac{1}{2}\hat{x}^T Q\hat{x} + q^T\hat{x} = -\tfrac{1}{2}\hat{x}^T Q\hat{x} = -\tfrac{1}{2} q^T PD^{-1}P^T PDP^T PD^{-1}P^T q = -\tfrac{1}{2} q^T PD^{-1}P^T q \ .$$

Now let x be given, and define $x^* := [I - PP^T]x - PD^{-1}P^Tq$. Then just as above it is straightforward to establish that $Qx^* = -q$ whereby x^* is an optimal solution. Furthermore, it holds that:

$$\begin{aligned}
\|x - x^*\|_2^2 &= (q^TPD^{-1} + x^TP)P^TP(D^{-1}P^Tq + P^Tx) \\
&= (q^TPD^{-\frac{1}{2}} + x^TPD^{\frac{1}{2}})D^{-1}(D^{-\frac{1}{2}}P^Tq + D^{\frac{1}{2}}P^Tx) \\
&\leq \tfrac{1}{\lambda_{\mathrm{pmin}}(Q)}(q^TPD^{-\frac{1}{2}} + x^TPD^{\frac{1}{2}})(D^{-\frac{1}{2}}P^Tq + D^{\frac{1}{2}}P^Tx) \\
&= \tfrac{1}{\lambda_{\mathrm{pmin}}(Q)}(q^TPD^{-1}P^Tq + x^TPDP^Tx + 2x^TPP^Tq) \\
&= \tfrac{1}{\lambda_{\mathrm{pmin}}(Q)}(-2h^* + x^TQx + 2x^Tq) \\
&= \tfrac{2}{\lambda_{\mathrm{pmin}}(Q)}(h(x) - h^*) \ ,
\end{aligned}$$

and taking square roots establishes the first inequality of the proposition.

Using the gradient inequality for convex functions, it holds that:

$$\begin{aligned}
h^* = h(x^*) &\geq h(x) + \nabla h(x)^T(x^* - x) \\
&\geq h(x) - \|\nabla h(x)\|_2\|x^* - x\|_2 \\
&\geq h(x) - \|\nabla h(x)\|_2\sqrt{\tfrac{2(h(x)-h^*)}{\lambda_{\mathrm{pmin}}(Q)}} \ ,
\end{aligned}$$

and rearranging the above proves the second inequality of the proposition. □

A.2.2. *Proof of Theorem 2.1.* We first prove part *(i)*. Utilizing (2.9), which states that $\hat{r}^{k+1} = \hat{r}^k - \varepsilon\left((\hat{r}^k)^T\mathbf{X}_{j_k}\right)\mathbf{X}_{j_k}$, we have:

$$\begin{aligned}
L_n(\hat{\beta}^{k+1}) &= \tfrac{1}{2n}\|\hat{r}^{k+1}\|_2^2 \\
&= \tfrac{1}{2n}\|\hat{r}^k - \varepsilon\left((\hat{r}^k)^T\mathbf{X}_{j_k}\right)\mathbf{X}_{j_k}\|_2^2 \\
&= \tfrac{1}{2n}\|\hat{r}^k\|_2^2 - \tfrac{1}{n}\varepsilon\left((\hat{r}^k)^T\mathbf{X}_{j_k}\right)^2 + \tfrac{1}{2n}\varepsilon^2\left((\hat{r}^k)^T\mathbf{X}_{j_k}\right)^2 \qquad \text{(A.2.1)} \\
&= L_n(\hat{\beta}^k) - \tfrac{1}{2n}\varepsilon(2-\varepsilon)\left((\hat{r}^k)^T\mathbf{X}_{j_k}\right)^2 \\
&= L_n(\hat{\beta}^k) - \tfrac{1}{2n}\varepsilon(2-\varepsilon)n^2\|\nabla L_n(\hat{\beta}^k)\|_\infty^2 \ ,
\end{aligned}$$

(where the last equality above uses (2.8)), which yields:

$$L_n(\hat{\beta}^{k+1}) - L_n^* = L_n(\hat{\beta}^k) - L_n^* - \tfrac{n}{2}\varepsilon(2-\varepsilon)\|\nabla L_n(\hat{\beta}^k)\|_\infty^2 \ . \tag{A.2.2}$$

We next seek to bound the right-most term above. We will do this by invoking Proposition A.2.1, which presents two important properties of convex quadratic functions. Because $L_n(\cdot)$ is a convex quadratic function of the same format as Proposition A.2.1 with $h(\cdot) \leftarrow L_n(\cdot)$, $Q \leftarrow \frac{1}{n}\mathbf{X}^T\mathbf{X}$, and $h^* \leftarrow L_n^*$, it follows from the second property of Proposition A.2.1 that

$$\|\nabla L_n(\beta)\|_2 \geq \sqrt{\frac{\lambda_{\text{pmin}}(\frac{1}{n}\mathbf{X}^T\mathbf{X})(L_n(\beta) - L_n^*)}{2}} = \sqrt{\frac{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})(L_n(\beta) - L_n^*)}{2n}} \ .$$

Therefore

$$\|\nabla L_n(\beta)\|_\infty^2 \geq \tfrac{1}{p}\|\nabla L_n(\beta)\|_2^2 \geq \frac{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})(L_n(\beta) - L_n^*)}{2np} \ .$$

Substituting this inequality into (A.2.2) yields after rearranging:

$$L_n(\hat{\beta}^{k+1}) - L_n^* \leq (L_n(\hat{\beta}^k) - L_n^*)\left(1 - \frac{\varepsilon(2-\varepsilon)\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}{4p}\right) = (L_n(\hat{\beta}^k) - L_n^*)\cdot\gamma \ . \tag{A.2.3}$$

Now note that $L_n(\hat{\beta}^0) = L_n(0) = \frac{1}{2n}\|\mathbf{y}\|_2^2$ and

$$\begin{aligned} L_n(\hat{\beta}^0) - L_n^* =& \tfrac{1}{2n}\|\mathbf{y}\|_2^2 - \tfrac{1}{2n}\|\mathbf{y} - \mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2 \\ =& \tfrac{1}{2n}\|\mathbf{y}\|_2^2 - \tfrac{1}{2n}(\|\mathbf{y}\|_2^2 - 2\mathbf{y}^T\mathbf{X}\hat{\beta}_{\text{LS}} + \|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2) \\ =& \tfrac{1}{2n}\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2 \ , \end{aligned}$$

where the last equality uses the normal equations (2.3). Then *(i)* follows by using elementary induction and combining the above with (A.2.3):

$$L_n(\hat{\beta}^k) - L_n^* \leq (L_n(\hat{\beta}^0) - L_n^*)\cdot\gamma^k = \tfrac{1}{2n}\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2 \cdot \gamma^k \ .$$

To prove *(ii)*, we invoke the first inequality of Proposition A.2.1, which in this context states that

$$\|\hat{\beta}^k - \hat{\beta}_{\text{LS}}\|_2 \leq \frac{\sqrt{2(L_n(\hat{\beta}^k) - L_n^*)}}{\sqrt{\lambda_{\text{pmin}}(\frac{1}{n}\mathbf{X}^T\mathbf{X})}} = \frac{\sqrt{2n(L_n(\hat{\beta}^k) - L_n^*)}}{\sqrt{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}} \ .$$

Part *(ii)* then follows by substituting the bound on $(L_n(\hat{\beta}^k) - L_n^*)$ from *(i)* and simplifying terms. Similarly, the proof of *(iii)* follows from the observation that $\|\mathbf{X}\hat{\beta}^k - \mathbf{X}\hat{\beta}_{\text{LS}}\|_2 = \sqrt{2n(L_n(\hat{\beta}^k) - L_n^*)}$ and then substituting the bound on $(L_n(\hat{\beta}^k) - L_n^*)$ from *(i)* and simplifying terms.

To prove *(iv)*, define the point $\tilde{\beta}^k := \hat{\beta}^k + \tilde{u}_{j_k} e_{j_k}$. Then using similar arithmetic as in (A.2.1) one obtains:

$$L_n^* \leq L_n(\tilde{\beta}^k) = L_n(\hat{\beta}^k) - \tfrac{1}{2n}\tilde{u}_{j_k}^2 \ ,$$

where we recall that $\tilde{u}_{j_k} = (\hat{r}^k)^T \mathbf{X}_{j_k}$. This inequality then rearranges to

$$|\tilde{u}_{j_k}| \ \leq \ \sqrt{2n(L_n(\hat{\beta}^k) - L_n^*)} \ \leq \ \|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2 \cdot \gamma^{k/2} \ , \tag{A.2.4}$$

where the second inequality follows by substituting the bound on $(L_n(\hat{\beta}^i) - L_n^*)$ from *(i)*. Recalling (2.4) and (2.8), the above is exactly part *(iv)*.

Part *(v)* presents two distinct bounds on $\|\hat{\beta}^k\|_1$, which we prove independently. To prove the first bound, let $\hat{\beta}_{\text{LS}}$ be any least-squares solution, which therefore satisfies (2.3). It is then elementary to derive using similar manipulation as in (A.2.1) that for all i the following holds:

$$\|\mathbf{X}(\hat{\beta}^{i+1} - \hat{\beta}_{\text{LS}})\|_2^2 = \|\mathbf{X}(\hat{\beta}^i - \hat{\beta}_{\text{LS}})\|_2^2 - (2\varepsilon - \varepsilon^2)\tilde{u}_{j_i}^2 \tag{A.2.5}$$

which implies that

$$(2\varepsilon - \varepsilon^2)\sum_{i=0}^{k-1} \tilde{u}_{j_i}^2 = \|\mathbf{X}(\hat{\beta}^0 - \hat{\beta}_{\text{LS}})\|_2^2 - \|\mathbf{X}(\hat{\beta}^k - \hat{\beta}_{\text{LS}})\|_2^2 = \|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2 - \|\mathbf{X}(\hat{\beta}^k - \hat{\beta}_{\text{LS}})\|_2^2 \ . \tag{A.2.6}$$

Then note that

$$\begin{aligned} \|\hat{\beta}^k\|_1 \leq & \|(\varepsilon\tilde{u}_{j_0}, \ldots, \varepsilon\tilde{u}_{j_{k-1}})\|_1 \\ \leq & \sqrt{k}\varepsilon\|(\tilde{u}_{j_0}, \ldots, \tilde{u}_{j_{k-1}})\|_2 \\ = & \sqrt{k}\sqrt{\tfrac{\varepsilon}{2-\varepsilon}}\sqrt{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2 - \|\mathbf{X}\hat{\beta}_{\text{LS}} - \mathbf{X}\hat{\beta}^k\|_2^2} \ , \end{aligned} \tag{A.2.7}$$

where the last equality is from (A.2.6).

To prove the second bound in *(v)*, noting that $\hat{\beta}^k = \sum_{i=0}^{k-1} \varepsilon\tilde{u}_{j_i} e_{j_i}$, we bound $\|\hat{\beta}^k\|_1$ as follows:

$$\begin{aligned} \|\hat{\beta}^k\|_1 \leq \varepsilon \sum_{i=0}^{k-1} |\tilde{u}_{j_i}| \ \leq \ & \varepsilon\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2 \sum_{i=0}^{k-1} \gamma^{i/2} \\ = & \frac{\varepsilon\|\mathbf{X}\hat{\beta}_{LS}\|_2}{1 - \sqrt{\gamma}} \left(1 - \gamma^{k/2}\right) \ , \end{aligned}$$

where the second inequality uses (A.2.4) for each $i \in \{0, \ldots, k-1\}$ and the final equality is a geometric series, which completes the proof of *(v)*. Part *(vi)* is simply the property of LS-Boost(ε) that derives from the fact that $\hat{\beta}^0 := 0$ and at every iteration at most one coordinate of β changes status from a zero to a non-zero value. □

A.2.3. *Additional properties of* LS-Boost(ε). We present two other interesting properties of the LS-Boost(ε) algorithm, namely an additional bound on the correlation between residuals and predictors, and a bound on the ℓ_2-shrinkage of the regression coefficients. Both are presented in the following proposition.

Proposition A.2.2. **(Two additional properties of LS-Boost(ε))** *Consider the iterates of the* LS-Boost(ε) *algorithm after k iterations and consider the linear convergence rate coefficient γ:*

$$\gamma := \left(1 - \frac{\varepsilon(2-\varepsilon)\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}{4p}\right).$$

(i) There exists an index $i \in \{0, \ldots, k\}$ for which the ℓ_∞ norm of the gradient vector of the least squares loss function evaluated at $\hat{\beta}^i$ satisfies:

$$\begin{aligned} & \|\nabla L_n(\hat{\beta}^i)\|_\infty \\ = \; & \tfrac{1}{n}\|\mathbf{X}^T\hat{r}^i\|_\infty \\ \leq \; & \min\left\{ \frac{\sqrt{\|\mathbf{X}\hat{\beta}_{LS}\|_2^2 - \|\mathbf{X}\hat{\beta}_{LS} - \mathbf{X}\hat{\beta}^{k+1}\|_2^2}}{n\sqrt{\varepsilon(2-\varepsilon)(k+1)}} \, , \tfrac{1}{n}\|\mathbf{X}\hat{\beta}_{LS}\|_2 \cdot \gamma^{k/2} \right\} . \end{aligned} \tag{A.2.8}$$

(ii) Let J_ℓ denote the number of iterations of LS-Boost(ε) *, among the first k iterations, where the algorithm takes a step in coordinate ℓ, for $\ell = 1, \ldots, p$, and let $J_{\max} := \max\{J_1, \ldots, J_p\}$. Then the following bound on the shrinkage of $\hat{\beta}^k$ holds:*

$$\|\hat{\beta}^k\|_2 \leq \sqrt{J_{\max}}\sqrt{\frac{\varepsilon}{2-\varepsilon}}\sqrt{\|\mathbf{X}\hat{\beta}_{LS}\|_2^2 - \|\mathbf{X}\hat{\beta}_{LS} - \mathbf{X}\hat{\beta}^k\|_2^2} \, . \tag{A.2.9}$$

□

Proof. We first prove part *(i)*. The first equality of (A.2.8) is a restatement of (2.8). For each $i \in \{0, \ldots, k\}$, recall that $\tilde{u}_{j_i} = (\hat{r}^i)^T\mathbf{X}_{j_i}$ and that $|\tilde{u}_{j_i}| = |(\hat{r}^i)^T\mathbf{X}_{j_i}| = \|\mathbf{X}^T\hat{r}^i\|_\infty$, from (2.8). Therefore:

$$\left(\min_{i\in\{0,\ldots,k\}} |\tilde{u}_{j_i}|\right)^2 = \min_{i\in\{0,\ldots,k\}} \tilde{u}_{j_i}^2 \leq \frac{1}{k+1}\sum_{i=0}^{k}\tilde{u}_{j_i}^2 \leq \frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2 - \|\mathbf{X}(\hat{\beta}^{k+1} - \hat{\beta}_{\text{LS}})\|_2^2}{\varepsilon(2-\varepsilon)(k+1)} \, , \tag{A.2.10}$$

where the final inequality follows from (A.2.6) in the proof of Theorem 2.1. Now letting i be an index achieving the minimum in the left hand side of

the above and taking square roots implies that

$$\|\mathbf{X}^T\hat{r}^i\|_\infty = |\tilde{u}_{j_i}| \leq \frac{\sqrt{\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2^2 - \|\mathbf{X}\hat{\beta}_{\mathrm{LS}} - \mathbf{X}\hat{\beta}^{k+1}\|_2^2}}{\sqrt{\varepsilon(2-\varepsilon)(k+1)}} \ ,$$

which is equivalent to the inequality in (A.2.8) for the first right-most term therein. Directly applying (A.2.4) from the proof of Theorem 2.1 and using the fact that i is an index achieving the minimum in the left hand side of (A.2.10) yields:

$$\|\mathbf{X}^T\hat{r}^i\|_\infty = |\tilde{u}_{j_i}| \leq |\tilde{u}_{j_k}| \leq \|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2 \left(1 - \frac{\varepsilon(2-\varepsilon)\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})}{4p}\right)^{k/2} \ ,$$

which is equivalent to the inequality in (A.2.8) for the second right-most term therein.

We now prove part *(ii)*. For fixed $k > 0$, let $\mathcal{J}(\ell)$ denote the set of iteration counters where LS-Boost(ε) modifies coordinate ℓ of β, namely

$$\mathcal{J}(\ell) := \{ i : i < k \text{ and } j_i = \ell \text{ in Step (2.) of Algorithm LS-Boost}(\varepsilon) \} \ ,$$

for $\ell = 1, \ldots, p$. Then $J_\ell = |\mathcal{J}(\ell)|$, and the sets $\mathcal{J}(1), \ldots, \mathcal{J}(p)$ partition the iteration index set $\{0, 1, \ldots, k-1\}$. We have:

(A.2.11)

$$\begin{aligned}
\|\hat{\beta}^k\|_2 &\leq \|(\textstyle\sum_{i\in\mathcal{J}(1)} \varepsilon\tilde{u}_{j_i}, \ldots, \sum_{i\in\mathcal{J}(p)} \varepsilon\tilde{u}_{j_i})\|_2 \\
&\leq \left\| \left(\sqrt{J(1)}\sqrt{\textstyle\sum_{i\in\mathcal{J}(1)} \varepsilon^2\tilde{u}_{j_i}^2}, \ldots, \sqrt{J(p)}\sqrt{\textstyle\sum_{i\in\mathcal{J}(p)} \varepsilon^2\tilde{u}_{j_i}^2} \right) \right\|_2 \\
&\leq \varepsilon\sqrt{J_{\max}} \left\| \left(\sqrt{\textstyle\sum_{i\in\mathcal{J}(1)} \tilde{u}_{j_i}^2}, \ldots, \sqrt{\textstyle\sum_{i\in\mathcal{J}(p)} \tilde{u}_{j_i}^2} \right) \right\|_2 \\
&= \varepsilon\sqrt{J_{\max}} \sqrt{\left(\tilde{u}_{j_0}^2 + \ldots + \tilde{u}_{j_{k-1}}^2 \right)} \ ,
\end{aligned}$$

and the proof is completed by applying inequality (A.2.6). □

Part *(i)* of Proposition A.2.2 describes the behavior of the gradient of the least squares loss function — indeed, recall that the dynamics of the gradient are closely linked to that of the LS-Boost(ε) algorithm and, in particular, to the evolution of the loss function values. To illustrate this connection, let us recall two simple characteristics of the LS-Boost(ε) algorithm:

$$\begin{aligned}
L_n(\hat{\beta}^k) - L_n(\hat{\beta}^{k+1}) &= \tfrac{n}{2}\varepsilon(2-\varepsilon)\|\nabla L_n(\hat{\beta}^k)\|_\infty^2 \\
\hat{r}^{k+1} - \hat{r}^k &= -\varepsilon \left((\hat{r}^k)^T \mathbf{X}_{j_k} \right) \mathbf{X}_{j_k} \ ,
\end{aligned}$$

which follow from (A.2.2) and Step (3.) of the LS-Boost(ε) algorithm, respectively. The above updates of the LS-Boost(ε) algorithm clearly show that smaller values of the ℓ_∞ norm of the gradient slows down the "progress" of the residuals and thus the overall algorithm. Larger values of the norm of the gradient, on the other hand, lead to rapid "progress" in the algorithm. Here, we use the term "progress" to measure the amount of decrease in training error and the norm of the changes in successive residuals. Informally speaking, the LS-Boost(ε) algorithm operationally works towards minimizing the unregularized least squares loss function — and the gradient of the least squares loss function is simultaneously shrunk towards zero. Equation (A.2.8) precisely quantifies the rate at which the ℓ_∞ norm of the gradient converges to zero. Observe that the bound is a minimum of two distinct rates: one which decays as $O(\frac{1}{\sqrt{k}})$ and another which is a linear rate of convergence with parameter $\sqrt{\gamma}$. This is similar to item *(v)* of Theorem 2.1. For small values of k the first rate will dominate, until a point is reached where the linear rate begins to dominate. Note that the dependence on the linear rate γ suggests that for large values of correlations among the samples, the gradient decays slower than for smaller pairwise correlations among the samples.

The behavior of the LS-Boost(ε) algorithm described above should be contrasted with the FS$_\varepsilon$ algorithm. In view of Step (3.) of the FS$_\varepsilon$ algorithm, the successive differences of the residuals in FS$_\varepsilon$ are indifferent to the magnitude of the gradient of the least squares loss function — as long as the gradient is non-zero, then for FS$_\varepsilon$ it holds that $\|\hat{r}^{k+1} - \hat{r}^k\|_2 = \varepsilon$. Thus FS$_\varepsilon$ undergoes a more erratic evolution, unlike LS-Boost(ε) where the convergence of the residuals is much more "smooth."

A.2.4. *Concentration Results for $\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})$ in the High-dimensional Case.*

Proposition A.2.3. *Suppose that $p > n$, let $\tilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ be a random matrix whose entries are i.i.d. standard normal random variables, and define $\mathbf{X} := \frac{1}{\sqrt{n}}\tilde{\mathbf{X}}$. Then it holds that:*

$$\mathbb{E}[\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})] \geq \frac{1}{n}\left(\sqrt{p} - \sqrt{n}\right)^2 \ .$$

Furthermore, for every $t \in [0, \sqrt{p}-\sqrt{n}]$, with probability at least $1-2\exp(-t^2/2)$ it holds that:

$$\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X}) \geq \frac{1}{n}\left(\sqrt{p} - \sqrt{n} - t\right)^2 \ .$$

PROOF. Let $\sigma_1(\tilde{\mathbf{X}}^T) \geq \sigma_2(\tilde{\mathbf{X}}^T) \geq \ldots \geq \sigma_n(\tilde{\mathbf{X}}^T)$ denote the ordered singular values of $\tilde{\mathbf{X}}^T$ (equivalently of $\tilde{\mathbf{X}}$). Then, Theorem 5.32 of [51] states that:

$$\mathbb{E}[\sigma_n(\tilde{\mathbf{X}}^T)] \geq \sqrt{p} - \sqrt{n} \ ,$$

which thus implies:

$$\begin{aligned}
\mathbb{E}[\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})] =& \mathbb{E}[(\sigma_n(\mathbf{X}^T))^2] \\
\geq& (\mathbb{E}[\sigma_n(\mathbf{X}^T)])^2 \\
=& \frac{1}{n}(\mathbb{E}[\sigma_n(\tilde{\mathbf{X}}^T)])^2 \\
\geq& \frac{1}{n}\left(\sqrt{p} - \sqrt{n}\right)^2 \ ,
\end{aligned}$$

where the first inequality is Jensen's inequality.

Corollary 5.35 of [51] states that for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$ it holds that:

$$\sigma_n(\tilde{\mathbf{X}}^T) \geq \sqrt{p} - \sqrt{n} - t \ ,$$

which implies for $t \in [0, \sqrt{p} - \sqrt{n}]$ that:

$$\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X}) = (\sigma_n(\mathbf{X}^T))^2 = \frac{1}{n}(\sigma_n(\tilde{\mathbf{X}}^T))^2 \geq \frac{1}{n}\left(\sqrt{p} - \sqrt{n} - t\right)^2 \ .$$

□

Note that in practice we standardize the model matrix $\mathbf{X}$ so that its columns have unit ℓ_2 norm. Supposing that the entries of $\mathbf{X}$ did originate from an i.i.d. standard normal matrix $\tilde{\mathbf{X}}$, standardizing the columns of $\tilde{\mathbf{X}}$ is not equivalent to setting $\mathbf{X} := \frac{1}{\sqrt{n}}\tilde{\mathbf{X}}$. But, for large enough n, standardizing is a valid approximation to normalizing by $\frac{1}{\sqrt{n}}$, i.e., $\mathbf{X} \approx \frac{1}{\sqrt{n}}\tilde{\mathbf{X}}$, and we may thus apply the above results.

Additional Details for Section 3.

A.2.5. *An Elementary Sequence Process Result, and a Proof of Proposition 3.1.* Consider the following elementary sequence process: $x^0 \in \mathbb{R}^n$ is given, and $x^{i+1} \leftarrow x^i - \alpha_i g^i$ for all $i \geq 0$, where $g^i \in \mathbb{R}^n$ and α_i is a nonnegative scalar, for all i. For this process there are *no* assumptions on how the vectors g^i might be generated.

PROPOSITION A.2.4. *For the elementary sequence process described above, suppose that the $\{g^i\}$ are uniformly bounded, namely $\|g^i\|_2 \leq G$ for all $i \geq 0$. Then for all $k \geq 0$ and for any $x \in \mathrm{R}^n$ it holds that:*

$$\text{(A.2.12)} \qquad \frac{1}{\sum_{i=0}^k \alpha_i} \sum_{i=0}^k \alpha_i (g^i)^T (x^i - x) \quad \leq \quad \frac{\|x^0 - x\|_2^2 + G^2 \sum_{i=0}^k \alpha_i^2}{2 \sum_{i=0}^k \alpha_i} \ .$$

Indeed, in the case when $\alpha_i = \varepsilon$ for all i, it holds that:

$$\text{(A.2.13)} \qquad \frac{1}{k+1} \sum_{i=0}^k (g^i)^T (x^i - x) \quad \leq \quad \frac{\|x^0 - x\|_2^2}{2(k+1)\varepsilon} + \frac{G^2 \varepsilon}{2} \ .$$

PROOF. Elementary arithmetic yields the following:

$$\begin{aligned} \|x^{i+1} - x\|_2^2 &= \|x^i - \alpha_i g^i - x\|_2^2 \\ &= \|x^i - x\|_2^2 + \alpha_i^2 \|g^i\|_2^2 + 2\alpha_i (g^i)^T (x - x^i) \\ &\leq \|x^i - x\|_2^2 + G^2 \alpha_i^2 + 2\alpha_i (g^i)^T (x - x^i) \ . \end{aligned}$$

Rearranging and summing these inequalities for $i = 0, \ldots, k$ then yields:

$$2 \sum_{i=0}^k \alpha_i (g^i)^T (x^i - x) \leq G^2 \sum_{i=0}^k \alpha_i^2 + \|x^0 - x\|_2^2 - \|x^{k+1} - x\|_2^2 \leq G^2 \sum_{i=0}^k \alpha_i^2 + \|x^0 - x\|_2^2 \, ,$$

which then rearranges to yield (A.2.12). (A.2.13) follows from (A.2.12) by direct substitution. □

Proof of Proposition 3.1: Consider the subgradient descent method (3.5) with arbitrary step-sizes α_i for all i. We will prove the following inequality:

$$\text{(A.2.14)} \qquad \min_{i \in \{0, \ldots, k\}} f(x^i) \quad \leq \quad f^* \ + \ \frac{\|x^0 - x^*\|_2^2 + G^2 \sum_{i=0}^k \alpha_i^2}{2 \sum_{i=0}^k \alpha_i} \ ,$$

from which the proof of Proposition 3.1 follows by substituting $\alpha_i = \alpha$ for all i and simplifying terms. Let us now prove (A.2.14). The subgradient descent method (3.5) is applied to instances of problem (3.3) where $f(\cdot)$ is convex,

and where g^i is a subgradient of $f(\cdot)$ at x^i, for all i. If x^* is an optimal solution of (3.3), it therefore holds from the subgradient inequality that

$$f^* = f(x^*) \geq f(x^i) + (g^i)^T(x^* - x^i) \ .$$

Substituting this inequality in (A.2.12) for the value of $x = x^*$ yields:

$$\begin{aligned}\frac{\|x^0 - x^*\|_2^2 + G^2 \sum_{i=0}^k \alpha_i^2}{2\sum_{i=0}^k \alpha_i} &\geq \frac{1}{\sum_{i=0}^k \alpha_i} \sum_{i=0}^k \alpha_i (g^i)^T (x^i - x^*) \\ &\geq \frac{1}{\sum_{i=0}^k \alpha_i} \sum_{i=0}^k \alpha_i (f(x^i) - f^*) \geq \min_{i \in \{0,\ldots,k\}} f(x^i) - f^* \ .\end{aligned}$$

□

A.2.6. *Proof of Theorem 3.1.* We first prove part *(i)*. Note that item *(i)* of Proposition 3.2 shows that FS_ε is a specific instance of subgradient descent to solve problem (3.7), using the constant step-size ε. Therefore we can apply the computational guarantees associated with the subgradient descent method, particularly Proposition 3.1, to the FS_ε algorithm. Examining Proposition 3.1, we need to work out the corresponding values of f^*, $\|x^0 - x^*\|_2$, α, and G in the context of FS_ε for solving the CM problem (3.7). Note that $f^* = 0$ for problem (3.7). We bound the distance from the initial residuals to the optimal least-squares residuals as follows:

$$\|\hat{r}^0 - r^*\|_2 = \|\hat{r}^0 - \hat{r}_{LS}\|_2 = \|\mathbf{y} - (\mathbf{y} - \mathbf{X}\hat{\beta}_{\text{LS}})\|_2 = \|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2 \ .$$

From Proposition 3.2 part *(i)* we have $\alpha = \varepsilon$. Last of all, we need to determine an upper bound G on the norms of subgradients. We have:

$$\|g^k\|_2 = \|\text{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k})\mathbf{X}_{j_k}\|_2 = \|\mathbf{X}_{j_k}\|_2 = 1 \ ,$$

since the covariates have been standardized, so we can set $G = 1$. Now suppose algorithm FS_ε is run for k iterations. Proposition 3.1 then implies that:

(A.2.15)

$$\min_{i \in \{0,\ldots,k\}} \|\mathbf{X}^T \hat{r}^i\|_\infty = \min_{i \in \{0,\ldots,k\}} f(\hat{r}^i) \leq f^* + \frac{\|\hat{r}^0 - r^*\|_2^2}{2\alpha(k+1)} + \frac{\alpha G^2}{2} = \frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2}{2\varepsilon(k+1)} + \frac{\varepsilon}{2} \ .$$

The above inequality provides a bound on the best (among the first k residual iterates) empirical correlation between the residuals $\hat{r}^i$ and each predictor variable, where the bound depends explicitly on the learning rate ε and the

number of iterations k. Furthermore, invoking (2.4), the above inequality implies the following upper bound on the norm of the gradient of the least squares loss $L_n(\cdot)$ for the model iterates $\{\hat{\beta}^i\}$ generated by FS_ε:

$$\min_{i\in\{0,\ldots,k\}} \|\nabla L_n(\hat{\beta}^i)\|_\infty \leq \frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2}{2n\varepsilon(k+1)} + \frac{\varepsilon}{2n} \ . \tag{A.2.16}$$

Let i be the index where the minimum is attained on the left side of the above inequality. In a similar vein as in the analysis in Section 2, we now use Proposition A.2.1 which presents two important properties of convex quadratic functions. Because $L_n(\cdot)$ is a convex quadratic function of the same format as Proposition A.2.1 with $h(\cdot) \leftarrow L_n(\cdot)$, $Q \leftarrow \frac{1}{n}\mathbf{X}^T\mathbf{X}$, and $h^* \leftarrow L_n^*$, it follows from the second property of Proposition A.2.1 that

$$\|\nabla L_n(\hat{\beta}^i)\|_2 \geq \sqrt{\frac{\lambda_{\text{pmin}}(\frac{1}{n}\mathbf{X}^T\mathbf{X})(L_n(\hat{\beta}^i) - L_n^*)}{2}} = \sqrt{\frac{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})(L_n(\hat{\beta}^i) - L_n^*)}{2n}} \ ,$$

where recall that $\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})$ denotes the smallest non-zero (hence positive) eigenvalue of $\mathbf{X}^T\mathbf{X}$. Therefore

$$\|\nabla L_n(\hat{\beta}^i)\|_\infty^2 \geq \tfrac{1}{p}\|\nabla L_n(\hat{\beta}^i)\|_2^2 \geq \frac{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})(L_n(\hat{\beta}^i) - L_n^*)}{2np} \ .$$

Substituting this inequality into (A.2.16) for the index i where the minimum is attained yields after rearranging:

$$L_n(\hat{\beta}^i) - L_n^* \leq \frac{p}{2n\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}\left[\frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2}{\varepsilon(k+1)} + \varepsilon\right]^2 \ , \tag{A.2.17}$$

which proves part *(i)*. The proof of part *(ii)* follows by noting from the first inequality of Proposition A.2.1 that there exists a least-squares solution $\hat{\beta}^*$ for which:

$$\begin{aligned}
\|\hat{\beta}^* - \hat{\beta}^i\|_2 \leq & \sqrt{\frac{2(L_n(\hat{\beta}^i) - L_n^*)}{\lambda_{\text{pmin}}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)}} \\
= & \sqrt{\frac{2n(L_n(\hat{\beta}^i) - L_n^*)}{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}} \\
\leq & \frac{\sqrt{p}}{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}\left[\frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2}{\varepsilon(k+1)} + \varepsilon\right] \ ,
\end{aligned}$$

where the second inequality in the above chain follows using (A.2.17). The proof of part *(iii)* follows by first observing that $\|\mathbf{X}(\hat{\beta}^i - \hat{\beta}_{\text{LS}})\|_2 = \sqrt{2n(L_n(\hat{\beta}^i) - L_n^*)}$ and then substituting the bound on $(L_n(\hat{\beta}^i) - L_n^*)$ from part *(i)* and simplifying terms. Part *(iv)* is a restatement of inequality (A.2.15). Finally, parts *(v)* and *(vi)* are simple and well-known structural properties of FS_ε that are re-stated here for completeness. □

A.2.7. *A deeper investigation of the computational guarantees for* LS-Boost(ε) *and* FS_ε. Here we show that, at least in theory, LS-Boost(ε) is much more efficient than FS_ε if the primary goal is to obtain a model with a certain (pre-specified) data-fidelity. To formalize this notion, we consider a parameter $\tau \in (0, 1]$. We say that $\bar{\beta}$ is at a τ-relative distance to the least squares predictions if $\bar{\beta}$ satisfies:

$$\|\mathbf{X}\bar{\beta} - \mathbf{X}\hat{\beta}_{\text{LS}}\|_2 \leq \tau \|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2 \ . \tag{A.2.18}$$

Now let us pose the following question: if both LS-Boost(ε) and FS_ε are allowed to run with an appropriately chosen learning rate ε for each algorithm, which algorithm will satisfy (A.2.18) in fewer iterations? We will answer this question by studying closely the computational guarantees of Theorems 2.1 and 3.1. Since our primary goal is to compute $\bar{\beta}$ satisfying (A.2.18), we may optimize the learning rate ε, for each algorithm, to achieve this goal with the smallest number of boosting iterations.

Let us first study LS-Boost(ε) . As we have seen, a learning rate of $\varepsilon = 1$ achieves the fastest rate of linear convergence for LS-Boost(ε) and is thus optimal with regard to the bound in part *(iii)* of Theorem 2.1. If we run LS-Boost(ε) with $\varepsilon = 1$ for $k^{\text{LS-Boost}(\varepsilon)} := \left\lceil \frac{4p}{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})} \ln\left(\frac{1}{\tau^2}\right) \right\rceil$ iterations, then it follows from part *(iii)* of Theorem 2.1 that we achieve (A.2.18). Furthermore, it follows from (3.9) that the resulting ℓ_1-shrinkage bound will satisfy:

$$\text{Sbound}^{\text{LS-Boost}(\varepsilon)} \leq \|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2 \sqrt{k^{\text{LS-Boost}(\varepsilon)}} \ .$$

For FS_ε, if one works out the arithmetic based on part *(iii)* of Theorem 3.1, the optimal number of boosting iterations to achieve (A.2.18) is given by: $k^{\text{FS}_\varepsilon} := \left\lceil \frac{4p}{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})} \left(\frac{1}{\tau^2}\right) \right\rceil - 1$ using the learning rate $\varepsilon = \frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2}{\sqrt{k^{\text{FS}_\varepsilon}+1}}$. Also, it follows from part *(v)* of Theorem 3.1 that the resulting shrinkage bound will satisfy:

$$\text{Sbound}^{\text{FS}_\varepsilon} \leq \varepsilon \cdot k^{\text{FS}_\varepsilon} \approx \|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2 \cdot \sqrt{k^{\text{FS}_\varepsilon}} \ .$$

Observe that $k^{\text{LS-Boost}(\varepsilon)} < k^{\text{FS}_\varepsilon}$, whereby LS-Boost$(\varepsilon)$ is able to achieve (A.2.18) in fewer iterations than FS_ε. Indeed, if we let η denote the ratio $k^{\text{LS-Boost}(\varepsilon)}/k^{\text{FS}_\varepsilon}$, then it holds that

$$\eta := \frac{k^{\text{LS-Boost}(\varepsilon)}}{k^{\text{FS}_\varepsilon}} \approx \frac{\ln\left(\frac{1}{\tau^2}\right)}{\frac{1}{\tau^2}} \leq \frac{1}{e} \;<\; 0.368\;. \tag{A.2.19}$$

The left panel of Figure A.2 shows the value of η as a function of τ. For small values of the tolerance parameter τ we see that η is itself close to zero, which means that LS-Boost(ε) will need significantly fewer iterations than FS_ε to achieve the condition (A.2.18).

We can also examine the ℓ_1-shrinkage bounds similarly. If we let ϑ denote the ratio of

$\textsc{Sbound}^{\text{LS-Boost}(\varepsilon)}$ to $\textsc{Sbound}^{\text{FS}_\varepsilon}$, then it holds that

(A.2.20)

$$\vartheta := \frac{\textsc{Sbound}^{\text{LS-Boost}(\varepsilon)}}{\textsc{Sbound}^{\text{FS}_\varepsilon}} = \sqrt{\frac{k^{\text{LS-Boost}(\varepsilon)}}{k^{\text{FS}_\varepsilon}}} = \frac{\sqrt{\ln\left(\frac{1}{\tau^2}\right)}}{\frac{1}{\tau}} \leq \frac{1}{\sqrt{e}} \;<\; 0.607\;.$$

This means that if both bounds are relatively tight, then the ℓ_1-shrinkage of the final model produced by LS-Boost(ε) is smaller than that of the final model produced by FS_ε, by at least a factor of 0.607. The right panel of Figure A.2 shows the value of ϑ as a function of τ. For small values of the relative predication error constant τ we see that ϑ is itself close to zero.

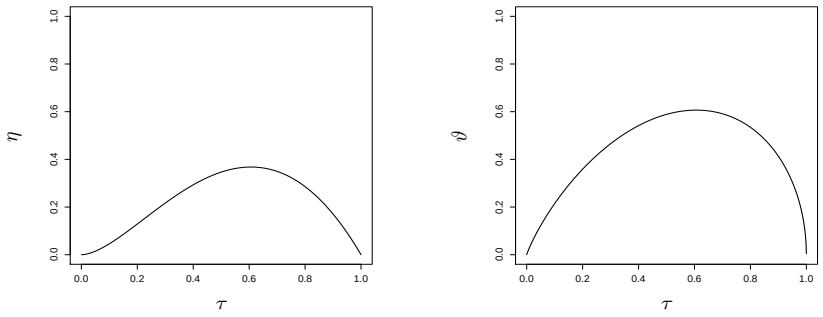

Fig A.2. *Plot showing the value of the ratio η of iterations of* LS-Boost(ε) *to FS_ε (equation (A.2.19)) versus the target relative prediction error τ [left panel], and the ratio ϑ of shrinkage bounds of* LS-Boost(ε) *to FS_ε (equation (A.2.20)) versus the target relative prediction error τ [right panel].*

We summarize the above analysis in the following remark.

REMARK A.2.1. **(Comparison of efficiency of LS-Boost(ε) and FS$_\varepsilon$)** *Suppose that the primary goal is to achieve a τ-relative prediction error as defined in (A.2.18), and that* LS-BOOST(ε) *and FS_ε are run with appropriately determined learning rates for each algorithm. Then the ratio of required number of iterations of these methods to achieve (A.2.18) satisfies*

$$\eta := \frac{k^{\text{LS-BOOST}(\varepsilon)}}{k^{FS_\varepsilon}} < 0.368 \ .$$

Also, the ratio of the shrinkage bounds from running these numbers of iterations satisfies

$$\vartheta := \frac{\text{SBOUND}^{\text{LS-BOOST}(\varepsilon)}}{\text{SBOUND}^{FS_\varepsilon}} < 0.607 \ ,$$

where all of the analysis derives from the bounds in Theorems 2.1 and 3.1.

We caution the reader that the analysis leading to Remark A.2.1 is premised on the singular goal of achieving (A.2.18) in as few iterations as possible. As mentioned previously, the models produced in the interior of the boosting profile are more statistically interesting than those produced at the end. Thus for both algorithms it may be beneficial, and may lessen the risk of over-fitting, to trace out a smoother profile by selecting the learning rate ε to be smaller than the prescribed values in this subsection ($\varepsilon = 1$ for LS-BOOST(ε) and $\varepsilon = \frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2}{\sqrt{k^{\text{FS}_\varepsilon}+1}}$ for FS$_\varepsilon$). Indeed, considering just LS-BOOST(ε) for simplicity, if our goal is to produce a τ-relative prediction error according to (A.2.18) with the largest value of the ℓ_1 norm of the coefficients, then Figure 3 suggests that this should be accomplished by selecting ε as small as possible (essentially very slightly larger than 0).

A.2.8. *Tightness of bounds.* Herein we show the results of some numerical experiments to assess the tightness of the bounds provided in Theorem 2.1. We considered two different classes of examples, Type 1 and Type 2, as described below.

Type 1: In this example, we set $\mathbf{X}$ such that $\mathbf{X}^T\mathbf{X} = \mathbf{\Sigma}_{p\times p}$ with the (i,j)-entry of $\mathbf{\Sigma}$ being given by: $\sigma_{ij} = \rho$ for $i \neq j$ and $\sigma_{ii} = 1$ otherwise. The response was generated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0$, without any noise. Here $\beta_1^0 = 3$ and $\beta_i^0 = 0$ otherwise.

Type 2: This example is similar to Type 1, with the exception of $\boldsymbol{\beta}^0$. Here $\beta_i^0 = 3$ for all $i = 1, \ldots, p$.

LS-BOOST(ε) was allowed to run for different values of ε for different datasets generated as per Types 1 and 2, for different values of ρ and p. Let $\hat{\gamma}_k := (L_n(\hat{\beta}^{k+1}) - L_n^*)/(L_n(\hat{\beta}^k) - L_n^*)$ denote the estimated rate of

change of the loss function optimality gap at iteration k. If the upper bounds implied by Theorem 2.1 are tight, the quantity $\hat{\gamma}_k \equiv \gamma$ for all values of k. We considered the evolution of LS-Boost(ε) for a fixed ε for up to $20,000$ iterations. We set $q_{10\%}, q_{50\%}$ and $q_{90\%}$ as the 10th, 50th and 90th quantiles of the observed values $\hat{\gamma}_k$ for $k \geq 1$, and we show the values of their deviations from γ in Table A.1. Table A.1 shows results for $\varepsilon \in \{0.1, 1\}$ — we see that the theoretical values and the observed values are quite close. The theoretical and empirical values are in good agreement for all values of ε. (Note that for smaller values of ε, the value of γ becomes progressively closer to one.)

We also study the evolution of the ℓ_1-shrinkage of the coefficients as per Theorem 2.1, part (v). We considered the following modified upper bound:

$$\|\hat{\beta}^k\|_1 \leq \min\left\{\sqrt{k}\sqrt{\tfrac{\varepsilon}{2-\varepsilon}}\sqrt{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2} \;\;,\;\; \frac{\varepsilon\|\mathbf{X}\hat{\beta}_{LS}\|_2}{1-\sqrt{\gamma}}\left(1-\gamma^{k/2}\right)\right\},$$

by dropping the term $(-\|\mathbf{X}\hat{\beta}_{\text{LS}} - \mathbf{X}\hat{\beta}^k\|_2^2)$ from the first component. Figure A.3 shows the results for LS-Boost(ε) for a fixed $\varepsilon \in \{0.001, 0.01\}$. The figure shows that the bounds on the ℓ_1-shrinkage of the coefficients are tight for early iterations k, but are not tight for large values of k. We observe empirically that the theoretical and empirical bounds are in good agreement for small/moderate values of ε – but the agreement deteriorates with larger values of ε that are close to one.

A.3. Additional Details for Section 4.

A.3.1. *Duality Between Regularized Correlation Minimization and the* Lasso. In this section, we precisely state the duality relationship between the RCM problem (4.2) and the Lasso optimization problem (1.1). We first prove the following property of the least squares loss function that will be useful in our analysis.

Proposition A.3.1. *The least squares loss function $L_n(\cdot)$ has the following max representation:*

$$L_n(\beta) = \max_{\tilde{r}\in P_{res}} \left\{-\tilde{r}^T(\tfrac{1}{n}\mathbf{X})\beta - \tfrac{1}{2n}\|\tilde{r}-\mathbf{y}\|_2^2 + \tfrac{1}{2n}\|\mathbf{y}\|_2^2\right\} \;, \tag{A.3.1}$$

where $P_{\text{res}} := \{r \in \mathbb{R}^n : r = \mathbf{y} - \mathbf{X}\beta$ for some $\beta \in \mathbb{R}^p\}$. Moreover, the unique optimal solution (as a function of β) to the subproblem in (A.3.1) is $\bar{r} := \mathbf{y} - \mathbf{X}\beta$.

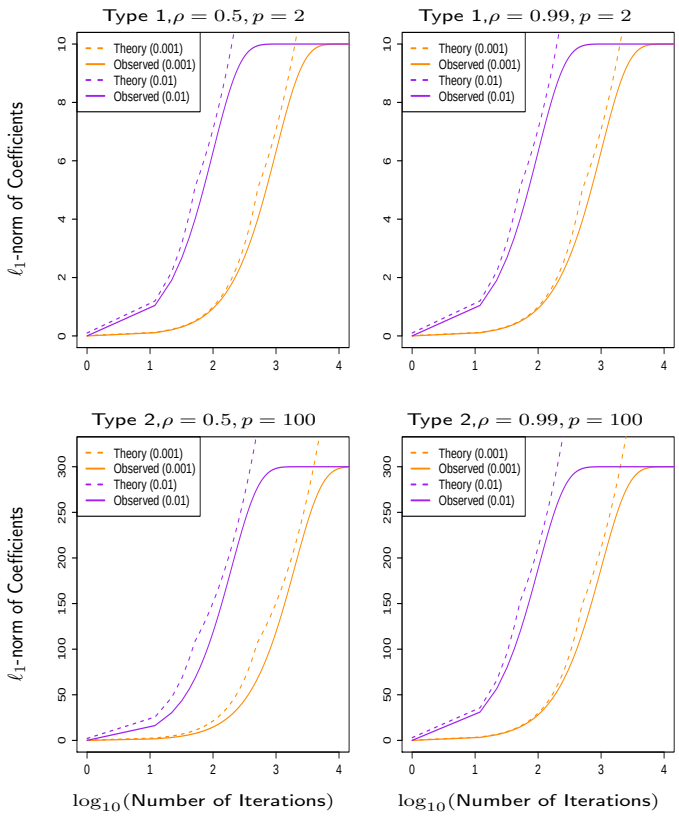

FIG A.3. *Figure showing observed versus theoretical bounds (as implied by Theorem 2.1) of the ℓ_1-norm of the regression coefficients obtained from* LS-BOOST(ε), *for different values of ε and different datasets. The horizontal axis is the number of* LS-BOOST(ε) *iterations (in the log-scale). The top panels show results for data of Type 1, and the bottom panels show results for data of Type 2.*

PROOF. For any $\beta \in \mathbb{R}^p$, it is easy to verify through optimality conditions (setting the gradient with respect to $\tilde{r}$ equal to 0) that $\bar{r}$ solves the subproblem in (A.3.1), i.e.,

$$\bar{r} = \arg\max_{\tilde{r} \in P_{\text{res}}} \left\{ -\tilde{r}^T (\tfrac{1}{n}\mathbf{X})\beta - \tfrac{1}{2n}\|\tilde{r} - \mathbf{y}\|_2^2 + \tfrac{1}{2n}\|\mathbf{y}\|_2^2 \right\} .$$

Thus, we have

$$\begin{aligned} \max_{\tilde{r} \in P_{\text{res}}} \left\{ -\tilde{r}^T (\tfrac{1}{n}\mathbf{X})\beta - \tfrac{1}{2n}\|\tilde{r} - \mathbf{y}\|_2^2 + \tfrac{1}{2n}\|\mathbf{y}\|_2^2 \right\} &= \tfrac{1}{n}\left(\tfrac{1}{2}\|\mathbf{y}\|_2^2 - \mathbf{y}^T\mathbf{X}\beta + \tfrac{1}{2}\|\mathbf{X}\beta\|_2^2 \right) \\ &= \tfrac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 . \end{aligned}$$

□

The next result demonstrates that RCM (4.2) has a direct interpretation as a (scaled) dual of the LASSO problem (1.1). Moreover, in part *(iii)* of the result below, we present a bound on the optimality gap for the LASSO problem in terms of a quantity that is closely related to the objective function of RCM.

PROPOSITION A.3.2. **(Duality Equivalence of Lasso and RCM$_\delta$, and Optimality Bounds)** *The* LASSO *problem* (1.1) *and the regularized correlation minimization problem* RCM$_\delta$ (4.2) *are dual optimization problems modulo the scaling factor* $\frac{n}{\delta}$. *In particular:*

(i) (Weak Duality) If β *is feasible for the* LASSO *problem* (1.1), *and if* $\tilde{r}$ *is feasible for the regularized correlation minimization problem* RCM$_\delta$ (4.2), *then*

$$L_n(\beta) + \tfrac{\delta}{n} f_\delta(\tilde{r}) \;\geq\; \tfrac{1}{2n}\|\mathbf{y}\|_2^2 \;.$$

(ii) (Strong Duality) It holds that:

$$L^*_{n,\delta} + \tfrac{\delta}{n} f^*_\delta = \tfrac{1}{2n}\|\mathbf{y}\|_2^2 \;.$$

Moreover, for any given parameter value $\delta \geq 0$, *there is a unique vector of residuals* $\hat{r}^*_\delta$ *associated with every* LASSO *solution* $\hat{\beta}^*_\delta$, *i.e.,* $\hat{r}^*_\delta = \mathbf{y} - \mathbf{X}\hat{\beta}^*_\delta$, *and* $\hat{r}^*_\delta$ *is the unique optimal solution to the RCM*$_\delta$ *problem* (4.2).

(iii) (Optimality Condition for LASSO*) If* β *is feasible for the* LASSO *problem* (1.1) *and* $r = \mathbf{y} - \mathbf{X}\beta$, *then*

$$\omega_\delta(\beta) := \|\mathbf{X}^T r\|_\infty - \frac{r^T \mathbf{X}\beta}{\delta} \;\geq 0 \;, \tag{A.3.2}$$

and

$$L_n(\beta) - L^*_{n,\delta} \;\leq\; \tfrac{\delta}{n} \cdot \omega_\delta(\beta) \;.$$

Hence, if $\omega_\delta(\beta) = 0$, *then* β *is an optimal solution of the* LASSO *problem* (1.1). □

PROOF. Let us first construct the problem RCM$_\delta$ using basic constructs of minmax duality. As demonstrated in Proposition A.3.1, the least-squares loss function $L_n(\cdot)$ has the following max representation:

$$L_n(\beta) = \max_{\tilde{r} \in P_{\text{res}}} \left\{ -\tilde{r}^T (\tfrac{1}{n}\mathbf{X})\beta - \tfrac{1}{2n}\|\tilde{r} - \mathbf{y}\|_2^2 + \tfrac{1}{2n}\|\mathbf{y}\|_2^2 \right\} \;.$$

Therefore the LASSO problem (1.1) can be written as

$$\min_{\beta \in B_\delta} \; \max_{\tilde{r} \in P_{\text{res}}} \left\{ -\tilde{r}^T (\tfrac{1}{n}\mathbf{X})\beta - \tfrac{1}{2n}\|\tilde{r} - \mathbf{y}\|_2^2 + \tfrac{1}{2n}\|\mathbf{y}\|_2^2 \right\}$$

where $B_\delta := \{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq \delta\}$. We construct a dual of the above problem by interchanging the min and max operators above, yielding the following dual optimization problem:

$$\max_{\tilde{r}\in P_{\text{res}}} \min_{\beta\in B_\delta} \left\{-\tilde{r}^T(\tfrac{1}{n}\mathbf{X})\beta - \tfrac{1}{2n}\|\tilde{r}-\mathbf{y}\|_2^2 + \tfrac{1}{2n}\|\mathbf{y}\|_2^2\right\} \ .$$

After negating, and dropping the constant term $\frac{1}{2n}\|\mathbf{y}\|_2^2$, the above dual problem is equivalent to:

$$\min_{\tilde{r}\in P_{\text{res}}} \max_{\beta\in B_\delta} \left\{\tilde{r}^T(\tfrac{1}{n}\mathbf{X})\beta\right\} + \tfrac{1}{2n}\|\tilde{r}-\mathbf{y}\|_2^2 \ . \tag{A.3.3}$$

Now notice that

$$\max_{\beta\in B_\delta} \left\{\tilde{r}^T(\tfrac{1}{n}\mathbf{X})\beta\right\} = \tfrac{\delta}{n}\left(\max_{j\in\{1,\ldots,p\}} |\tilde{r}^T\mathbf{X}_j|\right) = \tfrac{\delta}{n}\|\mathbf{X}^T\tilde{r}\|_\infty \ , \tag{A.3.4}$$

from which it follows after scaling by $\frac{n}{\delta}$ that (A.3.3) is equivalent to (4.2).

Let us now prove item (i). Let β be feasible for the Lasso problem (1.1) and $\tilde{r}$ be feasible for the regularized correlation minimization problem RCM_δ (4.2), and let $r = \mathbf{y} - \mathbf{X}\beta$ and let $\tilde{\beta}$ be such that $\tilde{r} = \mathbf{y} - \mathbf{X}\tilde{\beta}$. Then direct arithmetic manipulation yields the following equality:

$$L_n(\beta) + \tfrac{\delta}{n} f_\delta(\tilde{r}) = \tfrac{1}{2n}\|\mathbf{y}\|_2^2 + \tfrac{1}{2n}\|r-\tilde{r}\|_2^2 + \tfrac{\delta}{n}\left(\|\mathbf{X}^T\tilde{r}\|_\infty - \frac{\tilde{r}^T\mathbf{X}\beta}{\delta}\right) \ , \tag{A.3.5}$$

from which the result follows since $\|r-\tilde{r}\|_2^2 \geq 0$ and $\tilde{r}^T\mathbf{X}\beta \leq \|\mathbf{X}^T\tilde{r}\|_\infty\|\beta\|_1 \leq \delta\|\mathbf{X}^T\tilde{r}\|_\infty$ which implies that the last term above is also nonnegative.

To prove item (ii), notice that both the Lasso and RCM_δ can be re-cast as optimization problems with a convex quadratic objective function and with linear inequality constraints. That being the case, the classical strong duality results for linearly-constrained convex quadratic optimization apply, see [3] for example.

We now prove (iii). Since β is feasible for the Lasso problem, it follows from the Holder inequality that $r^T\mathbf{X}\beta \leq \|\mathbf{X}^T r\|_\infty\|\beta\|_1 \leq \delta\|\mathbf{X}^T r\|_\infty$, from which it then follows that $\omega_\delta(\beta) \geq 0$. Invoking (A.3.5) with $\tilde{r} \leftarrow r = \mathbf{y} - \mathbf{X}\beta$ yields:

$$L_n(\beta) + \tfrac{\delta}{n} f_\delta(r) = \tfrac{1}{2n}\|\mathbf{y}\|_2^2 + \tfrac{\delta}{n}\cdot\omega_\delta(\beta) \ .$$

Combining the above with strong duality (ii) yields:

$$L_n(\beta) + \tfrac{\delta}{n} f_\delta(r) = L_{n,\delta}^* + \tfrac{\delta}{n} f_\delta^* + \tfrac{\delta}{n}\cdot\omega_\delta(\beta) \ .$$

After rearranging we have:

$$L_n(\beta) - L_{n,\delta}^* \leq \tfrac{\delta}{n} f_\delta^* - \tfrac{\delta}{n} f_\delta(r) + \tfrac{\delta}{n}\cdot\omega_\delta(\beta) \leq \tfrac{\delta}{n}\cdot\omega_\delta(\beta) \ ,$$

where the last inequality follows since $f_\delta^* \leq f_\delta(r)$. □

A.3.2. *Proof of Proposition 4.1.* Recall the update formula for the residuals in R-FS$_{\varepsilon,\delta}$:

$$\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon \left[\operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) \mathbf{X}_{j_k} + \tfrac{1}{\delta}(\hat{r}^k - \mathbf{y}) \right] . \tag{A.3.6}$$

We first show that $g^k := \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) \mathbf{X}_{j_k} + \frac{1}{\delta}(\hat{r}^k - \mathbf{y})$ is a subgradient of $f_\delta(\cdot)$ at $\hat{r}^k$. Recalling the proof of Proposition 3.2, we have that $\operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) \mathbf{X}_{j_k}$ is a subgradient of $f(r) := \|\mathbf{X}^T r\|_\infty$ at $\hat{r}^k$ since $j_k \in \arg\max_{j \in \{1,\dots,p\}} |(\hat{r}^k)^T \mathbf{X}_j|$. Therefore, since $f_\delta(r) = f(r) + \frac{1}{2\delta}\|r - \mathbf{y}\|_2^2$, it follows from the additive property of subgradients (and gradients) that $g^k = \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) \mathbf{X}_{j_k} + \frac{1}{\delta}(\hat{r}^k - \mathbf{y})$ is a subgradient of $f_\delta(r)$ at $r = \hat{r}^k$. Therefore the update (A.3.6) is of the form $\hat{r}^{k+1} = \hat{r}^k - \varepsilon g^k$ where $g^k \in \partial f_\delta(\hat{r}^k)$. Finally note that $\hat{r}^k - \varepsilon g^k = \hat{r}^{k+1} = \mathbf{y} - \mathbf{X}\beta^{k+1} \in P_{\text{res}}$, hence $\Pi_{P_{\text{res}}}(\hat{r}^k - \varepsilon g^k) = \hat{r}^k - \varepsilon g^k$, i.e., the projection step is superfluous here. Therefore $\hat{r}^{k+1} = \Pi_{P_{\text{res}}}(\hat{r}^k - \varepsilon g^k)$, which shows that (A.3.6) is precisely the update for the subgradient descent method with step-size $\alpha_k := \varepsilon$. □

A.3.3. *Proof of Theorem 4.1.* Let us first use induction to demonstrate that the following inequality holds:

$$\|\hat{\beta}^k\|_1 \le \varepsilon \sum_{j=0}^{k-1} \left(1 - \tfrac{\varepsilon}{\delta}\right)^j \quad \text{for all } k \ge 0 . \tag{A.3.7}$$

Clearly, (A.3.7) holds for $k = 0$ since $\hat{\beta}^0 = 0$. Assuming that (A.3.7) holds for k, then the update for $\hat{\beta}^{k+1}$ in step (3.) of algorithm R-FS$_{\varepsilon,\delta}$ can be written as $\hat{\beta}^{k+1} = (1 - \frac{\varepsilon}{\delta})\hat{\beta}^k + \varepsilon \cdot \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) e_{j_k}$, from which it holds that

$$\begin{aligned}
\|\hat{\beta}^{k+1}\|_1 &= \|(1 - \tfrac{\varepsilon}{\delta})\hat{\beta}^k + \varepsilon \cdot \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) e_{j_k}\|_1 \\
&\le (1 - \tfrac{\varepsilon}{\delta})\|\hat{\beta}^k\|_1 + \varepsilon \|e_{j_k}\|_1 \\
&\le (1 - \tfrac{\varepsilon}{\delta})\varepsilon \sum_{j=0}^{k-1} \left(1 - \tfrac{\varepsilon}{\delta}\right)^j \; + \; \varepsilon \\
&= \varepsilon \sum_{j=0}^{k} \left(1 - \tfrac{\varepsilon}{\delta}\right)^j ,
\end{aligned}$$

which completes the induction. Now note that (A.3.7) is a geometric series and we have:

$$\text{(A.3.8)}\quad \|\hat{\beta}^k\|_1 \leq \varepsilon \sum_{j=0}^{k-1} \left(1-\tfrac{\varepsilon}{\delta}\right)^j = \delta\left[1-\left(1-\tfrac{\varepsilon}{\delta}\right)^k\right] \leq \delta \quad \text{for all } k \geq 0\ .$$

Recall that we developed the algorithm R-FS$_{\varepsilon,\delta}$ in such a way that it corresponds exactly to an instantiation of the subgradient descent method applied to the RCM problem (4.2). Indeed, the update rule for the residuals given in Step (3.) of R-FS$_{\varepsilon,\delta}$ is: $\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon g^k$ where $g^k = \left[\operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k})\mathbf{X}_{j_k} + \frac{1}{\delta}(\hat{r}^k - \mathbf{y})\right]$. We therefore can apply Proposition A.2.4, and more specifically the inequality (A.2.13). In order to do so we need to translate the terms of Proposition A.2.4 to our setting: here the variables x are now the residuals r, the iterates x^i are now the iterates $\hat{r}^i$, etc. The step-sizes of algorithm R-FS$_{\varepsilon,\delta}$ are fixed at ε, so we have $\alpha_i = \varepsilon$ for all $i \geq 0$. Setting the value of x in Proposition A.2.4 to be least-squares residual value, namely $x = \hat{r}_{LS}$, the left side of (A.2.13) is therefore:

(A.3.9)

$$\begin{aligned}
\tfrac{1}{k+1}\textstyle\sum_{i=0}^k (g^i)^T(x^i - x) &= \tfrac{1}{k+1}\textstyle\sum_{i=0}^k \left(\mathbf{X}\left[\operatorname{sgn}((\hat{r}^i)^T\mathbf{X}_{j_i})e_{j_i} - \tfrac{1}{\delta}\hat{\beta}^i\right]\right)^T (\hat{r}^i - \hat{r}_{LS}) \\
&= \tfrac{1}{k+1}\textstyle\sum_{i=0}^k \left(\operatorname{sgn}((\hat{r}^i)^T\mathbf{X}_{j_i})\mathbf{X}_{j_i} - \tfrac{1}{\delta}(\mathbf{X}\hat{\beta}^i)\right)^T \hat{r}^i \\
&= \tfrac{1}{k+1}\textstyle\sum_{i=0}^k \left[\|\mathbf{X}^T\hat{r}^i\|_\infty - \tfrac{1}{\delta}(\hat{r}^i)^T\mathbf{X}\hat{\beta}^i\right] \\
&= \tfrac{1}{k+1}\textstyle\sum_{i=0}^k \omega_\delta(\hat{\beta}^i)\ ,
\end{aligned}$$

where the second equality uses the fact that $\mathbf{X}^T\hat{r}_{LS} = 0$ from (2.3) and the fourth equality uses the definition of $\omega_\delta(\beta)$ from (A.3.2).

Let us now evaluate the right side of (A.2.13). We have $\|x^0 - x\|_2 = \|\hat{r}^0 - \hat{r}_{LS}\|_2 = \|\mathbf{y} - (\mathbf{y} - \mathbf{X}\hat{\beta}_{\mathrm{LS}})\|_2 = \|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2$. Also, it holds that

$$\begin{aligned}
\|g^i\|_2 =& \|\operatorname{sgn}((\hat{r}^i)^T\mathbf{X}_{j_i})\mathbf{X}_{j_i} - \tfrac{1}{\delta}(\mathbf{X}\hat{\beta}^i)\|_2 \\
\leq& \|\mathbf{X}_{j_i}\|_2 + \|\mathbf{X}(\tfrac{\hat{\beta}^i}{\delta})\|_2 \\
\leq& 1 + \tfrac{1}{\delta}\|\mathbf{X}\|_{1,2}\|\hat{\beta}^i\|_1 \\
\leq& 1 + \|\mathbf{X}\|_{1,2} \\
\leq& 2\ ,
\end{aligned}$$

where the third inequality follows since $\|\hat{\beta}^i\|_1 \leq \delta$ from (A.3.8) and the second and fourth inequalities follow from the assumption that the columns

of $\mathbf{X}$ have been normalized to have unit ℓ_2 norm. Therefore $G = 2$ is a uniform bound on $\|g^i\|_2$. Combining the above, inequality (A.2.13) implies that after running R-FS$_{\varepsilon,\delta}$ for k iterations, it holds that:

(A.3.10)

$$\min_{i\in\{0,\ldots,k\}} \omega_\delta(\hat\beta^i) \;\le\; \frac{1}{k+1}\sum_{i=0}^{k}\omega_\delta(\hat\beta^i) \;\le\; \frac{\|\mathbf{X}\hat\beta_{\mathrm{LS}}\|_2^2}{2(k+1)\varepsilon} + \frac{2^2\varepsilon}{2} \;=\; \frac{\|\mathbf{X}\hat\beta_{\mathrm{LS}}\|_2^2}{2\varepsilon(k+1)} + 2\varepsilon\;,$$

where the first inequality is elementary arithmetic and the second inequality is the application of (A.2.13). Now let i be the index obtaining the minimum in the left-most side of the above. Then it follows from part (iii) of Proposition A.3.2 that

$$L_n(\hat\beta^i) - L^*_{n,\delta} \;\le\; \tfrac{\delta}{n}\cdot\omega_\delta(\hat\beta^i) \le \frac{\delta\|\mathbf{X}\hat\beta_{\mathrm{LS}}\|_2^2}{2n\varepsilon(k+1)} + \frac{2\delta\varepsilon}{n}\;, \tag{A.3.11}$$

which proves item (i) of the theorem.

To prove item (ii), note first that if $\hat\beta^*_\delta$ is a solution of the Lasso problem (1.1), then it holds that $\|\hat\beta^*_\delta\|_1 \le \delta$ (feasibility) and $\omega_\delta(\hat\beta^*_\delta) = 0$ (optimality). This latter condition follows easily from the optimality conditions of linearly constrained convex quadratic problems, see [3] for example. Setting $\hat r^*_\delta = \mathbf{y} - \mathbf{X}\hat\beta^*_\delta$, the following holds true:

$$\begin{aligned}
\|\mathbf{X}\hat\beta^i - \mathbf{X}\hat\beta^*_\delta\|_2^2 &= 2n\left(L_n(\hat\beta^i) - L_n(\hat\beta^*_\delta) + (\hat r^*_\delta)^T\mathbf{X}(\hat\beta^i - \hat\beta^*_\delta)\right)\\
&= 2n\left(L_n(\hat\beta^i) - L^*_{n,\delta} - \delta\|\mathbf{X}^T\hat r^*_\delta\|_\infty + (\hat r^*_\delta)^T\mathbf{X}\hat\beta^i\right)\\
&\le 2n\left(L_n(\hat\beta^i) - L^*_{n,\delta} - \delta\|\mathbf{X}^T\hat r^*_\delta\|_\infty + \|\mathbf{X}^T\hat r^*_\delta\|_\infty\|\hat\beta^i\|_1\right)\\
&\le 2n\left(L_n(\hat\beta^i) - L^*_{n,\delta} - \delta\|\mathbf{X}^T\hat r^*_\delta\|_\infty + \delta\|\mathbf{X}^T\hat r^*_\delta\|_\infty\right)\\
&= 2n\left(L_n(\hat\beta^i) - L^*_{n,\delta}\right)\\
&\le \tfrac{\delta\|\mathbf{X}\hat\beta_{\mathrm{LS}}\|_2^2}{\varepsilon(k+1)} + 4\delta\varepsilon\;,
\end{aligned}$$

where the first equality is from direct arithmetic substitution, the second equality uses the fact that $\omega_\delta(\hat\beta^*_\delta) = 0$ whereby $(\hat r^*_\delta)^T\mathbf{X}\hat\beta^*_\delta = \delta\|\mathbf{X}^T\hat r^*_\delta\|_\infty$, the first inequality follows by applying Holder's inequality to the last term of the second equality, and the final inequality is an application of (A.3.11). Item (ii) then follows by taking square roots of the above.

Item (iii) is essentially just (A.3.8). Indeed, since $i \le k$ we have:

$$\|\hat{\beta}^i\|_1 \le \ \varepsilon \sum_{j=0}^{i-1} \left(1 - \tfrac{\varepsilon}{\delta}\right)^j \le \ \varepsilon \sum_{j=0}^{k-1} \left(1 - \tfrac{\varepsilon}{\delta}\right)^j = \ \delta \left[1 - \left(1 - \tfrac{\varepsilon}{\delta}\right)^k\right] \le \ \delta \ .$$

(Note that we emphasize the dependence on k rather than i in the above since we have direct control over the number of boosting iterations k.) Item (iv) of the theorem is just a restatement of the sparsity property of R-FS$_{\varepsilon,\delta}$. □

A.3.4. *Regularized Boosting: Related Work and Context.* As we have already seen, the FS$_\varepsilon$ algorithm leads to models that have curious similarities with the Lasso coefficient profile, but in general the profiles are different. Sufficient conditions under which the coefficient profiles of FS$_\varepsilon$ (for $\varepsilon \approx 0$) and Lasso are equivalent have been explored in [34]. A related research question is whether there are structurally similar algorithmic variants of FS$_\varepsilon$ that lead to Lasso solutions for arbitrary datasets? In this vein [53] propose BLasso, a corrective version of the forward stagewise algorithm. BLasso, in addition to taking incremental forward steps (as in FS$_\varepsilon$), also takes backward steps, the result of which is that the algorithm approximates the Lasso coefficient profile under certain assumptions on the data. The authors observe that BLasso often leads to models that are sparser and have better predictive accuracy than those produced by FS$_\varepsilon$.

In [9], the authors point out that models delivered by boosting methods need not be adequately sparse, and they highlight the importance of obtaining models that have more sparsity, better prediction accuracy, and better variable selection properties. They propose a sparse variant of $L2$-Boost (see also Section 1) which considers a regularized version of the squared error loss, penalizing the approximate degrees of freedom of the model.

In [26], the authors also point out that boosting algorithms often lead to a large collection of nonzero coefficients. They suggest reducing the complexity of the model by some form of "post-processing" technique—one such proposal is to apply a Lasso regularization on the selected set of coefficients.

A parallel line of work in machine learning [13] explores the scope of boosting-like algorithms on ℓ_1-regularized versions of different loss functions arising mainly in the context of classification problems. The proposal of [13], when adapted to the least squares regression problem with ℓ_1-regularization penalty, leads to the penalized version of the Lasso problem:

$$\min_\beta \ \ \tfrac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \ , \tag{A.3.12}$$

for which the authors [13] employ greedy coordinate descent methods. Like the boosting algorithms considered herein, at each iteration the algorithm studied by [13] selects a certain coefficient β_{j_k} to update, leaving all other coefficients β_i unchanged. The amount with which to update the coefficient β_{j_k} is determined by fully optimizing the loss function (A.3.12) with respect to β_{j_k}, again holding all other coefficients constant (note that one recovers LS-Boost(1) if $\lambda = 0$). This way of updating β_{j_k} leads to a simple soft-thresholding operation [12] and is *structurally* different from forward stagewise update rules. In contrast, the boosting algorithm R-FS$_{\varepsilon,\delta}$ that we propose here is based on subgradient descent on the dual of the Lasso problem (1.1), i.e., problem (4.2).

A.3.5. *Connecting R-FS$_{\varepsilon,\delta}$ to the Frank-Wolfe method.* Although we developed and analyzed R-FS$_{\varepsilon,\delta}$ from the perspective of subgradient descent, one can also interpret R-FS$_{\varepsilon,\delta}$ as the Frank-Wolfe algorithm in convex optimization [15, 35, 18] applied to the Lasso (1.1). This secondary interpretation can be derived directly from the structure of the updates in R-FS$_{\varepsilon,\delta}$ or as a special case of a more general primal-dual equivalence between subgradient descent and Frank-Wolfe developed in [1]. Other authors have commented on the qualitative similarities between boosting algorithms and the Frank-Wolfe method, see for instance [10] and [35].

We note that there are subtle differences between R-FS$_{\varepsilon,\delta}$ and a standard Frank-Wolfe method — the difference lies in the choice of the step-size sequence. Standard (popular) Frank Wolfe methods [35] (see also, references therein) use a step-size of $2/(k+2)$ (where k is the iteration index) or a step-size based on a line-search. The step-size in R-FS$_{\varepsilon,\delta}$, however, is *fixed* across iterations — this is done to make R-FS$_{\varepsilon,\delta}$ coincide with a boosting algorithm with fixed learning rate (and in particular, to make it coincide with FS$_\varepsilon$ when $\delta = \infty$). Analysis of this particular step-size choice in R-FS$_{\varepsilon,\delta}$ requires a different technique than that used in a more traditional analysis of Frank-Wolfe such as that presented in [35]. In particular, the analysis we present herein is based on tools from subgradient optimization, which we utilize since subgradient descent provides a natural unifying framework for a general class of boosting algorithms (including FS$_\varepsilon$ and R-FS$_{\varepsilon,\delta}$) via a single algorithm applied to a parametric class of objective functions. We note that [18] derive computational guarantees for Frank-Wolfe with a general step-size sequence – including a fixed step-size across iterations; however their proof is different than that presented herein.

A.4. Additional Details for Section 5.

A.4.1. *Proof of Theorem 5.1.* We first prove the feasibility of $\hat{\beta}^k$ for the LASSO problem with parameter $\bar{\delta}_k$. We do so by induction. The feasibility of $\hat{\beta}^k$ is obviously true for $k = 0$ since $\hat{\beta}^0 = 0$ and hence $\|\hat{\beta}^0\|_1 = 0 < \bar{\delta}_0$. Now suppose it is true for some iteration k, i.e., $\|\hat{\beta}^k\|_1 \leq \bar{\delta}_k$. Then the update for $\hat{\beta}^{k+1}$ in step (3.) of algorithm PATH-R-FS$_\varepsilon$ can be written as $\hat{\beta}^{k+1} = (1 - \frac{\varepsilon}{\bar{\delta}_k})\hat{\beta}^k + \frac{\varepsilon}{\bar{\delta}_k}(\bar{\delta}_k \text{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) e_{j_k})$, from which it follows that

$$\begin{aligned} \|\hat{\beta}^{k+1}\|_1 &= \|(1 - \tfrac{\varepsilon}{\bar{\delta}_k})\hat{\beta}^k + \tfrac{\varepsilon}{\bar{\delta}_k}(\bar{\delta}_k \text{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) e_{j_k})\|_1 \\ &\leq (1 - \tfrac{\varepsilon}{\bar{\delta}_k})\|\hat{\beta}^k\|_1 + \tfrac{\varepsilon}{\bar{\delta}_k}\|\bar{\delta}_k e_{j_k}\|_1 \leq (1 - \tfrac{\varepsilon}{\bar{\delta}_k})\bar{\delta}_k + \tfrac{\varepsilon}{\bar{\delta}_k}\bar{\delta}_k = \bar{\delta}_k \leq \bar{\delta}_{k+1} \ , \end{aligned}$$

which completes the induction.

We now prove the bound on the average training error in part *(i)*. In fact, we will prove something stronger than this bound, namely we will prove:

$$\frac{1}{k+1}\sum_{i=0}^{k} \frac{1}{\bar{\delta}_i}\left(L_n(\hat{\beta}^i) - L^*_{n,\bar{\delta}_i}\right) \ \leq \ \frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2}{2n\varepsilon(k+1)} + \frac{2\varepsilon}{n} \ , \tag{A.4.1}$$

from which average training error bound of part *(i)* follows since $\bar{\delta}_i \leq \bar{\delta}$ for all i. The update rule for the residuals given in Step (3.) of R-FS$_{\varepsilon,\delta}$ is: $\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon g^k$ where $g^k = \left[\text{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k})\mathbf{X}_{j_k} + \frac{1}{\bar{\delta}_k}(\hat{r}^k - \mathbf{y})\right]$. This update rule is precisely in the format of an elementary sequence process, see Appendix A.2.5, and we therefore can apply Proposition A.2.4, and more specifically the inequality (A.2.13). Similar in structure to the proof of Theorem 4.1, we first need to translate the terms of Proposition A.2.4 to our setting: once again the variables x are now the residuals r, the iterates x^i are now the iterates $\hat{r}^i$, etc. The step-sizes of algorithm PATH-R-FS$_\varepsilon$ are fixed at ε, so we have $\alpha_i = \varepsilon$ for all $i \geq 0$. Setting the value of x in Proposition A.2.4 to be least-squares residual value, namely $x = \hat{r}_{LS}$, and using the exact same logic as in the equations (A.3.9), one obtains the following result about the left side of (A.2.13):

$$\frac{1}{k+1}\sum_{i=0}^{k}(g^i)^T(x^i - x) = \frac{1}{k+1}\sum_{i=0}^{k}\omega_{\bar{\delta}_i}(\hat{\beta}^i) \ .$$

Let us now evaluate the right side of (A.2.13). We have $\|x^0 - x\|_2 = \|\hat{r}^0 -$

$\hat{r}_{LS}\|_2 = \|\mathbf{y} - (\mathbf{y} - \mathbf{X}\hat{\beta}_{\rm LS})\|_2 = \|\mathbf{X}\hat{\beta}_{\rm LS}\|_2$. Also, it holds that

$$\begin{aligned}\|g^i\|_2 =& \|\mathrm{sgn}((\hat{r}^i)^T\mathbf{X}_{j_i})\mathbf{X}_{j_i} - \tfrac{1}{\bar{\delta}_i}(\mathbf{X}\hat{\beta}^i)\|_2\\ \leq& \|\mathbf{X}_{j_i}\|_2 + \|\mathbf{X}(\tfrac{\hat{\beta}^i}{\bar{\delta}_i})\|_2\\ \leq& 1 + \tfrac{1}{\bar{\delta}_i}\|\mathbf{X}\|_{1,2}\|\hat{\beta}^i\|_1\\ \leq& 1 + \|\mathbf{X}\|_{1,2}\\ \leq& 2\end{aligned}$$

where the third inequality follows since $\|\hat{\beta}^i\|_1 \leq \bar{\delta}_i$ from the feasibility of $\hat{\beta}^i$ for the Lasso problem with parameter $\bar{\delta}_i$ proven at the outset, and the second and fourth inequalities follow from the assumption that the columns of $\mathbf{X}$ have been normalized to have unit ℓ_2 norm. Therefore $G = 2$ is a uniform bound on $\|g^i\|_2$. Combining the above, inequality (A.2.13) implies that after running PATH-R-FS$_\varepsilon$ for k iterations, it holds that:

$$\text{(A.4.2)}\qquad \frac{1}{k+1}\sum_{i=0}^{k}\omega_{\bar{\delta}_i}(\hat{\beta}^i) \;\leq\; \frac{\|\mathbf{X}\hat{\beta}_{\rm LS}\|_2^2}{2(k+1)\varepsilon} + \frac{2^2\varepsilon}{2} \;=\; \frac{\|\mathbf{X}\hat{\beta}_{\rm LS}\|_2^2}{2\varepsilon(k+1)} + 2\varepsilon\ ,$$

where the inequality is the application of (A.2.13). From Proposition A.3.2 we have $L_n(\hat{\beta}^i) - L^*_{n,\bar{\delta}_i} \;\leq\; \frac{\bar{\delta}_i}{n}\cdot\omega_{\bar{\delta}_i}(\hat{\beta}^i)$, which combines with (A.4.2) to yield:

$$\frac{1}{k+1}\sum_{i=0}^{k}\frac{1}{\bar{\delta}_i}\left(L_n(\hat{\beta}^i) - L^*_{n,\bar{\delta}_i}\right) \;\leq\; \frac{1}{(k+1)}\frac{1}{n}\sum_{i=0}^{k}\omega_{\bar{\delta}_i}(\hat{\beta}^i) \;\leq\; \frac{\|\mathbf{X}\hat{\beta}_{\rm LS}\|_2^2}{2n\varepsilon(k+1)} + \frac{2\varepsilon}{n}\ .$$

This proves (A.4.1) which then completes the proof of part *(i)* through the bounds $\bar{\delta}_i \leq \bar{\delta}$ for all i.

Part *(ii)* is a restatement of the feasibility of $\hat{\beta}^k$ for the Lasso problem with parameter $\bar{\delta}_k$ which was proved at the outset, and is re-written to be consistent with the format of, and for comparison with, Theorem 4.1. Last of all, part *(iii)* follows since at each iteration at most one new coefficient is introduced at a non-zero level. □

A.5. Additional Details on the Experiments. We describe here some additional details pertaining to the computational results performed in this paper. We first describe in some more detail the real datasets that have been considered in the paper.

Description of datasets considered.
We considered four different publicly available microarray datasets as described below.

Leukemia dataset. This dataset, taken from [11], has binary response with continuous covariates, with 72 samples and approximately 3500 covariates. We further processed the dataset by taking a subsample of $p = 500$ covariates, while retaining all $n = 72$ sample points. We artificially generated the response $\mathbf{y}$ via a linear model with the given covariates $\mathbf{X}$ (as described in Eg-A in Section 6). The true regression coefficient β^{pop} was taken as $\beta_i^{\text{pop}} = 1$ for all $i \leq 10$ and zero otherwise.

Golub dataset. The original dataset was taken from the R package `mpm`, which had 73 samples with approximately 5000 covariates. We reduced this to $p = 500$ covariates (all samples were retained). Responses $\mathbf{y}$ were generated via a linear model with β^{pop} as above.

Khan dataset. This dataset was taken from the dataset webpage `http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/` accompanying the book [33]. The original covariate matrix (`khan.xtest`), which had 73 samples with approximately 5000 covariates, was reduced to $p = 500$ covariates (all samples were retained). Responses $\mathbf{y}$ were generated via a linear model with β^{pop} as above.

Prostate cancer dataset. This dataset appears in [14] and is available from the R package `LARS`. The first column `lcavol` was taken as the response (no artificial response was created here). We generated multiple datasets from this dataset, as follows:

(a) One of the datasets is the original one with $n = 97$ and $p = 8$.
(b) We created another dataset, with $n = 97$ and $p = 44$ by enhancing the covariate space to include second order interactions.
(c) We created another dataset, with $n = 10$ and $p = 44$. We subsampled the dataset from (b), which again was enhanced to include second order interactions.

Note that in all the examples above we standardized $\mathbf{X}$ such that the columns have unit ℓ_2 norm, before running the different algorithms studied herein.

Sensitivity of the Learning Rate in LS-Boost(ε) *and FS*$_\varepsilon$. We performed several experiments running LS-Boost(ε) and FS$_\varepsilon$ on an array of real and synthetic datasets, to explore how the training and test errors change as a function of the number of boosting iterations and the learning rate. Some of the results appear in Figure A.4. The training errors were found to decrease with increasing number of boosting iterations. The rate of decay, however, is very sensitive to the value of ε, with smaller values of ε leading to slower

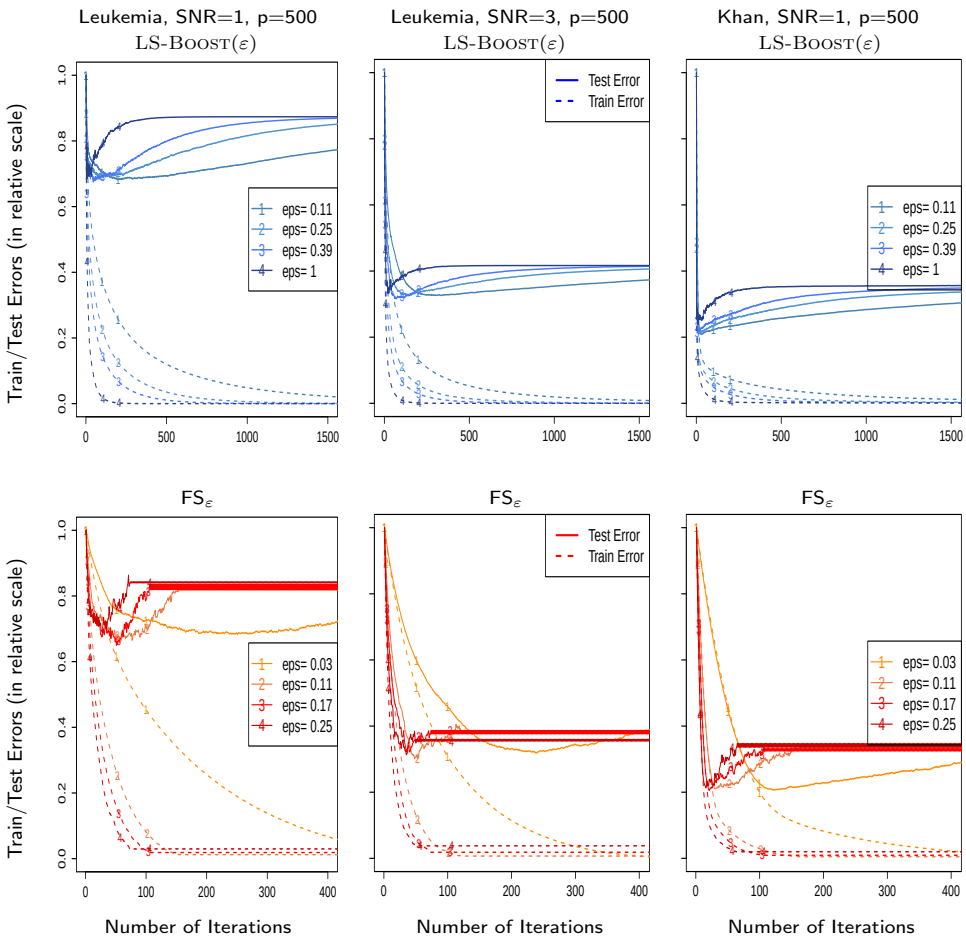

FIG A.4. *Figure showing the training and test errors (in relative scale) as a function of boosting iterations, for both* LS-BOOST(ε) *(top panel) and* FS_ε *(bottom panel). As the number of iterations increases, the training error shows a global monotone pattern. The test errors however, initially decrease and then start increasing after reaching a minimum. The best test errors obtained are found to be sensitive to the choice of* ε*. Two different datasets have been considered: the Leukemia dataset (left and middle panels) and the Khan dataset (right panel), as described in Section 6.*

convergence behavior to the least squares fit, as expected. The test errors were found to decrease and then increase after reaching a minimum; furthermore, the best predictive models were found to be sensitive to the choice of ε.

In addition to the above, we also performed a series of experiments on both real and synthetic datasets comparing the performance of LS-BOOST(ε) and

FS_ε to other sparse learning methods, namely solution via the Lasso, stepwise regression [14] and FS_0 [14]. Our results are presented in Table A.2. In all the cases, we found that the performance of FS_ε and LS-Boost(ε) were at least as good as the Lasso solution. And in some cases, the performances of FS_ε and LS-Boost(ε) were superior. The best predictive models achieved by LS-Boost(ε) and FS_ε correspond to values of ε that are larger than zero or even close to one – this suggests that a proper choice of ε can lead to superior models.

Statistical properties of R-FS$_{\varepsilon,\delta}$, Lasso and FS$_\varepsilon$: an empirical study. We performed some experiments to evaluate the performance of R-$\text{FS}_{\varepsilon,\delta}$, in terms of predictive accuracy and sparsity of the optimal model, versus the more widely known methods FS_ε and solution via the Lasso. In all the cases, we took a small value of $\varepsilon = 10^{-3}$. We ran R-$\text{FS}_{\varepsilon,\delta}$ on a grid of twenty δ values, with the limiting solution corresponding to the Lasso solution estimate at the particular value of δ selected. In all cases, we found that when δ was large, i.e., larger than the best δ for the Lasso (in terms of obtaining a model with the best predictive performance), R-$\text{FS}_{\varepsilon,\delta}$ delivered a model with excellent statistical properties – R-$\text{FS}_{\varepsilon,\delta}$ led to sparse solutions (the sparsity was similar to that of the best Lasso model) and the predictive performance was as good as, and in some cases better than, the Lasso solution. This suggests that the choice of δ does not play a very crucial role in the R-$\text{FS}_{\varepsilon,\delta}$ algorithm, once it is chosen to be reasonably large; indeed the number of boosting iterations plays a more important role in obtaining good quality statistical estimates. When compared to FS_ε (i.e., the version of R-$\text{FS}_{\varepsilon,\delta}$ with $\delta = \infty$) we observed that the best models delivered by R-$\text{FS}_{\varepsilon,\delta}$ were more sparse (i.e., with fewer non-zeros) than the best FS_ε solutions. This complements a popular belief about boosting in that it delivers models that are quite dense – see the discussion herein in Section A.3.4. Furthermore, it shows that the particular form of regularized boosting that we consider, R-$\text{FS}_{\varepsilon,\delta}$, does indeed induce sparser solutions. Our detailed results are presented in Table A.3.

Comments on Table A.2. In this experiment, we ran FS_ε and LS-Boost(ε) for thirty different values of ε in the range 0.001 to 0.8. The entire regularization paths for Lasso, FS_0, and the more aggressive Stepwise regression were computed with the `LARS` package. First, we observe that Stepwise regression, which is quite fast in reaching an unconstrained least squares solution, does not perform well in terms of obtaining a model with good predictive performance. The slowly learning boosting methods perform quite well – in fact their performances are quite similar to the best Lasso solutions. A closer

inspection shows that FS_ε almost always delivers the best predictive models when ε is allowed to be flexible. While a good automated method to find the optimal value of ε is certainly worth investigating, we leave this for future work (of course, there are excellent heuristics for choosing the optimal ε in practice, such as cross validation, etc.). However, we do highlight that in practice a strictly non-zero learning rate ε may lead to better models than its limiting version $\varepsilon = 0+$.

For Eg-A ($\rho = 0.8$), both LS-Boost(ε) and FS_ε achieved the best model at $\varepsilon = 10^{-3}$. For Eg-A ($\rho = 0$), LS-Boost(ε) achieved the best model at $\varepsilon = 0.1, 0.7, 0.8$ and FS_ε achieved the best model at $\varepsilon = 10^{-3}, 0.7, 0.8$ (both for SNR values 1, 3, 10 respectively). For the Leukemia dataset, LS-Boost(ε) achieved the best model at $\varepsilon = 0.6, 0.7, 0.02$ and FS_ε achieved the best model at $\varepsilon = 0.6, 0.02, 0.02$ (both for SNR values 1, 3, 10 respectively). For the Khan dataset, LS-Boost(ε) achieved the best model at $\varepsilon = 0.001, 0.001, 0.02$ and FS_ε achieved the best model at $\varepsilon = 0.001, 0.02, 0.001$ (both for SNR values 1, 3, 10 respectively).

Type	p	ρ	ε	$\lvert q_{10\%} - \gamma\rvert$	$\lvert q_{50\%} - \gamma\rvert$	$\lvert q_{90\%} - \gamma\rvert$	γ
1	2	-0.99	0.1	2.375e-04	2.375e-04	2.375e-04	9.99762e-01
1	2	-0.99	1.0	1.250e-03	1.250e-03	1.250e-03	9.98750e-01
1	2	-0.5	0.1	1.187e-02	1.187e-02	1.187e-02	9.88125e-01
1	2	-0.5	1.0	6.250e-02	6.250e-02	6.250e-02	9.37500e-01
1	2	0.5	0.1	1.187e-02	1.187e-02	1.187e-02	9.88125e-01
1	2	0.5	1.0	6.250e-02	6.250e-02	6.250e-02	9.37500e-01
1	2	0.99	0.1	2.375e-04	2.375e-04	2.375e-04	9.99762e-01
1	2	0.99	1.0	1.250e-03	1.250e-03	1.250e-03	9.98750e-01
2	2	-0.99	0.1	8.146e-04	8.146e-04	8.146e-04	9.99762e-01
2	2	-0.99	1.0	1.865e-02	1.250e-03	1.250e-03	9.98750e-01
2	2	-0.5	0.1	1.187e-02	1.187e-02	1.187e-02	9.88125e-01
2	2	-0.5	1.0	6.250e-02	6.250e-02	6.250e-02	9.37500e-01
2	2	0.5	0.1	1.187e-02	1.187e-02	1.187e-02	9.88125e-01
2	2	0.5	1.0	6.250e-02	6.250e-02	6.250e-02	9.37500e-01
2	2	0.99	0.1	2.375e-04	2.375e-04	2.375e-04	9.99762e-01
2	2	0.99	1.0	1.865e-02	1.250e-03	1.250e-03	9.98750e-01
2	50	-0.99	0.1	2.804e-03	2.615e-03	6.518e-04	9.99348e-01
2	50	-0.99	1.0	3.431e-03	3.431e-03	3.431e-03	9.96569e-01
2	50	-0.5	0.1	4.738e-03	7.345e-04	7.345e-04	9.99265e-01
2	50	-0.5	1.0	3.866e-03	3.866e-03	3.866e-03	9.96134e-01
2	50	0.5	0.1	1.869e-03	1.750e-03	1.716e-03	9.99525e-01
2	50	0.5	1.0	2.500e-03	2.500e-03	2.500e-03	9.97500e-01
2	50	0.99	0.1	9.676e-05	6.597e-05	5.665e-05	9.99991e-01
2	50	0.99	1.0	1.013e-02	5.000e-05	5.000e-05	9.99950e-01
2	100	-0.99	0.1	1.330e-03	1.217e-03	1.153e-03	9.99679e-01
2	100	-0.99	1.0	1.692e-03	1.692e-03	1.692e-03	9.98308e-01
2	100	-0.5	0.1	1.464e-03	1.352e-03	1.270e-03	9.99638e-01
2	100	-0.5	1.0	1.904e-03	1.904e-03	1.904e-03	9.98096e-01
2	100	0.5	0.1	1.120e-03	9.784e-04	9.459e-04	9.99762e-01
2	100	0.5	1.0	1.753e-02	1.250e-03	1.250e-03	9.98750e-01
2	100	0.99	0.1	7.874e-05	5.335e-05	4.507e-05	9.99995e-01
2	100	0.99	1.0	1.008e-02	2.500e-05	2.500e-05	9.99975e-01
2	200	-0.99	0.1	8.898e-04	7.217e-04	6.814e-04	9.99840e-01
2	200	-0.99	1.0	1.509e-02	8.403e-04	8.403e-04	9.99160e-01
2	200	-0.5	0.1	8.952e-04	7.415e-04	6.965e-04	9.99821e-01
2	200	-0.5	1.0	1.523e-02	9.446e-04	9.446e-04	9.99055e-01
2	200	0.5	0.1	8.429e-04	6.466e-04	6.208e-04	9.99881e-01
2	200	0.5	1.0	1.128e-02	6.250e-04	6.250e-04	9.99375e-01
2	200	0.99	0.1	7.186e-05	4.833e-05	4.043e-05	9.99998e-01
2	200	0.99	1.0	3.417e-04	2.325e-04	2.161e-04	9.99988e-01
2	500	-0.99	0.1	6.974e-04	4.662e-04	4.331e-04	9.99936e-01
2	500	-0.99	1.0	8.227e-03	3.348e-04	3.348e-04	9.99665e-01
2	500	-0.5	0.1	6.554e-04	4.443e-04	4.135e-04	9.99929e-01
2	500	-0.5	1.0	9.550e-03	3.761e-04	3.761e-04	9.99624e-01
2	500	0.5	0.1	7.124e-04	4.565e-04	4.281e-04	9.99953e-01
2	500	0.5	1.0	5.371e-03	3.637e-03	2.500e-04	9.99750e-01
2	500	0.99	0.1	6.848e-05	4.582e-05	3.807e-05	9.99999e-01
2	500	0.99	1.0	1.914e-04	1.201e-04	1.048e-04	9.99995e-01

TABLE A.1

Table showing the observed convergence rate versus the theoretical linear convergence parameter γ (as given by Theorem 2.1) for the LS-BOOST(ε) *algorithm for different values of ε, for the datasets Types 1 and 2. The bounds are observed to be fairly tight until the ℓ_1-norm of the regression coefficients are found to stabilize, which corresponds to the unregularized fit. Since statistically interesting solutions appear typically in the interior of the* LS-BOOST(ε) *path, the agreement between the observed and theoretical bounds are encouraging.*

Dataset	SNR	n	LS-Boost(ε) $\times 10^{-2}$	FS_ε $\times 10^{-2}$	FS_0 $\times 10^{-2}$	Stepwise $\times 10^{-2}$	Lasso $\times 10^{-2}$
Leukemia	1	72	65.9525 (1.8221)	66.7713 (1.8097)	68.1869 (1.4971)	74.5487 (2.6439)	68.3471 (1.584)
	3	72	35.4844 (1.1973)	35.5704 (0.898)	35.8385 (0.7165)	38.9429 (1.8030)	35.3673 (0.7924)
	10	72	13.5424 (0.4267)	13.3690 (0.3771)	13.6298 (0.3945)	14.8802 (0.4398)	13.4929 (0.4276)
Khan	1	63	22.3612 (1.1058)	22.6185 (1.0312)	22.9128 (1.1209)	25.2328 (1.0734)	23.5145 (1.2044)
	3	63	9.3988 (0.4856)	9.4851 (0.4721)	9.6571 (0.3813)	10.8495 (0.3627)	9.2339 (0.404)
	10	63	3.4061 (0.1272)	3.4036 (0.1397)	3.4812 (0.1093)	3.7986 (0.0914)	3.1118 (0.1229)
Eg-A, $\rho = 0.8$	1	50	53.1406 (1.5943)	52.1377 (1.6559)	53.6286 (1.4464)	60.3266 (1.9341)	53.7675 (1.2415)
	3	50	29.1960 (1.2555)	29.2814 (1.0487)	30.0654 (1.0066)	33.4318 (0.8780)	29.8000 (1.2662)
	10	50	12.2688 (0.3359)	12.0845 (0.3668)	12.6034 (0.5052)	15.9408 (0.7939)	12.4262 (0.3660)
Eg-A, $\rho = 0$	1	50	74.1228 (2.1494)	73.8503 (2.0983)	75.0705 (2.5759)	92.8779 (2.7025)	75.0852 (2.1039)
	3	50	38.1357 (2.7795)	40.0003 (1.8576)	41.0643 (1.5503)	43.9425 (3.9180)	41.4932 (2.2092)
	10	50	14.8867 (0.6994)	12.9090 (0.5553)	15.2174 (0.7086)	12.5502 (0.8256)	15.0877 (0.7142)

TABLE A.2

Table showing the prediction errors (in percentages) of different methods: LS-Boost(ε) , FS_ε *(both for different values of* ε*),* FS_0*, (forward) Stepwise regression, and* Lasso*. The numbers within parentheses denote standard errors. (The test errors were all standardized by the test error corresponding to the null model.)* LS-Boost(ε) , FS_ε *are found to exhibit similar statistical performances as the* Lasso*, in fact in some examples the boosting methods seem to be marginally better than* Lasso*. The predictive performance of the models were also found to be sensitive to the choice of the learning rate* ε*. For* FS_0 *and Stepwise we used the* `R` *package* `LARS` *[14] to compute the solutions. For all the cases,* $p = 500$*. For Eg-A, we took* $n = 50$*. Both* LS-Boost(ε) *and* FS_ε *were run for a few values of* ε *in the range* $[0.001 - 0.8]$ *– in all cases, the optimal models (see the text for details) for* LS-Boost(ε) *and* FS_ε *were achieved at a value of* ε *larger than its limiting version* $\varepsilon = 0+$*, thereby suggesting the sensitivity of the best predictive model to the learning rate* ε*.*

Real Data Example: Leukemia							
Method	n	p	SNR	Test Error	Sparsity	$\|\hat{\beta}^{\text{opt}}\|_1/\|\hat{\beta}^*\|_1$	$\delta/\delta_{\max}$
FS_ε	72	500	1	0.3431 (0.0087)	28	0.2339	-
R-$\text{FS}_{\varepsilon,\delta}$	72	500	1	0.3411 (0.0086)	25	0.1829	0.56
Lasso	72	500	1	0.3460 (0.0086)	30	1	0.11
FS_ε	72	500	10	0.0681 (0.0014)	67	0.7116	-
R-$\text{FS}_{\varepsilon,\delta}$	72	500	10	0.0659 (0.0014)	60	0.5323	0.56
Lasso	72	500	10	0.0677 (0.0015)	61	1	0.29

Synthetic Data Examples: Eg-B (SNR=1)							
Method	n	p	ρ	Test Error	Sparsity	$\|\hat{\beta}^{\text{opt}}\|_1/\|\hat{\beta}^*\|_1$	$\delta/\delta_{\max}$
FS_ε	50	500	0	0.19001 (0.0057)	56	0.9753	-
R-$\text{FS}_{\varepsilon,\delta}$	50	500	0	0.18692 (0.0057)	51	0.5386	0.71
Lasso	50	500	0	0.19163 (0.0059)	47	1	0.38
FS_ε	50	500	0.5	0.20902 (0.0057)	14	0.9171	-
R-$\text{FS}_{\varepsilon,\delta}$	50	500	0.5	0.20636 (0.0055)	10	0.1505	0.46
Lasso	50	500	0.5	0.21413 (0.0059)	13	1	0.07
FS_ε	50	500	0.9	0.05581 (0.0015)	4	0.9739	-
R-$\text{FS}_{\varepsilon,\delta}$	50	500	0.9	0.05507 (0.0015)	4	0.0446	0.63
Lasso	50	500	0.9	0.09137 (0.0025)	5	1	0.04

Table A.3

Table showing the statistical properties of R-$FS_{\varepsilon,\delta}$ as compared to Lasso *and FS_ε. Both R-$FS_{\varepsilon,\delta}$ and FS_ε use $\varepsilon = 0.001$. The model that achieved the best predictive performance (test-error) corresponds to $\hat{\beta}^{opt}$. The limiting model (as the number of boosting iterations is taken to be infinitely large) for each method is denoted by $\hat{\beta}^*$. "Sparsity" denotes the number of coefficients in $\hat{\beta}^{opt}$ larger than 10^{-5} in absolute value. $\delta_{\max}$ is the ℓ_1-norm of the least squares solution with minimal ℓ_1-norm. Both R-$FS_{\varepsilon,\delta}$ and* Lasso *were run for a few δ values of the form $\eta\delta_{\max}$, where η takes on twenty values in $[0.01, 0.8]$. For the real data instances, R-$FS_{\varepsilon,\delta}$ and* Lasso *were run for a maximum of 30,000 iterations, and FS_ε was run for 20,000 iterations. For the synthetic examples, all methods were run for a maximum of 10,000 iterations. The best models for R-$FS_{\varepsilon,\delta}$ and FS_ε were all obtained in the interior of the path. The best models delivered by R-$FS_{\varepsilon,\delta}$ are seen to be more sparse and have better predictive performance than the best models obtained by FS_ε. The performances of* Lasso *and R-$FS_{\varepsilon,\delta}$ are found to be quite similar, though in some cases R-$FS_{\varepsilon,\delta}$ is seen to be at an advantage in terms of better predictive accuracy. The test errors were all standardized by the test error corresponding to the null model.*