The Annals of Statistics
2017, Vol. 45, No. 6, 2328–2364
https://doi.org/10.1214/16-AOS1505

A NEW PERSPECTIVE ON BOOSTING IN LINEAR REGRESSION VIA SUBGRADIENT OPTIMIZATION AND RELATIVES

BY ROBERT M. FREUND[1,*], PAUL GRIGAS[2,†] AND RAHUL MAZUMDER[3,*]

Massachusetts Institute of Technology[*] *and University of California, Berkeley*[†]

We analyze boosting algorithms [*Ann. Statist.* **29** (2001) 1189–1232; *Ann. Statist.* **28** (2000) 337–407; *Ann. Statist.* **32** (2004) 407–499] in linear regression from a new perspective: that of modern first-order methods in convex optimization. We show that classic boosting algorithms in linear regression, namely the incremental forward stagewise algorithm (FS_ε) and least squares boosting [LS-BOOST(ε)], can be viewed as subgradient descent to minimize the loss function defined as the maximum absolute correlation between the features and residuals. We also propose a minor modification of FS_ε that yields an algorithm for the LASSO, and that may be easily extended to an algorithm that computes the LASSO path for different values of the regularization parameter. Furthermore, we show that these new algorithms for the LASSO may also be interpreted as the same master algorithm (subgradient descent), applied to a regularized version of the maximum absolute correlation loss function. We derive novel, comprehensive computational guarantees for several boosting algorithms in linear regression (including LS-BOOST(ε) and FS_ε) by using techniques of first-order methods in convex optimization. Our computational guarantees inform us about the statistical properties of boosting algorithms. In particular, they provide, for the first time, a precise theoretical description of the amount of data-fidelity and regularization imparted by running a boosting algorithm with a prespecified learning rate for a fixed but arbitrary number of iterations, for *any* dataset.

1. Introduction.

Boosting [17, 21, 31, 43, 44] is an extremely successful and popular supervised learning method that combines multiple weak[4] learners into a powerful “committee.” AdaBoost [18, 31, 44], one of the earliest boosting algorithms developed in the context of classification, may be viewed as an optimization algorithm: a form of gradient descent in a certain function space [4, 5]. In an influ-

Received December 2015; revised August 2016.

[1] Supported by AFOSR Grant No. FA9550-11-1-0141 and the MIT-Chile-Pontificia Universidad Catolica de Chile Seed Fund.

[2] Supported in part by NSF Graduate Research Fellowship and the MIT-Chile-Pontificia Universidad Catolica de Chile Seed Fund.

[3] Supported by ONR Grant N000141512342 and a grant from the Betty Moore-Sloan foundation.

MSC2010 subject classifications. Primary 62J05, 62J07k; secondary 90C25.

Key words and phrases. Boosting, forward stagewise regression, linear regression, bias-variance tradeoff, shrinkage, convex optimization, computational guarantees.

[4] This term originates in the context of boosting for classification, where a “weak” classifier is slightly better than random guessing.

ential paper, [21] nicely interpreted boosting methods used in classification problems as instances of stagewise additive modeling [32]. Friedman [23] provided a unified view of stagewise additive modeling and steepest descent minimization methods in function space to explain boosting methods. For related perspectives from the machine learning community, we refer the reader to [35, 40] and the references therein.

An important instantiation of boosting, and the topic of the present paper, is its application in linear regression [6, 7, 23, 31]. We use the usual notation with model matrix $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_p] \in \mathbb{R}^{n \times p}$, response vector $\mathbf{y} \in \mathbb{R}^{n \times 1}$, and regression coefficients $\beta \in \mathbb{R}^p$. We assume that the $\mathbf{X}_i$'s have been centered to have zero mean and unit ℓ_2 norm and $\mathbf{y}$ is also centered to have zero mean. For a regression coefficient β, the predicted value of the response is $\mathbf{X}\beta$ and $r = \mathbf{y} - \mathbf{X}\beta$ denotes the residuals.

Boosting and implicit regularization. We begin our study with a popular algorithm: Least Squares Boosting—also known as LS-BOOST(ε) [23]—which is formally described herein in Section 2. LS-BOOST(ε) has been studied by several authors [6–8, 22, 34]. Starting from the null model $\hat{\beta}^0 = 0$, at the kth iteration LS-BOOST(ε) determines the covariate index j_k with the best univariate fit to the current residuals $\hat{r}^k = \mathbf{y} - \mathbf{X}\hat{\beta}^k$:

$$j_k \in \mathop{\arg\min}_{1 \le m \le p} \sum_{i=1}^n (\hat{r}_i^k - x_{im}\tilde{u}_m)^2 \qquad \text{where, } \tilde{u}_m = \mathop{\arg\min}_{u \in \mathbb{R}} \left(\sum_{i=1}^n (\hat{r}_i^k - x_{im}u)^2 \right).$$

The algorithm then updates the j_kth regression coefficient with a shrinkage factor $\varepsilon > 0$: $\hat{\beta}_{j_k}^{k+1} \leftarrow \hat{\beta}_{j_k}^k + \varepsilon \tilde{u}_{j_k}$, with all other regression coefficients unchanged. A close cousin of the LS-BOOST(ε) algorithm is the Incremental Forward Stagewise algorithm [12, 31]—also known as FS$_\varepsilon$—which is formally described herein in Section 3. FS$_\varepsilon$ chooses the covariate most correlated (in absolute value) with the residual $\hat{r}^k$ and performs the update

$$\begin{aligned} &\hat{\beta}_{j_k}^{k+1} \leftarrow \hat{\beta}_{j_k}^k + \varepsilon \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}), \\ &\hat{\beta}_j^{k+1} \leftarrow \hat{\beta}_j^k, \qquad j \neq j_k, \qquad \text{where } j_k \in \mathop{\arg\max}_{j \in \{1,\dots,p\}} |(\hat{r}^k)^T \mathbf{X}_j|. \end{aligned}$$

[Since the covariates are standardized, both LS-BOOST(ε) and FS$_\varepsilon$ lead to the same variable selection for a given $\hat{r}^k$.] LS-BOOST(ε) and FS$_\varepsilon$ have curious similarities but subtle differences as we characterize in Section 3 (see also [7]). In both algorithms, the shrinkage factor ε, also known as the learning rate, counterbalances the greedy selection strategy of choosing the *best* covariate. Qualitatively speaking, a greedy fitting procedure may overfit quickly—a small value of ε slows down the learning rate as compared to a larger choice of ε (that is fine-tuned to minimize the training error) and leads to slower overfitting. With a small ε it is possible to explore a larger class of models, with varying degrees of shrinkage—this often leads

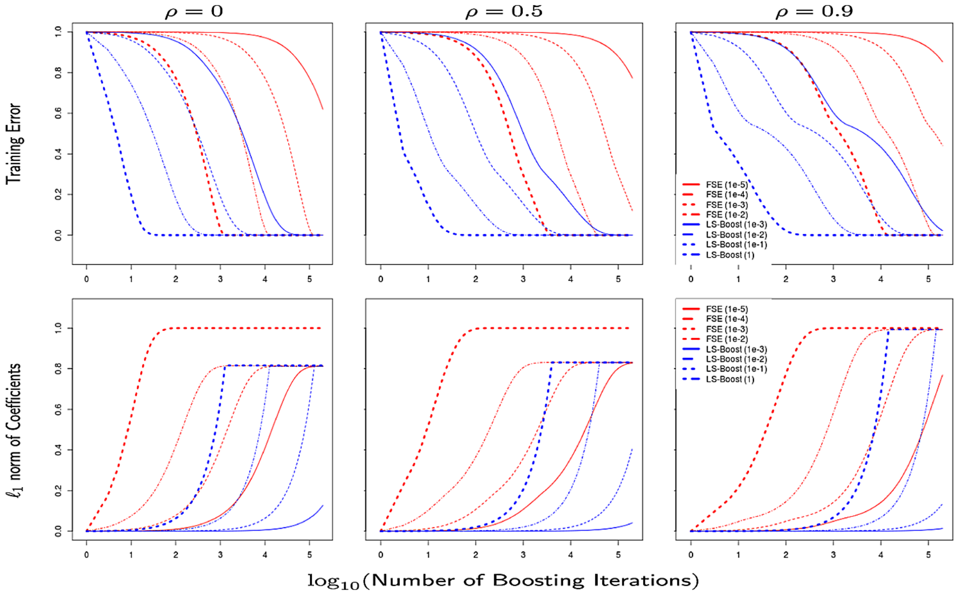

FIG. 1. *Evolution of* LS-BOOST(ε) *and* FS$_\varepsilon$ *versus iterations* (*in log-scale*), *for a synthetic dataset with* $n = 50$, $p = 500$: *the covariates are Gaussian with pairwise correlations* ρ, *the true* β *has ten nonzeros with* $\beta_i = 1, i \leq 10$ *and SNR* $= 1$. *Different* ρ *and* ε *values are considered.* [*Top Row*] *Shows the training errors for different learning rates*, [*Bottom Row*] *shows the* ℓ_1 *norm of coefficients produced by the algorithms for different learning rates* (*the y-axis values have been re-scaled to lie in* [0, 1]).

to models with better predictive power [23]. Let M denote the number of boosting iterations. Then both M and ε (the shrinkage factor) together control the training error and the amount of shrinkage. We refer the reader to Figure 1, which illustrates the evolution of the algorithmic properties of the LS-BOOST(ε) algorithm as a function of k and ε. Up until now, as pointed out by [31], the understanding of the tradeoff between regularization and data-fidelity for these boosting methods has been rather qualitative. One of the contributions of this paper is a precise quantification of this tradeoff. In Sections 2 and 3, we will derive comprehensive computational guarantees for these algorithms which provide a formal description of how M and ε control the amount of training error and regularization in FS$_\varepsilon$ and LS-BOOST(ε), as well as precise bounds on their tradeoffs between regularization and data-fidelity. Furthermore, in Section 3.3 we will provide a unified treatment of LS-BOOST(ε), FS$_\varepsilon$, and a generalization with adaptive step-sizes FS$_{\varepsilon_k}$—wherein we will show that all of these methods can be viewed as special instances of (convex) subgradient optimization.

Both LS-BOOST(ε) and FS$_\varepsilon$ may be interpreted as "cautious" versions of the classical Forward Selection or Forward Stepwise regression [36, 50]. This algorithm identifies the variable most correlated (in absolute value) with the current residual, includes it in the model, and updates the *joint least squares* fit based on the current set of predictors—this update strategy makes stepwise regression *aggressive*, and hence different from FS_ε and LS-BOOST(ε).

LASSO *and explicit regularization.* All of the algorithms described above impart regularization in an implicit fashion through the choice of ε and M. In contrast, let us consider the constraint version of the LASSO [46]:

$$\text{LASSO}: \quad L^*_{n,\delta} := \min_{\beta} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 \qquad (1.1)$$
$$\text{s.t.} \quad \|\beta\|_1 \le \delta,$$

with regularization parameter $\delta \ge 0$. The nature of regularization via the LASSO is explicit—it is set up to find the best least squares solution subject to a constraint on the ℓ_1 norm of the regression coefficients. We let $\{\hat{\beta}_\delta\}$ denote the path of solutions of (1.1) for all $\delta \ge 0$, otherwise known as the "LASSO path."

Boosting and LASSO: *Path properties.* Although LASSO and the above boosting methods originate from different perspectives, there are curious similarities between the two as is nicely explored in [12, 30, 31]. For certain datasets, the coefficient profiles[5] of LASSO and FS_0 (defined to be the limiting case of the FS_ε algorithm as $\varepsilon \to 0+$) are exactly the same (see Figure 2, top panel) [31]. However, they are different in general (Figure 2, bottom panel). Efforts to understand the FS_ε algorithm paved the way for the Least Angle Regression algorithm—also known as LAR [12] (see also [31]). The LAR is a unified framework: one instance of LAR computes the LASSO path and another delivers a coefficient profile for FS_0.

The similarities between the LASSO and boosting coefficient profiles motivate us to develop a minor modification of boosting that generates the LASSO path, which we will accomplish in Sections 4 and 5. In a different line of approach, [51] describes BLASSO, a modification of the FS_ε algorithm with the inclusion of additional "backward steps" so that the resultant coefficient profile mimics the LASSO path.

Boosting and LASSO: *Computation.* While solving the LASSO is computationally very attractive for small to moderate-sized datasets, efficient implementations of boosting (e.g., FS_ε) are equally efficient[6] [19]. With regard to LAR, for example, computing the FS_0 and LASSO profiles have comparable cost. In examples

[5] By a coefficient profile, we mean the map $\lambda \mapsto \hat{\beta}_\lambda$ where, $\lambda \in \Lambda$ indexes a family of coefficients $\hat{\beta}_\lambda$. For example, the regression coefficients delivered by FS_0 delivers a coefficient profile as a function of their ℓ_1-norms.

[6] Friedman [19] shows that an optimized implementation of FS_ε leads to an entire solution path for a problem with 10,000 features and 200 samples in 0.5 seconds, whereas solving the LASSO can take up to 6–8 times longer to compute a path of solutions.

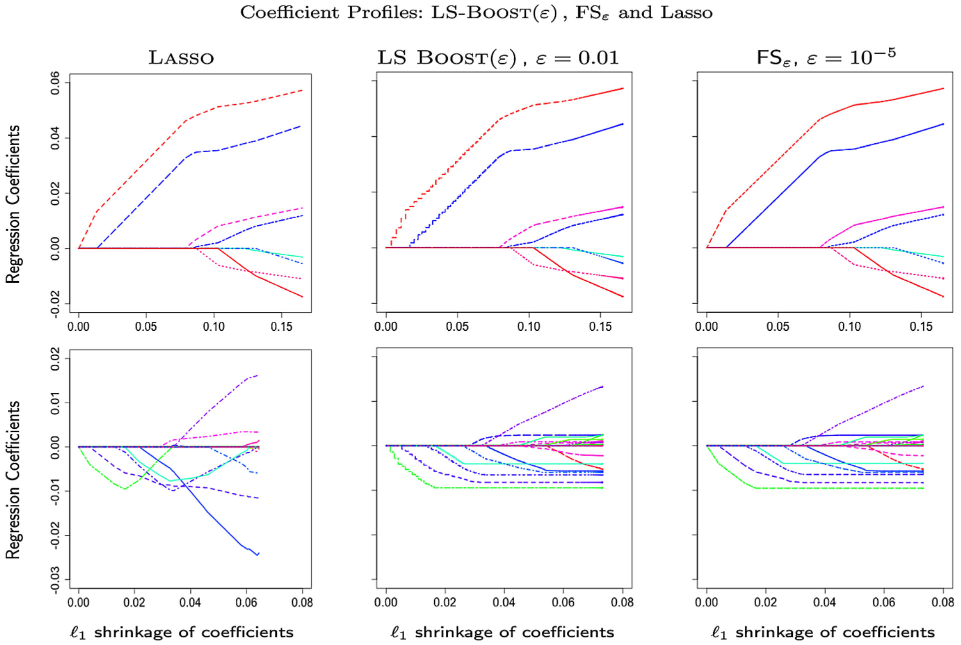

FIG. 2. *Coefficient Profiles of several algorithms as as function of the ℓ_1-norm of coefficients on different datasets.* [*Top Panel*] *Prostate Cancer dataset described in Section* 6 *with $n = 98$ and $p = 8$. All profiles look similar.* [*Bottom Panel*] *A subset of samples of the Prostate Cancer dataset with $n = 10$; we also included all second order interactions to get $p = 44$. The coefficient profile of* LASSO *is seen to be different from* FS$_\varepsilon$ *and* LS-BOOST(ε). *Figure* A.1 [14] *shows training error vis-à-vis the ℓ_1-shrinkage of the models for the same data.*

where the number of possible features is extremely large or possibly infinite [41, 42] a boosting algorithm like FS$_\varepsilon$ is computationally more attractive than solving the LASSO.

Subgradient optimization as a unifying viewpoint of boosting and the LASSO. In spite of the various nice perspectives on FS$_\varepsilon$ and its connections to the LASSO as described above, the present understanding about the relationships between the LASSO path, FS$_\varepsilon$, and LS-BOOST(ε) for arbitrary datasets and $\varepsilon > 0$ has nevertheless been fairly limited. A chief goal of this paper is to contribute some substantial further understanding of the relationship between these objects. Somewhat like the LAR algorithm can be viewed as a master algorithm with special instances yielding the LASSO path and FS$_0$, we establish herein that FS$_\varepsilon$ and LS-BOOST(ε) can be viewed as special instances of one grand algorithm: the subgradient descent method (of convex optimization) applied to the following parametric class of optimization problems:

$$(1.2)\quad P_\delta : \underset{r}{\text{minimize}}\ \|\mathbf{X}^T r\|_\infty + \frac{1}{2\delta}\|r - \mathbf{y}\|_2^2 \qquad \text{where } r = \mathbf{y} - \mathbf{X}\beta \text{ for some } \beta,$$

where $\delta \in (0, \infty]$ is a regularization parameter. Here, the first term is the maximum absolute inner-product between the features and residuals, and the second term acts as a regularizer by penalizing residuals that are far from the response. As we describe herein in Section 4, Problem (1.2) is intimately connected to the LASSO problem (1.1) through duality. We then show that the subgradient descent algorithm applied to Problem (1.2) leads to a new boosting algorithm—dubbed R-FS$_{\varepsilon,\delta}$ (for Regularized incremental Forward Stagewise regression)—that solves the LASSO, and that is almost identical to FS$_\varepsilon$ except that it includes an additional simple re-scaling of the coefficients. Section 4 develops a variety of properties of the new algorithm R-FS$_{\varepsilon,\delta}$ related to regularization, data-fidelity, etc. And in Section 5 we present an adaptive version PATH-R-FS$_\varepsilon$ of R-FS$_{\varepsilon,\delta}$ which approximates the LASSO path, with associated approximation guarantees as well. We also observe empirically that R-FS$_{\varepsilon,\delta}$ has statistical properties similar to LASSO and FS$_\varepsilon$, and it often leads to models that are more sparse than FS$_\varepsilon$.

Summary of contributions. We view the contributions of this paper as falling into two main streams as follows:

- *Current Boosting Methods and the Subgradient Descent Method.* We show that both LS-BOOST(ε) and FS$_\varepsilon$ are instances of the subgradient descent method of convex optimization applied to the problem of minimizing the maximum absolute correlation between features and residuals. This leads to the first-ever computational guarantees for the behavior of these boosting methods as well as precise bounds on the tradeoffs between regularization and data-fidelity for these methods, which hold for *any* dataset. See Theorem 2.1, Proposition 3.2 and Theorem 3.1.
- *New Boosting Method* R-FS$_{\varepsilon,\delta}$ *connecting* FS$_\varepsilon$ *and the* LASSO. We present a new boosting algorithm named R-FS$_{\varepsilon,\delta}$—for Regularized Forward Stagewise regression—that is identical to FS$_\varepsilon$ except for a simple rescaling of the coefficients at each iteration, and that specializes to FS$_\varepsilon$ as well as to an algorithm for solving the LASSO, depending on the choice of δ. We present computational guarantees for convergence of R-FS$_{\varepsilon,\delta}$ to LASSO solutions, and we present a path version of the algorithm that computes an approximation of the LASSO path with associated approximation bounds. See Proposition 4.1, Theorem 4.1, Theorem 5.1 and Corollary 5.1.

Organization of the paper. The paper is organized as follows. In Section 2, we analyze the convergence behavior of the LS-BOOST(ε) algorithm. In Section 3, we present a unifying algorithmic framework for FS$_\varepsilon$, FS$_{\varepsilon_k}$ and LS-BOOST(ε) as subgradient descent. In Section 4, we introduce R-FS$_{\varepsilon,\delta}$ as a boosting algorithm naturally associated with Problem (1.2). In Section 5, we further expand R-FS$_{\varepsilon,\delta}$ into a method for computing approximate solutions of the LASSO path. Section 6 contains computational experiments. To improve readability, most of the technical details are placed in the supplementary section [14].

1.1. *Notation.* For a vector $x \in \mathbb{R}^m$, we use x_i to denote the ith coordinate of x. We use superscripts to index vectors in a sequence $\{x^k\}$. Let e_j denote the jth unit vector in $\mathbb{R}^m$, and let $e = (1, \dots, 1)$ denote the vector of ones. Let $\| \cdot \|_q$ denote the ℓ_q norm for $q \in [1, \infty]$ with unit ball B_q, and let $\|v\|_0$ denote the number of nonzero coefficients of the vector v. For $A \in \mathbb{R}^{m \times n}$, let $\|A\|_{q_1,q_2} := \max_{x : \|x\|_{q_1} \le 1} \|Ax\|_{q_2}$ be the operator norm. In particular, $\|A\|_{1,2} = \max(\|A_1\|_2, \dots, \|A_n\|_2)$ is the maximum ℓ_2 norm of the columns of A. For a scalar α, $\operatorname{sgn}(\alpha)$ denotes the sign of α. The notation "$\tilde{v} \leftarrow \arg\max_{v \in S}\{f(v)\}$" denotes assigning $\tilde{v}$ to be any optimal solution of the problem $\max_{v \in S}\{f(v)\}$. For a convex set P, let $\Pi_P(\cdot)$ denote the Euclidean projection operator onto P, namely $\Pi_P(\bar{x}) := \arg\min_{x \in P} \|x - \bar{x}\|_2$. Let $\partial f(\cdot)$ denote the subdifferential operator of a convex function $f(\cdot)$. If $Q \neq 0$ is a symmetric positive semidefinite matrix, let $\lambda_{\max}(Q)$, $\lambda_{\min}(Q)$, and $\lambda_{\text{pmin}}(Q)$ denote the largest, smallest and smallest nonzero (and hence positive) eigenvalue of Q, respectively. We use the notation $L_n(\beta) := \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2$ for the least squares loss function.

2. LS-Boost(ε): Computational guarantees and statistical implications.

Roadmap. We begin our formal study by examining the LS-Boost(ε) algorithm. We study the rate at which the coefficients generated by LS-Boost(ε) converge to the set of unregularized least square solutions. This characterizes the amount of data-fidelity as a function of the number of iterations and ε. In particular, we show (global) linear convergence of the regression coefficients to the set of least squares coefficients, with similar convergence rates derived for the prediction estimates and the boosting training errors delivered by LS-Boost(ε). We also present bounds on the shrinkage of the regression coefficients $\hat{\beta}^k$ as a function of k and ε, thereby describing how the amount of shrinkage of the regression coefficients changes as a function of the number of iterations k.

We present below a formal description of LS-Boost(ε) following [23]:

Algorithm: Least Squares Boosting—LS-Boost(ε)

Fix the learning rate $\varepsilon > 0$, the number of iterations M and initialize $\hat{\beta}^0 = 0$ and $\hat{r}^0 = \mathbf{y}$.

1. For $0 \le k \le M$, do the following:

2. Select the covariate index j_k and $\tilde{u}_{j_k}$ as follows:

$$j_k \in \arg\min_{1 \le m \le p} \sum_{i=1}^{n} (\hat{r}_i^k - x_{im}\tilde{u}_m)^2, \qquad \text{where, } \tilde{u}_m = \arg\min_{u \in \mathbb{R}} \left(\sum_{i=1}^{n} (\hat{r}_i^k - x_{im}u)^2 \right),$$

for $m = 1, \dots, p$.

3. Update the regression coefficients and residuals as

$$\hat{\beta}_{j_k}^{k+1} \leftarrow \hat{\beta}_{j_k}^k + \varepsilon \tilde{u}_{j_k}, \qquad \hat{\beta}_j^{k+1} \leftarrow \hat{\beta}_j^k, \qquad j \neq j_k \quad \text{and}$$

$$\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon \mathbf{X}_{j_k} \tilde{u}_{j_k}.$$

A special instance of the LS-BOOST(ε) algorithm with $\varepsilon = 1$ is known as LS-BOOST [23] or Forward Stagewise [31]—it is essentially a method of repeated simple least squares fitting of the residuals [7]. In the signal processing literature, LS-BOOST is known as Matching Pursuit [34], a scheme used to approximate a signal (herein, response) as a sparse linear sum of dictionary elements (herein, features). In words, the LS-BOOST algorithm at the kth iteration determines the covariate index j_k resulting in the maximal decrease in the univariate regression fit to the current residuals. If $\mathbf{X}_{j_k}\tilde{u}_{j_k}$ denotes the *best* univariate fit for the current residuals, LS-BOOST updates the residuals: $\hat{r}^{k+1} \leftarrow \hat{r}^k - \mathbf{X}_{j_k}\tilde{u}_{j_k}$ and the j_kth regression coefficient: $\hat{\beta}^{k+1}_{j_k} \leftarrow \hat{\beta}^k_{j_k} + \tilde{u}_{j_k}$, in an attempt to minimize the training error. LS-BOOST(ε) has old roots—as noted by [7], LS-BOOST with $M = 2$ is known as "twicing," a method proposed by Tukey [48]. The papers [6–8] present very interesting perspectives on LS-BOOST(ε), where they refer to the algorithm as $L2$-BOOST. Bühlmann and Hothorn [7] also obtains approximate expressions for the effective degrees of freedom of the $L2$-BOOST algorithm. LS-BOOST(ε) is also closely related to Friedman's MART algorithm [22]. LS-BOOST(ε) is a slow-learning variant of LS-BOOST, which diminishes the fast and greedy learning style of LS-BOOST with an additional damping factor of ε, which consequently leads to a richer family of models, as we study in this section.

2.1. *Computational guarantees and intuition.* We first review some useful properties associated with the familiar least squares regression problem

$$\begin{aligned} \text{LS}: \quad L_n^* := \min_\beta L_n(\beta) &:= \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 \\ \text{s.t.} \quad \beta &\in \mathbb{R}^p, \end{aligned} \tag{2.1}$$

where $L_n(\cdot)$ is the least squares loss, whose gradient is

$$\nabla L_n(\beta) = -\frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = -\frac{1}{n}\mathbf{X}^T r, \tag{2.2}$$

where $r = \mathbf{y} - \mathbf{X}\beta$ is the vector of residuals corresponding to the regression coefficients β. It follows that β is a least-squares solution of LS if and only if $\nabla L_n(\beta) = 0$, which leads to the well-known normal equations

$$0 = -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = -\mathbf{X}^T r. \tag{2.3}$$

It also holds that

$$n \cdot \|\nabla L_n(\beta)\|_\infty = \|\mathbf{X}^T r\|_\infty = \max_{j \in \{1,\ldots,p\}}\{|r^T \mathbf{X}_j|\}. \tag{2.4}$$

The following theorem describes precise computational guarantees for LS-BOOST(ε): linear convergence of LS-BOOST(ε) with respect to (2.1), and bounds on the ℓ_1 shrinkage of the coefficients produced. Note that the theorem uses the quantity $\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})$ which denotes the smallest nonzero (and hence positive) eigenvalue of $\mathbf{X}^T\mathbf{X}$.

THEOREM 2.1 [Linear Convergence of LS-BOOST(ε) for Least Squares]. *Consider the* LS-BOOST(ε) *algorithm with learning rate* $\varepsilon \in (0,1]$, *and define the linear convergence rate coefficient* γ:

$$\gamma := \left(1 - \frac{\varepsilon(2-\varepsilon)\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})}{4p}\right) < 1. \tag{2.5}$$

For all $k \geq 0$, *the following bounds hold* ($\hat{\beta}_{\mathrm{LS}}$ *denotes a least squares solution*):

(i) (*training error*): $L_n(\hat{\beta}^k) - L_n^* \leq \frac{1}{2n}\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2^2 \cdot \gamma^k$;

(ii) (*regression coefficients*): *there exists a least squares solution* $\hat{\beta}_{LS}^k$ *such that*

$$\|\hat{\beta}^k - \hat{\beta}_{LS}^k\|_2 \leq \frac{\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2}{\sqrt{\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})}} \cdot \gamma^{k/2};$$

(iii) (*predictions*): *for every least-squares solution* $\hat{\beta}_{\mathrm{LS}}$ *it holds that*

$$\|\mathbf{X}\hat{\beta}^k - \mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2 \leq \|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2 \cdot \gamma^{k/2};$$

(iv) (*gradient norm/correlation values*): $\|\nabla L_n(\hat{\beta}^k)\|_\infty = \frac{1}{n}\|\mathbf{X}^T\hat{r}^k\|_\infty \leq \frac{1}{n}\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2 \cdot \gamma^{k/2}$;

(v) (ℓ_1-*shrinkage of coefficients*):

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

(vi) (*sparsity of coefficients*): $\|\hat{\beta}^k\|_0 \leq k$.

Before remarking on the various parts of Theorem 2.1, we first discuss the quantity γ defined in (2.5), which is called the linear convergence rate coefficient. We can write $\gamma = 1 - \frac{\varepsilon(2-\varepsilon)}{4\kappa(\mathbf{X}^T\mathbf{X})}$ where $\kappa(\mathbf{X}^T\mathbf{X})$ is defined to be the ratio $\kappa(\mathbf{X}^T\mathbf{X}) := \frac{p}{\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})}$. Note that $\kappa(\mathbf{X}^T\mathbf{X}) \in [1,\infty)$. To see this, let $\tilde{\beta}$ be an eigenvector associated with the largest eigenvalue of $\mathbf{X}^T\mathbf{X}$, then

$$0 < \lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X}) \leq \lambda_{\max}(\mathbf{X}^T\mathbf{X}) = \frac{\|\mathbf{X}\tilde{\beta}\|_2^2}{\|\tilde{\beta}\|_2^2} \leq \frac{\|\mathbf{X}\|_{1,2}^2\|\tilde{\beta}\|_1^2}{\|\tilde{\beta}\|_2^2} \leq p, \tag{2.6}$$

where the last inequality uses our assumption that the columns of $\mathbf{X}$ have been normalized (whereby $\|\mathbf{X}\|_{1,2} = 1$), and the fact that $\|\tilde{\beta}\|_1 \leq \sqrt{p}\|\tilde{\beta}\|_2$. This then implies that $\gamma \in [0.75, 1.0)$—independent of any assumption on the dataset—and most importantly it holds that $\gamma < 1$.

Let us now make the following immediate remarks on Theorem 2.1:

- The bounds in parts (i)–(iv) state that the training errors, regression coefficients, predictions, and correlation values produced by LS-BOOST(ε) converge linearly (also known as geometric or exponential convergence) to their least squares counterparts: they decrease by at least the constant multiplicative factor $\gamma < 1$ for part (i), and by $\sqrt{\gamma}$ for parts (ii)–(iv), at every iteration. The bounds go to zero at this linear rate as $k \to \infty$.
- The computational guarantees in parts (i)–(vi) provide characterizations of the data-fidelity and shrinkage of the LS-BOOST(ε) algorithm for any given specifications of the learning rate ε and the number of boosting iterations k. Moreover, the quantities appearing in the bounds can be computed from simple characteristics of the data that can be obtained a priori without even running the boosting algorithm. (And indeed, one can even substitute $\|\mathbf{y}\|_2$ in place of $\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2$ throughout the bounds if desired since $\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2 \leq \|\mathbf{y}\|_2$.)

Some intuition behind Theorem 2.1. Let us now study the LS-BOOST(ε) algorithm and build intuition regarding its progress with respect to solving the unconstrained least squares problem (2.1), which will inform the results in Theorem 2.1. Since the predictors are all standardized to have unit ℓ_2 norm, it follows that the coefficient index j_k and corresponding step-size $\tilde{u}_{j_k}$ selected by LS-BOOST(ε) satisfy

$$j_k \in \underset{j\in\{1,\dots,p\}}{\arg\max} \left|(\hat{r}^k)^T \mathbf{X}_j\right| \quad \text{and} \quad \tilde{u}_{j_k} = (\hat{r}^k)^T \mathbf{X}_{j_k}. \tag{2.7}$$

Combining (2.4) and (2.7), we see that

$$|\tilde{u}_{j_k}| = \left|(\hat{r}^k)^T \mathbf{X}_{j_k}\right| = n \cdot \|\nabla L_n(\hat{\beta}^k)\|_\infty. \tag{2.8}$$

Using the formula for $\tilde{u}_{j_k}$ in (2.7), we have the following convenient way to express the change in residuals at each iteration of LS-BOOST(ε):

$$\hat{r}^{k+1} = \hat{r}^k - \varepsilon\big((\hat{r}^k)^T \mathbf{X}_{j_k}\big)\mathbf{X}_{j_k}. \tag{2.9}$$

Intuitively, since (2.9) expresses $\hat{r}^{k+1}$ as the difference of two correlated variables, $\hat{r}^k$ and $\operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k})\mathbf{X}_{j_k}$, we expect the squared ℓ_2 norm of $\hat{r}^{k+1}$ to be smaller than that of $\hat{r}^k$. On the other hand, as we will see from (3.2), convergence of the residuals is ensured by the dependence of the change in residuals on $|(\hat{r}^k)^T \mathbf{X}_{j_k}|$, which goes to 0 as we approach a least squares solution. In the proof of Theorem 2.1 in Section A.2.2 (in [14]), we make this intuition precise by using (2.9) to quantify the amount of decrease in the least squares objective function at each iteration of LS-BOOST(ε). The final ingredient of the proof uses properties of convex quadratic functions (Section A.2.1 in [14]) to relate the exact amount of the decrease from iteration k to $k+1$ to the current optimality gap $L_n(\hat{\beta}^k) - L_n^*$, which yields the following strong linear convergence property:

$$L_n(\hat{\beta}^{k+1}) - L_n^* \leq \gamma \cdot \big(L_n(\hat{\beta}^k) - L_n^*\big). \tag{2.10}$$

The above states that the training error gap decreases at each iteration by at least the multiplicative factor of γ, and clearly implies item (i) of Theorem 2.1. The bounds in Theorem 2.1 are indeed tight, as addressed in Section A.2.8 [14].

Comments on the global linear convergence rate in Theorem 2.1. The global linear convergence of LS-BOOST(ε) proved in Theorem 2.1, while novel, is not at odds with the present understanding of such convergence for optimization problems. One can view LS-BOOST(ε) as performing steepest descent optimization steps with respect to the ℓ_1 norm unit ball (rather than the ℓ_2 norm unit ball which is the canonical version of the steepest descent method; see [39]). It is known [39] that canonical steepest decent exhibits global linear convergence for convex quadratic optimization so long as the Hessian matrix Q of the quadratic objective function is positive definite, that is, $\lambda_{\min}(Q) > 0$. And for the least squares loss function $Q = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, which yields the condition that $\lambda_{\min}(\mathbf{X}^T\mathbf{X}) > 0$. As discussed in [3], this result extends to other norms defining steepest descent as well. Hence, what is modestly surprising herein is not the linear convergence *per se*, but rather that LS-BOOST(ε) exhibits global linear convergence even when $\lambda_{\min}(\mathbf{X}^T\mathbf{X}) = 0$, that is, even when $\mathbf{X}$ does not have full column rank [essentially replacing $\lambda_{\min}(\mathbf{X}^T\mathbf{X})$ with $\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})$ in our analysis]. This derives specifically from the structure of the least squares loss function, whose function values (and whose gradient) are invariant in the null space of $\mathbf{X}$, that is, $L_n(\beta + d) = L_n(\beta)$ for all d satisfying $\mathbf{X}d = 0$, and is thus rendered "immune" to changes in β in the null space of $\mathbf{X}^T\mathbf{X}$.

2.2. *Statistical insights from the computational guarantees.* Note that in most noisy problems, the limiting least squares solution is statistically less interesting than an estimate obtained in the interior of the boosting profile, since the latter typically corresponds to a model with better bias-variance tradeoff. We thus caution the reader that the bounds in Theorem 2.1 should *not* be merely interpreted as statements about how rapidly the boosting iterations reach the least squares fit. We rather intend for these bounds to inform us about the *evolution* of the training errors and the amount of shrinkage of the coefficients as the LS-BOOST(ε) algorithm progresses and when k is at most moderately large. When the training errors are paired with the profile of the ℓ_1-shrinkage values of the regression coefficients, they lead to the ordered pairs

$$\left(\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\hat{\beta}^k\|_2^2, \|\hat{\beta}^k\|_1\right), \qquad k \geq 1, \tag{2.11}$$

which describes the data-fidelity and ℓ_1-shrinkage tradeoff as a function of k, for the given learning rate $\varepsilon > 0$. This profile is described in Figure A.1 in Section A.1.1 [14] for several data instances. The bounds in Theorem 2.1 provide estimates for the two components of the ordered pair (2.11), and they can be computed

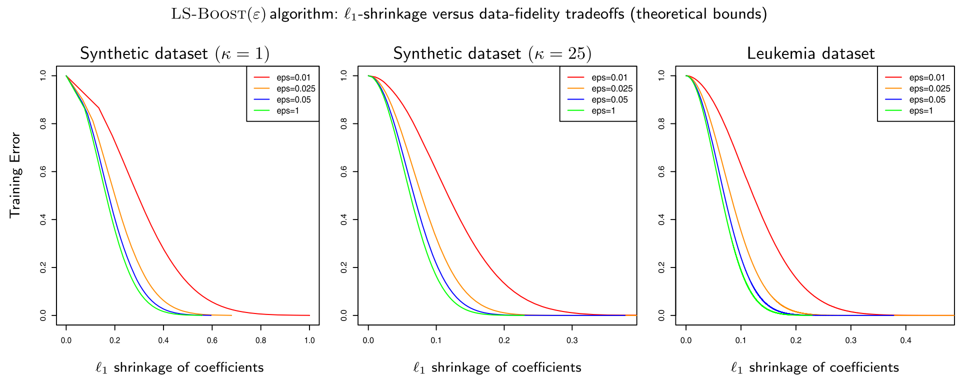

FIG. 3. *Figure showing profiles of ℓ_1 shrinkage of the regression coefficients versus training error for the* LS-BOOST(ε) *algorithm, for different values of the learning rate* ε (*denoted by the moniker "eps" in the legend*). *The profiles have been obtained from the computational bounds in Theorem* 2.1. *The left and middle panels correspond to synthetic values of the ratio* $\kappa = \frac{p}{\lambda_{\text{pmin}}}$, *and for the right panel profiles the value of* κ (*here*, $\kappa = 270.05$) *is extracted from the Leukemia dataset, described in Section* 6. *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.*

prior to running the boosting algorithm. For simplicity, let us use the following crude estimate:

$$\ell_k := \min\left\{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2\sqrt{\frac{k\varepsilon}{2-\varepsilon}}, \frac{\varepsilon\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2}{1-\sqrt{\gamma}}\left(1-\gamma^{\frac{k}{2}}\right)\right\},$$

which is an upper bound of the bound in part (v) of the theorem, to provide an upper approximation of $\|\hat{\beta}_k\|_1$. Combining the above estimate with the guarantee in part (i) of Theorem 2.1 in (2.11), we obtain the following ordered pairs:

$$\left(\frac{1}{2n}\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2 \cdot \gamma^k + L_n^*, \ell_k\right), \qquad k \geq 1, \tag{2.12}$$

which describe the *entire* profile of the training error bounds and the ℓ_1-shrinkage bounds as a function of k as suggested by Theorem 2.1. These profiles, as described above in (2.12), are illustrated in Figure 3.

It is interesting to consider the profiles of Figure 3 alongside the *explicit* regularization framework of the LASSO (1.1) which also traces out a profile of the form (2.11), namely,

$$\left(\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\hat{\beta}_\delta^*\|_2^2, \|\hat{\beta}_\delta^*\|_1\right), \qquad \delta \geq 0, \tag{2.13}$$

as a function of δ, where, $\hat{\beta}_\delta^*$ is a solution to the LASSO problem (1.1). For a value of $\delta := \ell_k$ the optimal objective value of the LASSO problem will serve as a lower bound of the corresponding LS-BOOST(ε) loss function value at iteration

k. Thus, the training error of $\hat{\beta}^k$ delivered by the LS-BOOST(ε) algorithm will be sandwiched between the following lower and upper bounds:

$$L_{i,k} := \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\hat{\beta}^*_{\ell_k}\|_2^2 \leq \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\hat{\beta}^k\|_2^2 \leq \frac{1}{2n}\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2^2 \cdot \gamma^k + L_n^* =: U_{i,k}$$

for every k. Note that the difference between the upper and lower bounds above, given by $U_{i,k} - L_{i,k}$, converges to zero as $k \to \infty$. Figure A.1, Section A.1.1 in [14] shows the training error versus shrinkage profiles for LS-BOOST(ε) and LASSO solutions, for different datasets.

For the bounds in parts (i) and (iii) of Theorem 2.1, the asymptotic limits (as $k \to \infty$) are the unregularized least squares training error and predictions—which are quantities that are uniquely defined even in the underdetermined case.

The bound in part (ii) of Theorem 2.1 is a statement concerning the regression coefficients. In this case, the notion of convergence needs to be appropriately modified from parts (i) and (iii), since the *natural* limiting object $\hat{\beta}_{\mathrm{LS}}$ is not necessarily unique. In this case, perhaps not surprisingly, the regression coefficients $\hat{\beta}^k$ need not converge. The result in part (ii) of the theorem states that $\hat{\beta}^k$ converges at a linear rate to the *set* of least squares solutions. In other words, at every LS-BOOST(ε) boosting iteration, there exists a least squares solution $\hat{\beta}^k_{\mathrm{LS}}$ for which the presented bound holds. Here, $\hat{\beta}^k_{\mathrm{LS}}$ is in fact the closest least squares solution to $\hat{\beta}^k$ in the ℓ_2 norm—and the particular candidate least squares solution $\hat{\beta}^k_{\mathrm{LS}}$ may be different for each iteration.

Interpreting the parameters and algorithm dynamics. There are several determinants of the quality of the bounds in the different parts of Theorem 2.1 which can be grouped into:

- algorithmic parameters: this includes the learning rate ε and the number of iterations k, and
- data dependent quantities: $\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2$, $\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})$ and p.

The coefficient of linear convergence is given by the quantity $\gamma := 1 - \frac{\varepsilon(2-\varepsilon)}{4\kappa(\mathbf{X}^T\mathbf{X})}$, where $\kappa(\mathbf{X}^T\mathbf{X}) := \frac{p}{\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})}$. Note that γ is monotone decreasing in ε for $\varepsilon \in (0, 1]$, and is minimized at $\varepsilon = 1$. This simple observation confirms the general intuition about LS-BOOST(ε): $\varepsilon = 1$ corresponds to the most aggressive model fitting behavior in the LS-BOOST(ε) family, with smaller values of ε corresponding to a slower model fitting process. The ratio $\kappa(\mathbf{X}^T\mathbf{X})$ is a close cousin of the condition number associated with the data matrix $\mathbf{X}$—and smaller values of $\kappa(\mathbf{X}^T\mathbf{X})$ imply a faster rate of convergence.

In the overdetermined case with $n \geq p$ and $\mathrm{rank}(\mathbf{X}) = p$, the condition number $\bar{\kappa}(\mathbf{X}^T\mathbf{X}) := \frac{\lambda_{\max}(\mathbf{X}^T\mathbf{X})}{\lambda_{\min}(\mathbf{X}^T\mathbf{X})}$ plays a key role in determining the stability of the least-squares solution $\hat{\beta}_{\mathrm{LS}}$ and in measuring the degree of multicollinearity present.

Note that $\bar{\kappa}(\mathbf{X}^T\mathbf{X}) \in [1, \infty)$, and that the problem is better conditioned for smaller values of this ratio. Furthermore, since $\operatorname{rank}(\mathbf{X}) = p$ it holds that $\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X}) = \lambda_{\min}(\mathbf{X}^T\mathbf{X})$, and thus $\bar{\kappa}(\mathbf{X}^T\mathbf{X}) \le \kappa(\mathbf{X}^T\mathbf{X})$ by (2.6). Thus, the condition number $\kappa(\mathbf{X}^T\mathbf{X})$ always upper bounds the classical condition number $\bar{\kappa}(\mathbf{X}^T\mathbf{X})$, and if $\lambda_{\max}(\mathbf{X}^T\mathbf{X})$ is close to p, then $\bar{\kappa}(\mathbf{X}^T\mathbf{X}) \approx \kappa(\mathbf{X}^T\mathbf{X})$ and the two measures essentially coincide. Finally, since in this setup $\hat{\beta}_{\text{LS}}$ is unique, part (ii) of Theorem 2.1 implies that the sequence $\{\hat{\beta}^k\}$ converges linearly to the unique least squares solution $\hat{\beta}_{\text{LS}}$.

In the underdetermined case with $p > n$, $\lambda_{\min}(\mathbf{X}^T\mathbf{X}) = 0$, and thus $\bar{\kappa}(\mathbf{X}^T\mathbf{X}) = \infty$. On the other hand, $\kappa(\mathbf{X}^T\mathbf{X}) < \infty$ since $\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})$ is the smallest *nonzero* (hence positive) eigenvalue of $\mathbf{X}^T\mathbf{X}$. Therefore, the condition number $\kappa(\mathbf{X}^T\mathbf{X})$ is similar to the classical condition number $\bar{\kappa}(\cdot)$ restricted to the subspace $\mathcal{S}$ spanned by the columns of $\mathbf{X}$ [whose dimension is $\operatorname{rank}(\mathbf{X})$]. Interestingly, the linear rate of convergence enjoyed by LS-BOOST(ε) is in a sense adaptive—the algorithm automatically adjusts itself to the convergence rate dictated by the parameter γ "as if" it knows that the null space of $\mathbf{X}$ is not relevant.

As the dataset is varied, the value of γ can change substantially from one dataset to another, thereby leading to differences in the convergence behavior bounds in parts (i)–(v) of Theorem 2.1. To settle all of these ideas, we can derive some simple bounds on γ using tools from random matrix theory. Towards this end, let us suppose that the entries of $\mathbf{X}$ are drawn from a standard Gaussian ensemble, which are subsequently standardized such that every column of $\mathbf{X}$ has unit ℓ_2 norm. Then it follows from random matrix theory [49] that $\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X}) \gtrapprox \frac{1}{n}(\sqrt{p} - \sqrt{n})^2$ with high probability. (See Section A.2.4 in [14] for a more detailed discussion of this fact.) To gain better insights into the behavior of γ and how it depends on the values of pairwise correlations of the features, we performed some computational experiments, the results of which are shown in Figure 4. Figure 4 shows the behavior of γ as a function of p for a fixed $n = 50$ and $\varepsilon = 1$, for different datasets $\mathbf{X}$ simulated as follows. We first generated a multivariate data matrix from a Gaussian distribution with mean zero and covariance $\Sigma_{p\times p} = (\sigma_{ij})$, where, $\sigma_{ij} = \rho$ for all $i \neq j$; and then all of the columns of the data matrix were standardized to have unit ℓ_2 norm. The resulting matrix was taken as $\mathbf{X}$. We considered different cases by varying the magnitude of pairwise correlations of the features ρ—when ρ is small, the rate of convergence is typically faster (smaller γ) and the rate becomes slower (higher γ) for higher values of ρ. Figure 4 shows that the coefficient of linear convergence γ is quite close to 1.0—which suggests a slowly converging algorithm and confirms our intuition about the algorithmic behavior of LS-BOOST(ε). Indeed, LS-BOOST(ε), like any other boosting algorithm, should indeed converge slowly to the unregularized least squares solution. The slowly converging nature of the LS-BOOST(ε) algorithm provides, for the first time, a precise theoretical justification of the empirical observation made in [31] that stagewise regression is widely considered ineffective as a tool to obtain the unregularized least squares

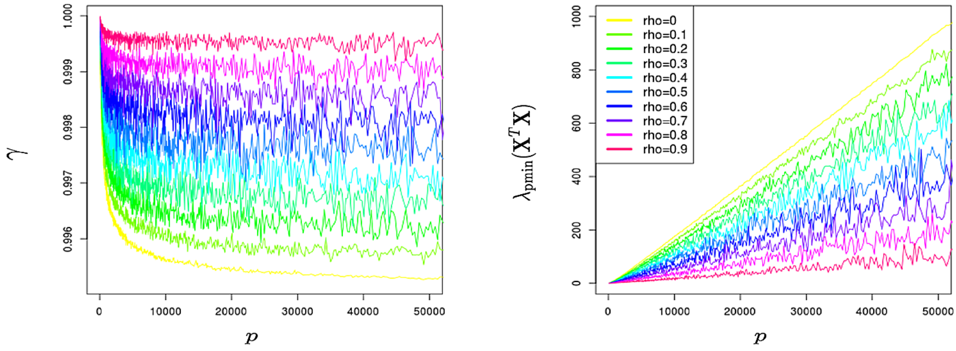

FIG. 4. *Figure showing the behavior of γ [left panel] and $\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})$ [right panel] for different values of ρ (denoted by the moniker "rho" in the legend) and p, with $\varepsilon = 1$. There are ten profiles in each panel corresponding to different values of ρ for $\rho = 0, 0.1, \dots, 0.9$. Each profile documents the change in γ as a function of p, the smallest value of p appearing in the display is $p = 73$. Here, the data matrix $\mathbf{X}$ is comprised of $n = 50$ samples from a p-dimensional multivariate Gaussian distribution with mean zero, and all pairwise correlations equal to ρ, and the features are then standardized to have unit ℓ_2 norm. The left panel shows that γ exhibits a phase of rapid decay (as a function of p) after which it stabilizes into the regime of* fastest *convergence. Interestingly, the behavior shows a monotone trend in ρ: the rate of progress of* LS-BOOST(ε) *becomes slower for larger values of ρ and faster for smaller values of ρ.*

fit, as compared to other stepwise model fitting procedures like Forward Stepwise regression (discussed in Section 1).

The above discussion sheds some interesting insight into the behavior of the LS-BOOST(ε) algorithm. For larger values of ρ, the observed covariates tend to be even more highly correlated (since $p \gg n$). Whenever a pair of features are highly correlated, the LS-BOOST(ε) algorithm finds it *difficult* to prefer one over the other, and thus takes turns in updating both coefficients, thereby distributing the effects of a covariate to all of its correlated cousins. Since a group of correlated covariates are all competing to be updated by the LS-BOOST(ε) algorithm, the progress made by the algorithm in decreasing the loss function is naturally slowed down. In contrast, when ρ is small, the LS-BOOST(ε) algorithm brings in a covariate and in a sense completes the process by doing the exact line-search on that feature. This heuristic explanation attempts to explain the slower rate of convergence of the LS-BOOST(ε) algorithm for large values of ρ—a phenomenon that we observe in practice and which is also substantiated by the computational guarantees in Theorem 2.1. We refer the reader to Figures 1 and 5 which further illustrate the above justification. Statement (v) of Theorem 2.1 provides upper bounds on the ℓ_1 shrinkage of the coefficients. Figure 3 illustrates the evolution of the data-fidelity versus ℓ_1-shrinkage as obtained from the computational bounds in Theorem 2.1. Some additional discussion and properties of LS-BOOST(ε) are presented in the Supplementary Material Section A.2.3 [14].

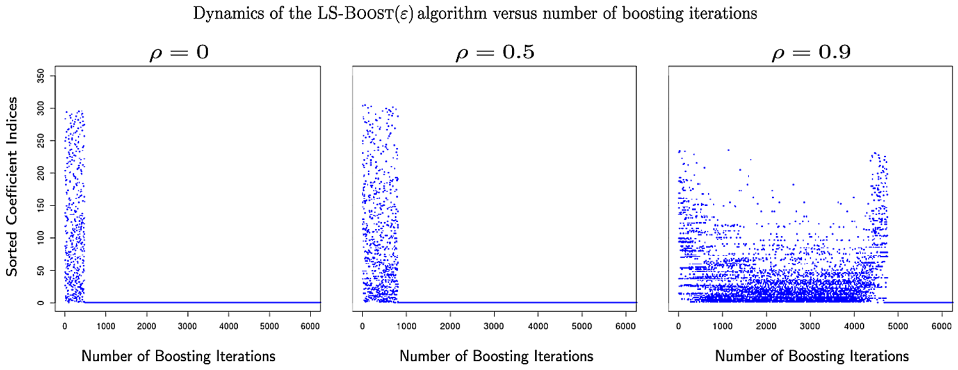

FIG. 5. *Showing the* LS-BOOST(ε) *algorithm run on the same synthetic dataset as was used in Figure* 1 *with* $\varepsilon = 1$, *for three different values of the pairwise correlation* ρ. *A point is "on" if the corresponding regression coefficient is updated at iteration* k. *Here, the vertical axes have been reoriented so that the coefficients that are updated the maximum number of times appear lower on the axes. For larger values of* ρ, *we see that the* LS-BOOST(ε) *algorithm aggressively updates the coefficients for a large number of iterations, whereas the dynamics of the algorithm for smaller values of* ρ *are less pronounced. For larger values of* ρ *the* LS-BOOST(ε) *algorithm takes longer to reach the least squares fit and this is reflected in the above figure from the update patterns in the regression coefficients. The dynamics of the algorithm evident in this figure nicely complements the insights gained from Figure* 1.

We briefly discuss computational guarantees for Forward Stepwise regression (also known as Forward Selection) [12, 31, 50]. The Forward Stepwise algorithm, at the kth iteration, selects a covariate $\mathbf{X}_{j_k}$ maximally correlated with the current residual:[7] $j_k \in \arg\max_j |(\hat{r}^k)^T \mathbf{X}_j|$. If the set of active (i.e., nonzero) regression coefficient indices at iteration k is denoted by $\mathcal{I}_k$, then the algorithm appends j_k to this set, $\mathcal{I}_{k+1} \leftarrow \mathcal{I}_k \cup \{j_k\}$, and updates the regression coefficients by performing a *joint* least squares regression of $\mathbf{y}$ restricted to the covariates in $\mathcal{I}_{k+1}$:

$$\hat{\beta}^{k+1} \in \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \quad \text{s.t.} \quad \beta_i = 0 \qquad \forall i \notin \mathcal{I}_{k+1}.$$

Notice that the least-squares loss function value for Forward Stepwise is bounded by the corresponding value if one was instead using LS-BOOST, that is, $L_n(\hat{\beta}^{k+1}) \leq L_n(\hat{\beta}^k + \tilde{u}_{j_k} e_{j_k})$. Thus, it is straightforward to show that parts (i)–(iv) [and also part (vi)] of Theorem 2.1 (with $\varepsilon = 1$) hold in their current form for Forward Stepwise as well. Interestingly, part (v) does not hold for Forward Stepwise nor should we expect Forward Stepwise—which is the most aggressive model fitting procedure described herein—to have any such guarantees about the shrinkage of coefficients. In the signal processing literature, Forward Stepwise is popularly known as Orthogonal Matching Pursuit [38].

[7] Recall that we assume all the covariates to have unit ℓ_2 norm.

3. Boosting algorithms as subgradient descent.

Roadmap. In this section, we present a new unifying framework for interpreting the three boosting algorithms that were discussed in Section 1, namely FS_ε, its nonuniform learning rate extension $\text{FS}_{\varepsilon_k}$, and LS-BOOST($\varepsilon$). We show herein that all three algorithmic families can be interpreted as instances of the subgradient descent method of convex optimization, applied to the problem of minimizing the largest correlation between residuals and predictors. Interestingly, this unifying lens will also result in a natural generalization of FS_ε with very strong ties to the LASSO problem and its solution, as we will present in Sections 4 and 5. The framework presented in this section leads to convergence guarantees for FS_ε and $\text{FS}_{\varepsilon_k}$. In Theorem 3.1 herein, we present a theoretical description of the evolution of the FS_ε algorithm, in terms of its data-fidelity and shrinkage guarantees as a function of the number of boosting iterations. These results are a consequence of the computational guarantees for FS_ε that inform us about the rate at which the FS_ε training error, regression coefficients, and predictions make their way to their least squares counterparts. In order to develop these results, we motivate and briefly review the subgradient descent method of convex optimization.

3.1. *Boosting algorithms* FS_ε, LS-BOOST(ε) *and* $\text{FS}_{\varepsilon_k}$.

We present a formal description of the FS_ε algorithm introduced in Section 1.

Algorithm: Incremental Forward Stagewise Regression—FS_ε

Fix the learning rate $\varepsilon > 0$, the number of iterations M, and initialize $\hat{\beta}^0 = 0$ and $\hat{r}^0 = \mathbf{y}$.

```
1. For $0 \le k \le M$ do the following:
2. Compute $j_k \in \arg\max_{j \in \{1,\dots,p\}} |(\hat{r}^k)^T \mathbf{X}_j|$.
3. Update the regression coefficients and residuals as
```

$$\begin{aligned} &\hat{\beta}_{j_k}^{k+1} \leftarrow \hat{\beta}_{j_k}^{k} + \varepsilon \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}), \qquad \hat{\beta}_{j}^{k+1} \leftarrow \hat{\beta}_{j}^{k}, j \neq j_k, \quad \text{and} \\ &\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) \mathbf{X}_{j_k}. \end{aligned} \tag{3.1}$$

The FS_ε algorithm in the kth iteration greedily chooses a covariate $\mathbf{X}_{j_k}$ that is the most correlated (in absolute value) with the current residual and updates the j_kth regression coefficient, along with the residuals, with a shrinkage factor ε. Firstly, since all of the covariates are standardized to have unit ℓ_2 norm, for the same given residual value $\hat{r}^k$ it is simple to derive that LS-BOOST(ε) and FS_ε lead to the same choice of j_k. Qualitatively speaking, as in the case of LS-BOOST(ε), a smaller value of ε corresponds to a slower learning procedure. However, in contrast to LS-BOOST(ε), where ε lies naturally in $(0, 1]$, the choice of ε in FS_ε is more sensitive to the scale of the problem. Indeed LS-BOOST(ε) and FS_ε in spite of their similarities contain subtle differences, for example, in their residual updates:

$$\begin{aligned} \text{LS-BOOST}(\varepsilon): \quad & \|\hat{r}^{k+1} - \hat{r}^k\|_2 = \varepsilon|(\hat{r}^k)^T\mathbf{X}_{j_k}| = \varepsilon \cdot n \cdot \|\nabla L_n(\hat{\beta}^k)\|_\infty \\ \text{FS}_\varepsilon: \quad & \|\hat{r}^{k+1} - \hat{r}^k\|_2 = \varepsilon|s_k| \qquad \text{where } s_k = \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k}), \end{aligned} \tag{3.2}$$

where $\nabla L_n(\cdot)$ is the gradient of $L_n(\beta)$. Note that for both of the algorithms, the quantity $\|\hat{r}^{k+1} - \hat{r}^k\|_2$ involves the shrinkage factor ε. Their difference thus lies in the multiplicative factor, which is $n \cdot \|\nabla L_n(\hat{\beta}^k)\|_\infty$ for LS-BOOST(ε) and is $|\operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})|$ for FS_ε. The norm of the successive residual differences for LS-BOOST(ε) is proportional to the ℓ_∞ norm of the gradient of the least squares loss function [see herein equations (2.2) and (2.4)]. For FS_ε, the norm of the successive residual differences depends on the absolute value of the sign of the j_kth coordinate of the gradient. Note that $s_k \in \{-1, 0, 1\}$ depending upon whether $(\hat{r}^k)^T\mathbf{X}_{j_k}$ is negative, zero or positive; and $s_k = 0$ only when $(\hat{r}^k)^T\mathbf{X}_{j_k} = 0$, that is, only when $\|\nabla L_n(\hat{\beta}^k)\|_\infty = 0$ and hence $\hat{\beta}^k$ is a least squares solution. Thus, for FS_ε the ℓ_2 norm of the difference in residuals is almost always ε during the course of the algorithm. For the LS-BOOST(ε) algorithm, progress is considerably more sensitive to the norm of the gradient—as the algorithm makes its way to the unregularized least squares fit, one should expect the norm of the gradient to also shrink to zero, as we have established formally in Section 2. Qualitatively speaking, this means that the updates of LS-BOOST(ε) are more well-behaved when compared to the updates of FS_ε, which are more erratically behaved. Of course, the additional shrinkage factor ε further dampens the progress for both algorithms.

While Section 2 shows that the predicted values $\mathbf{X}\hat{\beta}^k$ obtained from LS-BOOST(ε) converge (at a globally linear rate) to the least squares fit as $k \to \infty$ (for any value of $\varepsilon \in (0, 1]$); on the other hand, for FS_ε with $\varepsilon > 0$, the iterates $\mathbf{X}\hat{\beta}^k$ need not necessarily converge to the least squares fit as $k \to \infty$. Indeed, the FS_ε algorithm, by its operational definition, has a uniform learning rate ε which remains fixed for all iterations; this makes it impossible to always guarantee convergence to a least squares solution with accuracy less than $O(\varepsilon)$. We show in this section that the predictions from the FS_ε algorithm converges to an approximate least squares solution, albeit at a global sublinear rate.[8]

Since the main difference between FS_ε and LS-BOOST(ε) lies in the choice of the step-size used to update the coefficients, let us therefore consider a nonconstant step-size/nonuniform learning rate version of FS_ε, which we call $\text{FS}_{\varepsilon_k}$. $\text{FS}_{\varepsilon_k}$ replaces update (3.1) of FS_ε by

$$\textit{residual update}: \quad \hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon_k \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})\mathbf{X}_{j_k}$$

[8] For the purposes of this paper, linear convergence of a sequence $\{a_i\}$ will mean that $a_i \to \bar{a}$ and there exists a scalar $\gamma < 1$ for which $(a_i - \bar{a})/(a_{i-1} - \bar{a}) \le \gamma$ for all i. Sublinear convergence will mean that there is no such $\gamma < 1$ that satisfies the above property. For much more general versions of linear and sublinear convergence, see [2] for example.

coefficient update: $\hat{\beta}^{k+1}_{j_k} \leftarrow \hat{\beta}^{k}_{j_k} + \varepsilon_k \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k})$ and $\hat{\beta}^{k+1}_{j} \leftarrow \hat{\beta}^{k}_{j}, j \neq j_k$,

where $\{\varepsilon_k\}$ is a sequence of learning-rates (or step-sizes) which depend upon the iteration index k. LS-BOOST(ε) can thus be thought of as a version of $\text{FS}_{\varepsilon_k}$, where the step-size ε_k is given by $\varepsilon_k := \varepsilon \tilde{u}_{j_k} \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k})$.

In Section 3.3, we provide a unified treatment of LS-BOOST(ε), FS_ε and $\text{FS}_{\varepsilon_k}$, wherein we show that all these methods can be viewed as special instances of subgradient optimization.

3.2. *Brief review of subgradient descent.* We briefly motivate and review the subgradient descent method for nondifferentiable convex optimization problems. Consider the following optimization problem:

$$\begin{aligned} f^* := \min_x \; & f(x) \\ \text{s.t.} \quad & x \in P, \end{aligned} \tag{3.3}$$

where $P \subseteq \mathbb{R}^n$ is a closed convex set and $f(\cdot) : P \to \mathbb{R}$ is a convex function. If $f(\cdot)$ is differentiable, then $f(\cdot)$ will satisfy the following gradient inequality:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \qquad \text{for any } x, y \in P,$$

which states that $f(\cdot)$ lies above its first-order (linear) approximation at x. One of the most intuitive optimization schemes for solving (3.3) is the method of gradient descent. This method is initiated at a given point $x^0 \in P$. If x^k is the current iterate, then the next iterate is given by the update formula: $x^{k+1} \leftarrow \Pi_P(x^k - \alpha_k \nabla f(x^k))$. In this method, the potential new point is $x^k - \alpha_k \nabla f(x^k)$, where $\alpha_k > 0$ is called the step-size at iteration k, and the step is taken in the direction of the negative of the gradient. If this potential new point lies outside of the feasible region P, it is then projected back onto P. Here, recall that $\Pi_P(\cdot)$ is the Euclidean projection operator, namely $\Pi_P(x) := \arg\min_{y \in P} \|x - y\|_2$.

Now suppose that $f(\cdot)$ is not differentiable. By virtue of the fact that $f(\cdot)$ is convex, $f(\cdot)$ will have a *subgradient* at each point x. Recall that g is a subgradient of $f(\cdot)$ at x if the following subgradient inequality holds:

$$f(y) \geq f(x) + g^T (y - x) \qquad \text{for all } y \in P, \tag{3.4}$$

which generalizes the gradient inequality above and states that $f(\cdot)$ lies above the linear function on the right-hand side of (3.4). Because there may exist more than one subgradient of $f(\cdot)$ at x, let $\partial f(x)$ denote the set of subgradients of $f(\cdot)$ at x. Then "$g \in \partial f(x)$" denotes that g is a subgradient of $f(\cdot)$ at the point x, and so g satisfies (3.4) for all y. The subgradient descent method (see, e.g., [45]) is a simple generalization of the method of gradient descent to the case when $f(\cdot)$ is not differentiable. One simply replaces the gradient by the subgradient, yielding the following update scheme:

$$\begin{aligned} \text{Compute a subgradient of } f(\cdot) \text{ at } x^k: \quad & g^k \in \partial f(x^k), \\ \text{Peform update at } x^k: \quad & x^{k+1} \leftarrow \Pi_P(x^k - \alpha_k g^k). \end{aligned} \tag{3.5}$$

The following proposition summarizes a well-known computational guarantee associated with the subgradient descent method.

PROPOSITION 3.1 (Convergence bound for subgradient descent [37, 39]). *Consider the subgradient descent method* (3.5), *using a constant step-size* $\alpha_i = \alpha$ *for all* i. *Let* x^* *be an optimal solution of* (3.3) *and suppose that the subgradients are uniformly bounded, namely* $\|g^i\|_2 \leq G$ *for all* $i \geq 0$. *Then for each* $k \geq 0$, *the following inequality holds*:

$$\min_{i \in \{0,\dots,k\}} f(x^i) \leq f^* + \frac{\|x^0 - x^*\|_2^2}{2(k+1)\alpha} + \frac{\alpha G^2}{2}. \tag{3.6}$$

The left-hand side of (3.6) is simply the best objective function value obtained among the first k iterations. The right-hand side of (3.6) bounds the best objective function value from above, namely the optimal value f^* plus a nonnegative quantity that is a function of the number of iterations k, the constant step-size α, the bound G on the norms of subgradients and the distance from the initial point to an optimal solution x^* of (3.3). Note that for a fixed step-size $\alpha > 0$, the right-hand side of (3.6) goes to $\frac{\alpha G^2}{2}$ as $k \to \infty$. In the interest of completeness, we include a proof of Proposition 3.1 in the Supplementary Material Section A.2.5 [14].

3.3. *A subgradient descent framework for boosting.* We now show that the boosting algorithms discussed in Section 1, namely FS_ε and its relatives $\text{FS}_{\varepsilon_k}$ and LS-BOOST(ε), can all be interpreted as instantiations of the subgradient descent method to minimize the largest absolute correlation between the residuals and predictors.

Let $P_{\text{res}} := \{r \in \mathbb{R}^n : r = \mathbf{y} - \mathbf{X}\beta \text{ for some } \beta \in \mathbb{R}^p\}$ denote the affine space of residuals and consider the following convex optimization problem:

$$\begin{aligned} \text{Correlation Minimization (CM)}: \quad f^* := \min_r \; & f(r) := \|\mathbf{X}^T r\|_\infty \\ \text{s.t.} \quad & r \in P_{\text{res}}, \end{aligned} \tag{3.7}$$

which we dub the "Correlation Minimization" problem, or CM for short. Note an important subtlety in the CM problem, namely that the optimization variable in CM is the *residual* r and *not* the regression coefficient vector β.

Since the columns of $\mathbf{X}$ have unit ℓ_2 norm by assumption, $f(r)$ is the largest absolute correlation between the residual vector r and the predictors. Therefore, (3.7) is the convex optimization problem of minimizing the largest correlation between the residuals and the predictors, over all possible values of the residuals. From (2.3) with $r = \mathbf{y} - \mathbf{X}\beta$, we observe that $\mathbf{X}^T r = 0$ if and only if β is a least squares solution, whereby $f(r) = \|\mathbf{X}^T r\|_\infty = 0$ for the least squares residual vector $r = \hat{r}_{\text{LS}} = \mathbf{y} - \mathbf{X}\hat{\beta}_{\text{LS}}$. Since the objective function in (3.7) is nonnegative, we conclude that $f^* = 0$ and the least squares residual vector $\hat{r}_{\text{LS}}$ is also the unique optimal solution of the CM problem (3.7). Thus, CM can be viewed as an optimization problem which also produces the least squares solution.

The following proposition states that the three boosting algorithms FS_ε, $\text{FS}_{\varepsilon_k}$ and LS-BOOST(ε) can all be viewed as instantiations of the subgradient descent method to solve the CM problem (3.7).

PROPOSITION 3.2. *Consider the subgradient descent method* (3.5) *with step-size sequence* $\{\alpha_k\}$ *to solve the correlation minimization* (*CM*) *problem* (3.7), *initialized at* $\hat{r}^0 = \mathbf{y}$. *Then*:

(i) *the* FS_ε *algorithm is an instance of subgradient descent*, *with a constant step-size* $\alpha_k := \varepsilon$ *at each iteration*,

(ii) *the* $\text{FS}_{\varepsilon_k}$ *algorithm is an instance of subgradient descent*, *with nonuniform step-sizes* $\alpha_k := \varepsilon_k$ *at iteration* k, *and*

(iii) *the* LS-BOOST(ε) *algorithm is an instance of subgradient descent*, *with nonuniform step-sizes* $\alpha_k := \varepsilon|\tilde{u}_{j_k}|$ *at iteration* k, *where* $\tilde{u}_{j_k} := \arg\min_u \|\hat{r}^k - \mathbf{X}_{j_k}u\|_2^2$.

PROOF. We first prove (i). Recall the update of the residuals in FS_ε:

$$\hat{r}^{k+1} = \hat{r}^k - \varepsilon \cdot \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})\mathbf{X}_{j_k}.$$

We first show that $g^k := \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})\mathbf{X}_{j_k}$ is a subgradient of the objective function $f(r) = \|\mathbf{X}^T r\|_\infty$ of the correlation minimization problem CM (3.7) at $r = \hat{r}^k$. At iteration k, FS_ε chooses the coefficient to update by selecting $j_k \in \arg\max_{j\in\{1,\dots,p\}} |(\hat{r}^k)^T\mathbf{X}_j|$, whereby $\operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})((\hat{r}^k)^T\mathbf{X}_{j_k}) = \|\mathbf{X}^T(\hat{r}^k)\|_\infty$, and, therefore, for any r it holds that

$$\begin{aligned} f(r) = \|\mathbf{X}^T r\|_\infty &\geq \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})((\mathbf{X}_{j_k})^T r) \\ &= \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})((\mathbf{X}_{j_k})^T(\hat{r}^k + r - \hat{r}^k)) \\ &= \|\mathbf{X}^T(\hat{r}^k)\|_\infty + \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})((\mathbf{X}_{j_k})^T(r - \hat{r}^k)) \\ &= f(\hat{r}^k) + \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})((\mathbf{X}_{j_k})^T(r - \hat{r}^k)). \end{aligned}$$

Therefore, using the definition of a subgradient in (3.4), it follows that $g^k := \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})\mathbf{X}_{j_k}$ is a subgradient of $f(r) = \|\mathbf{X}^T r\|_\infty$ at $r = \hat{r}^k$. Therefore, the update $\hat{r}^{k+1} = \hat{r}^k - \varepsilon \cdot \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})\mathbf{X}_{j_k}$ is of the form $\hat{r}^{k+1} = \hat{r}^k - \varepsilon g^k$ where $g^k \in \partial f(\hat{r}^k)$. Last of all notice that the update can also be written as $\hat{r}^k - \varepsilon g^k = \hat{r}^{k+1} = \mathbf{y} - \mathbf{X}\hat{\beta}^{k+1} \in P_{\text{res}}$, hence $\Pi_{P_{\text{res}}}(\hat{r}^k - \varepsilon g^k) = \hat{r}^k - \varepsilon g^k$, that is, the projection step is superfluous here and, therefore, $\hat{r}^{k+1} = \Pi_{P_{\text{res}}}(\hat{r}^k - \varepsilon g^k)$, which is precisely the update for the subgradient descent method with step-size $\alpha_k := \varepsilon$.

The proof of (ii) is the same as (i) with a step-size choice of $\alpha_k = \varepsilon_k$ at iteration k. Furthermore, as discussed earlier, LS-BOOST(ε) may be thought of as a specific instance of $\text{FS}_{\varepsilon_k}$, whereby the proof of (iii) follows as a special case of (ii). □

Proposition 3.2 presents a new interpretation of the boosting algorithms FS_ε and its cousins as subgradient descent. This is interesting especially since FS_ε and LS-BOOST(ε) have been traditionally interpreted as greedy coordinate descent or steepest descent type procedures [22, 31]. This has the following consequences of note:

- We take recourse to existing tools and results about subgradient descent optimization to inform us about the computational guarantees of these methods. When translated to the setting of linear regression, these results will shed light on the data fidelity versus shrinkage characteristics of FS_ε and its cousins— all using quantities that can be easily obtained prior to running the boosting algorithm. We will show the details of this in Theorem 3.1 below.
- The subgradient optimization viewpoint provides a unifying algorithmic theme which we will also apply to a regularized version of problem CM (3.7), and that we will show is very strongly connected to the LASSO. This will be developed in Section 4. Indeed, the regularized version of the CM problem that we will develop in Section 4 will lead to a new family of boosting algorithms which are a seemingly minor variant of the basic FS_ε algorithm but deliver [$O(\varepsilon)$-approximate] solutions to the LASSO.

3.4. *Deriving and interpreting computational guarantees for* FS_ε. The following theorem presents the convergence properties of FS_ε, which are a consequence of the interpretation of FS_ε as an instance of the subgradient descent method.

THEOREM 3.1 (Convergence Properties of FS_ε). *Consider the* FS_ε *algorithm with learning rate* ε. *Let* $k \geq 0$ *be the total number of iterations. Then there exists an index* $i \in \{0, \dots, k\}$ *for which the following bounds hold*:

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

(ii) (*regression coefficients*): *there exists a least squares solution* $\hat{\beta}^i_{LS}$ *such that*

$$\|\hat{\beta}^i - \hat{\beta}^i_{LS}\|_2 \leq \frac{\sqrt{p}}{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}\left[\frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2}{\varepsilon(k+1)} + \varepsilon\right];$$

(iii) (*predictions*): *for every least-squares solution* $\hat{\beta}_{\text{LS}}$ *it holds that*

$$\|\mathbf{X}\hat{\beta}^i - \mathbf{X}\hat{\beta}_{\text{LS}}\|_2 \leq \frac{\sqrt{p}}{\sqrt{\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}}\left[\frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2}{\varepsilon(k+1)} + \varepsilon\right];$$

(iv) (*correlation values*) $\|\mathbf{X}^T\hat{r}^i\|_\infty \leq \frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2}{2\varepsilon(k+1)} + \frac{\varepsilon}{2}$;

(v) (ℓ_1-*shrinkage of coefficients*): $\|\hat{\beta}^i\|_1 \leq k\varepsilon$;

(vi) (*sparsity of coefficients*): $\|\hat{\beta}^i\|_0 \leq k$.

The proof of Theorem 3.1 is presented in Section A.2.6, in [14].

Interpreting the computational guarantees. Theorem 3.1 accomplishes for FS_ε what Theorem 2.1 did for LS-BOOST(ε)—parts (i)–(iv) of the theorem describe the rate in which the training error, regression coefficients and related quantities make their way toward their [$O(\varepsilon)$-approximate] unregularized least squares counterparts. Part (v) of the theorem also describes the rate at which the shrinkage of the regression coefficients evolve as a function of the number of boosting iterations. The rate of convergence of FS_ε is sublinear, unlike the linear rate of convergence for LS-BOOST(ε). Note that this type of sublinear convergence implies that the rate of decrease of the training error (for instance) is dramatically faster in the very early iterations as compared to later iterations. Taken together, Theorems 3.1 and 2.1 highlight an important difference between the behavior of algorithms LS-BOOST(ε) and FS_ε:

- the limiting solution of the LS-BOOST(ε) algorithm (as $k \to \infty$) corresponds to the unregularized least squares solution, but
- the limiting solution of the FS_ε algorithm (as $k \to \infty$) corresponds to an $O(\varepsilon)$ approximate least squares solution.

As demonstrated in Theorems 2.1 and 3.1, both LS-BOOST(ε) and FS_ε have nice convergence properties with respect to the unconstrained least squares problem (2.1). However, unlike the convergence results for LS-BOOST(ε) in Theorem 2.1, FS_ε exhibits a *sublinear* rate of convergence towards a *suboptimal* least squares solution. For example, part (i) of Theorem 3.1 implies in the limit as $k \to \infty$ that FS_ε identifies a model with training error at most

$$L_n^* + \frac{p\varepsilon^2}{2n(\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X}))}. \tag{3.8}$$

In addition, part (ii) of Theorem 3.1 implies that as $k \to \infty$, FS_ε identifies a model whose distance to the set of least squares solutions $\{\hat{\beta}_{\mathrm{LS}} : \mathbf{X}^T\mathbf{X}\hat{\beta}_{\mathrm{LS}} = \mathbf{X}^T\mathbf{y}\}$ is at most: $\frac{\varepsilon\sqrt{p}}{\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})}$.

Note that the computational guarantees in Theorem 3.1 involve the quantities $\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})$ and $\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2$, assuming n and p are fixed. To settle ideas, let us consider the synthetic datasets used in Figures 4 and 1, where the covariates were generated from a multivariate Gaussian distribution with pairwise correlation ρ. Figure 4 suggests that $\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X})$ decreases with increasing ρ values. Thus, controlling for other factors appearing in the computational bounds[9], it follows from the statements of Theorem 3.1 that the training error decreases much more rapidly for smaller ρ values, as a function of k. This is nicely validated by the computational results in Figure 1 (the three top panel figures), which show that the training errors decay at a faster rate for smaller values of ρ.

[9] To control for other factors, for example, we may assume that $p > n$ and for different values of ρ we have $\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2 = \|\mathbf{y}\|_2 = 1$ with ε fixed across the different examples.

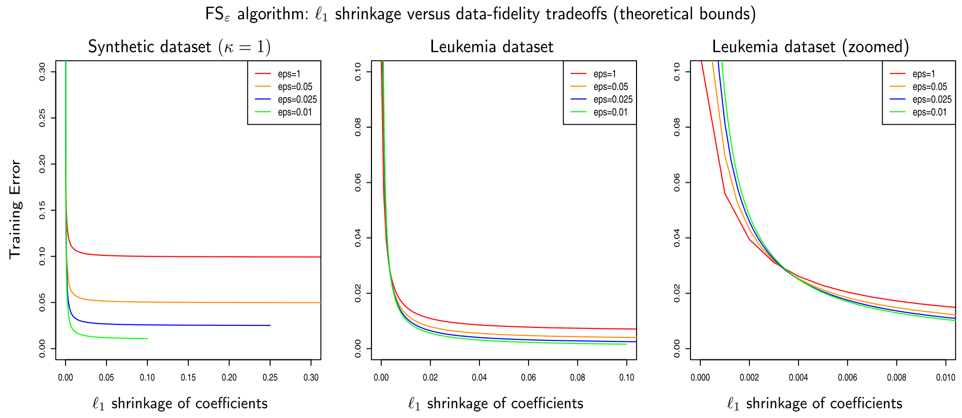

FIG. 6. *Figure showing profiles of ℓ_1 shrinkage bounds of the regression coefficients versus training error bounds for the* FS$_\varepsilon$ *algorithm, for different values of the learning rate ε. The profiles have been obtained from the bounds in parts* (i) *and* (v) *of Theorem* 3.1. *The left panel corresponds to a hypothetical dataset using $\kappa = \frac{p}{\lambda_{\text{pmin}}} = 1$, and the middle and right panels use the parameters of the Leukemia dataset.*

Let us examine more carefully the properties of the sequence of models explored by FS$_\varepsilon$ and the corresponding tradeoffs between data fidelity and model complexity. Let TBOUND and SBOUND denote the training error bound and shrinkage bound in parts (i) and (v) of Theorem 3.1, respectively. Then simple manipulation of the arithmetic in these two bounds yields the following tradeoff equation:

$$\text{TBOUND} = \frac{p}{2n\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}\left[\frac{\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2^2}{\text{SBOUND} + \varepsilon} + \varepsilon\right]^2.$$

The above tradeoff between the training error bound and the shrinkage bound is illustrated in Figure 6, which shows this tradeoff curve for four different values of the learning rate ε. Except for very small shrinkage levels, lower values of ε produce smaller training errors. But unlike the corresponding tradeoff curves for LS-BOOST(ε), there is a range of values of the shrinkage for which smaller values of ε actually produce larger training errors, though admittedly this range is for very small shrinkage values. For more reasonable shrinkage values, smaller values of ε will correspond to smaller values of the training error.

Part (v) of Theorems 2.1 and 3.1 presents shrinkage bounds for FS$_\varepsilon$ and LS-BOOST(ε), respectively. Let us briefly compare these bounds. Examining the shrinkage bound for LS-BOOST(ε), we can bound the left term from above by $\sqrt{k}\sqrt{\varepsilon}\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2$. We can also bound the right term from above by $\varepsilon\|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2/(1 - \sqrt{\gamma})$ where recall from Section 2 that γ is the linear convergence rate coefficient $\gamma := 1 - \frac{\varepsilon(2-\varepsilon)\lambda_{\text{pmin}}(\mathbf{X}^T\mathbf{X})}{4p}$. We may therefore alternatively write the following shrinkage bound for LS-BOOST(ε):

$$\|\hat{\beta}^k\|_1 \le \|\mathbf{X}\hat{\beta}_{\text{LS}}\|_2 \min\{\sqrt{k}\sqrt{\varepsilon}, \varepsilon/(1 - \sqrt{\gamma})\}. \tag{3.9}$$

The shrinkage bound for FS_ε is simply $k\varepsilon$. Comparing these two bounds, we observe that not only does the shrinkage bound for FS_ε grow at a faster rate as a function of k for large enough k, but also the shrinkage bound for FS_ε grows unbounded in k, unlike the right term above for the shrinkage bound of LS-BOOST(ε).

One can also compare FS_ε and LS-BOOST(ε) in terms of the efficiency with which these two methods achieve a certain pre-specified data-fidelity. In Section A.2.7 [14] we show, at least in theory, that LS-BOOST(ε) is much more efficient than FS_ε at achieving such data-fidelity and, furthermore, it does so with much better shrinkage.

Generalizations of FS_ε. Here, we briefly mention some recent work that generalize the FS_ε algorithm: [24] study extensions to incorporate non-convex penalization schemes, and [47] propose a framework generalizing FS_ε to a flexible family of convex loss functions and sparsity inducing convex regularizers.

4. Regularized correlation minimization, boosting and LASSO.

Roadmap. In this section, we introduce a new boosting algorithm, parameterized by a scalar $\delta \geq 0$, which we denote by R-$\text{FS}_{\varepsilon,\delta}$ (for Regularized incremental Forward Stagewise regression), that is obtained by incorporating a simple rescaling step to the coefficient updates in FS_ε. We then introduce a regularized version of the Correlation Minimization (CM) problem (3.7) which we refer to as RCM. We show that the adaptation of the subgradient descent algorithmic framework to the Regularized Correlation Minimization problem RCM exactly yields the algorithm R-$\text{FS}_{\varepsilon,\delta}$. The new algorithm R-$\text{FS}_{\varepsilon,\delta}$ may be interpreted as a natural extension of popular boosting algorithms like FS_ε, and has the following notable properties:

- Whereas FS_ε updates the coefficients in an additive fashion by adding a small amount ε to the coefficient most correlated with the current residuals, R-$\text{FS}_{\varepsilon,\delta}$ first shrinks *all* of the coefficients by a scaling factor $1 - \frac{\varepsilon}{\delta} < 1$ and then updates the selected coefficient in the same additive fashion as FS_ε.
- R-$\text{FS}_{\varepsilon,\delta}$ delivers $O(\varepsilon)$-accurate solutions to the LASSO in the limit as $k \to \infty$, unlike FS_ε which delivers $O(\varepsilon)$-accurate solutions to the unregularized least squares problem.
- R-$\text{FS}_{\varepsilon,\delta}$ has computational guarantees similar in spirit to the ones described in the context of FS_ε—these quantities directly inform us about the data-fidelity *vis-à-vis* shrinkage tradeoffs as a function of the number of boosting iterations and the learning rate ε.

The notion of using additional regularization along with the implicit shrinkage imparted by boosting is not new in the literature. Various interesting notions have been proposed in [9, 11, 19, 25, 51]; see also the discussion in Section A.3.4 [14]. However, the framework we present here is new. We present a unified subgradient descent framework for a class of regularized CM problems that results in algorithms that have appealing structural similarities with forward stagewise regression-type algorithms, while also being very strongly connected to the LASSO.

Boosting with additional shrinkage—R-FS$_{\varepsilon,\delta}$. Here, we give a formal description of the R-FS$_{\varepsilon,\delta}$ algorithm. R-FS$_{\varepsilon,\delta}$ is controlled by two parameters: the learning rate ε, which plays the same role as the learning rate in FS$_\varepsilon$, and the "regularization parameter" $\delta \geq \varepsilon$. Our reason for referring to δ as a regularization parameter is due to the connection between R-FS$_{\varepsilon,\delta}$ and the LASSO, which will be made clear later. The shrinkage factor, that is, the amount by which we shrink the coefficients before updating the selected coefficient, is determined as $1 - \frac{\varepsilon}{\delta}$. Supposing that we choose to update the coefficient indexed by j_k at iteration k, then the coefficient update may be written as

$$\hat{\beta}^{k+1} \leftarrow \left(1 - \frac{\varepsilon}{\delta}\right)\hat{\beta}^k + \varepsilon \cdot \operatorname{sgn}((\hat{r}^k)^T \mathbf{X}_{j_k}) e_{j_k}.$$

Below we give a concise description of R-FS$_{\varepsilon,\delta}$, including the update for the residuals that corresponds to the update for the coefficients stated above.

Algorithm: R-FS$_{\varepsilon,\delta}$

Fix a learning rate $\varepsilon > 0$, regularization parameter $\delta > 0$ (with $\varepsilon \leq \delta$), number of iterations M; and initialize at $\hat{\beta}^0 = 0$.

For $0 \leq k \leq M$, select $j_k \in \arg\max_{j\in\{1,\dots,p\}} |(\hat{r}^k)^T \mathbf{X}_j|$ and perform the update:

$$\begin{aligned} &\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon\left[\operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k})\mathbf{X}_{j_k} + \frac{1}{\delta}(\hat{r}^k - \mathbf{y})\right] \\ &\hat{\beta}^{k+1}_{j_k} \leftarrow \left(1 - \frac{\varepsilon}{\delta}\right)\hat{\beta}^k_{j_k} + \varepsilon \operatorname{sgn}((\hat{r}^k)^T\mathbf{X}_{j_k}) \quad \text{and} \quad \hat{\beta}^{k+1}_j \leftarrow \left(1 - \frac{\varepsilon}{\delta}\right)\hat{\beta}^k_j, j \neq j_k. \end{aligned} \tag{4.1}$$

Note that R-FS$_{\varepsilon,\delta}$ and FS$_\varepsilon$ are structurally very similar—and indeed when $\delta = \infty$ then R-FS$_{\varepsilon,\delta}$ is exactly FS$_\varepsilon$. Note also that R-FS$_{\varepsilon,\delta}$ shares the same upper bound on the sparsity of the regression coefficients as FS$_\varepsilon$, namely for all k it holds that: $\|\hat{\beta}^k\|_0 \leq k$. When $\delta < \infty$ then, as previously mentioned, the main structural difference between R-FS$_{\varepsilon,\delta}$ and FS$_\varepsilon$ is the additional rescaling of the coefficients by the factor $1 - \frac{\varepsilon}{\delta}$. This rescaling better controls the growth of the coefficients and, as will be demonstrated next, plays a key role in connecting R-FS$_{\varepsilon,\delta}$ to the LASSO.

Regularized correlation minimization (*RCM*) *and* LASSO. The starting point of our formal analysis of R-FS$_{\varepsilon,\delta}$ is the Correlation Minimization (CM) problem (3.7), which we now modify by introducing a regularization term that penalizes residuals that are far from the vector of observations $\mathbf{y}$. This modification leads to the following parametric family of optimization problems indexed by

$\delta \in (0, \infty]$:

$$\begin{aligned} \mathrm{RCM}_\delta : \quad & f_\delta^* := \min_r f_\delta(r) := \|\mathbf{X}^T r\|_\infty + \frac{1}{2\delta}\|r - \mathbf{y}\|_2^2 \\ & \text{s.t.} \quad r \in P_{\mathrm{res}} := \{r \in \mathbb{R}^n : r = \mathbf{y} - \mathbf{X}\beta \text{ for some } \beta \in \mathbb{R}^p\}, \end{aligned} \tag{4.2}$$

where "RCM" connotes Regularlized Correlation Minimization. Note that RCM reduces to the correlation minimization problem CM (3.7) when $\delta = \infty$. RCM may be interpreted as the problem of minimizing, over the space of residuals, the largest correlation between the residuals and the predictors plus a regularization term that penalizes residuals that are far from the response $\mathbf{y}$ (which itself can be interpreted as the residuals associated with the model $\beta = 0$).

Interestingly, as we show in [14] Section A.3.1, RCM (4.2) is equivalent to the LASSO (1.1) via duality. This equivalence provides further insight about the regularization used to obtain RCM_δ. Comparing the LASSO and RCM, notice that the space of the variables of the LASSO is the space of regression coefficients β, namely $\mathbb{R}^p$, whereas the space of the variables of RCM is the space of model residuals, namely $P_{\mathrm{res}} \subset \mathbb{R}^n$. The duality relationship shows that RCM_δ (4.2) is an equivalent characterization of the LASSO problem, just like the correlation minimization (CM) problem (3.7) is an equivalent characterization of the (unregularized) least squares problem. Recall that Proposition 3.2 showed that subgradient descent applied to the CM problem (4.2) (which is RCM_δ with $\delta = \infty$) leads to the well-known boosting algorithm FS_ε. We now extend this theme with the following proposition, which states that R-$\mathrm{FS}_{\varepsilon,\delta}$ is equivalent to subgradient descent applied to RCM_δ.

PROPOSITION 4.1. *The* R-$\mathrm{FS}_{\varepsilon,\delta}$ *algorithm is an instance of subgradient descent to solve the regularized correlation minimization* (RCM_δ) *problem* (4.2), *initialized at* $\hat{r}^0 = \mathbf{y}$, *with a constant step-size* $\alpha_k := \varepsilon$ *at each iteration.*

The proof of Proposition 4.1 is presented in [14] Section A.3.2.

4.1. R-$\mathrm{FS}_{\varepsilon,\delta}$: *Computational guarantees and their implications.*

In this subsection, we present computational guarantees and convergence properties of the boosting algorithm R-$\mathrm{FS}_{\varepsilon,\delta}$. Due to the structural equivalence between R-$\mathrm{FS}_{\varepsilon,\delta}$ and subgradient descent applied to the RCM_δ problem (4.2) (Proposition 4.1) and the close connection between RCM_δ and the LASSO (see [14], Section A.3.1), the convergence properties of R-$\mathrm{FS}_{\varepsilon,\delta}$ are naturally stated with respect to the LASSO problem (1.1). Similar to Theorem 3.1 which described such properties for FS_ε (with respect to the unregularized least squares problem), we have the following properties for R-$\mathrm{FS}_{\varepsilon,\delta}$.

THEOREM 4.1 (Convergence Properties of R-$\mathrm{FS}_{\varepsilon,\delta}$ for the LASSO). *Consider the* R-$\mathrm{FS}_{\varepsilon,\delta}$ *algorithm with learning rate* ε *and regularization parameter* $\delta \in (0, \infty)$, *where* $\varepsilon \leq \delta$. *Then the regression coefficient* $\hat{\beta}^k$ *is feasible for the*

LASSO *problem* (1.1) *for all* $k \geq 0$. *Let* $k \geq 0$ *denote a specific iteration counter. Then there exists an index* $i \in \{0, \dots, k\}$ *for which the following bounds hold*:

(i) (*training error*): $L_n(\hat{\beta}^i) - L^*_{n,\delta} \leq \frac{\delta}{n}\left[\frac{\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2^2}{2\varepsilon(k+1)} + 2\varepsilon\right]$;

(ii) (*predictions*): *for every* LASSO *solution* $\hat{\beta}^*_\delta$ *it holds that*

$$\|\mathbf{X}\hat{\beta}^i - \mathbf{X}\hat{\beta}^*_\delta\|_2 \leq \sqrt{\frac{\delta\|\mathbf{X}\hat{\beta}_{\mathrm{LS}}\|_2^2}{\varepsilon(k+1)} + 4\delta\varepsilon};$$

(iii) (ℓ_1-*shrinkage of coefficients*): $\|\hat{\beta}^i\|_1 \leq \delta\left[1 - \left(1 - \frac{\varepsilon}{\delta}\right)^k\right] \leq \delta$;

(iv) (*sparsity of coefficients*): $\|\hat{\beta}^i\|_0 \leq k$.

The proof of Theorem 4.1 is presented in the Supplementary Material [14], Section A.3.3.

Interpreting the computational guarantees. The statistical interpretations implied by the computational guarantees presented in Theorem 4.1 are analogous to those previously discussed for LS-BOOST(ε) (Theorem 2.1) and FS$_\varepsilon$ (Theorem 3.1). These guarantees inform us about the data-fidelity versus shrinkage tradeoffs as a function of the number of boosting iterations, as nicely demonstrated in Figure 7. There is, however, an important differentiation between the properties of R-FS$_{\varepsilon,\delta}$ and the properties of LS-BOOST(ε) and FS$_\varepsilon$, namely:

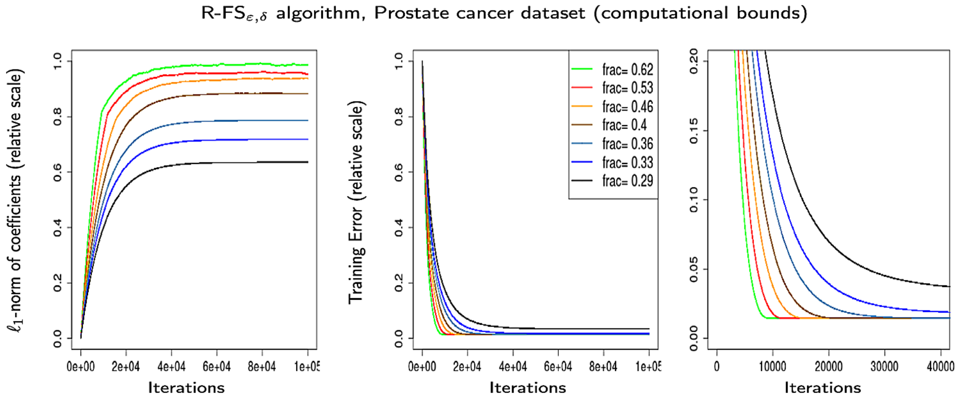

FIG. 7. *Figure showing the evolution of the* R-FS$_{\varepsilon,\delta}$ *algorithm* (*with* $\varepsilon = 10^{-4}$) *for different values of* δ, *as a function of the number of boosting iterations for the Prostate cancer dataset, with* $n = 10$, $p = 44$ (*see also Figure* 8). [*Left panel*] *shows the change of the* ℓ_1-*norm of the regression coefficients.* [*Middle panel*] *shows the evolution of the training errors, and* [*Right panel*] *is a zoomed-in version of the middle panel. Here, we took different values of* δ *given by* $\delta =$ *frac* $\times\ \delta_{\max}$, *where,* $\delta_{\max}$ *denotes the* ℓ_1-*norm of the minimum* ℓ_1-*norm least squares solution, for* 7 *different values of frac.*

- For LS-BOOST(ε) and FS$_\varepsilon$, the computational guarantees (Theorems 2.1 and 3.1) describe how the estimates make their way to an unregularized [$O(\varepsilon)$-approximate] least squares solution as a function of the number of boosting iterations.
- For R-FS$_{\varepsilon,\delta}$, our results (Theorem 4.1) characterize how the estimates approach a [$O(\varepsilon)$-approximate] LASSO solution.

Notice that like FS$_\varepsilon$, R-FS$_{\varepsilon,\delta}$ traces out a profile of regression coefficients. This is reflected in item (iii) of Theorem 4.1 which bounds the ℓ_1-shrinkage of the coefficients as a function of the number of boosting iterations k. Due to the rescaling of the coefficients, the ℓ_1-shrinkage may be bounded by a geometric series that approaches δ as k grows. Thus, there are two important aspects of the bound in item (iii): (a) the dependence on the number of boosting iterations k which characterizes model complexity during early iterations, and (b) the uniform bound of δ which applies even in the limit as $k \to \infty$ and implies that all regression coefficient iterates $\hat{\beta}^k$ are feasible for the LASSO problem (1.1).

On the other hand, item (i) characterizes the quality of the coefficients with respect to the LASSO solution, as opposed to the unregularized least squares problem as in FS$_\varepsilon$. In the limit as $k \to \infty$, item (i) implies that R-FS$_{\varepsilon,\delta}$ identifies a model with training error at most $L^*_{n,\delta} + \frac{2\delta\varepsilon}{n}$. This upper bound on the training error may be set to any prescribed error level by appropriately tuning ε; in particular, for $\varepsilon \approx 0$ and fixed $\delta > 0$ this limit is essentially $L^*_{n,\delta}$. Thus, combined with the uniform bound of δ on the ℓ_1-shrinkage, we see that the R-FS$_{\varepsilon,\delta}$ algorithm delivers the LASSO solution in the limit as $k \to \infty$.

It is important to emphasize that R-FS$_{\varepsilon,\delta}$ should not just be interpreted as an algorithm to solve the LASSO. Indeed, like FS$_\varepsilon$, the trajectory of the algorithm is important and R-FS$_{\varepsilon,\delta}$ may identify a more statistically interesting model in the interior of its profile. Thus, even if the LASSO solution for δ leads to overfitting, the R-FS$_{\varepsilon,\delta}$ updates may visit a model with better predictive performance by trading off bias and variance in a more desirable fashion suitable for the particular problem at hand.

Figure 8 shows the profiles of R-FS$_{\varepsilon,\delta}$ for different values of $\delta \leq \delta_{\max}$, where $\delta_{\max}$ is the ℓ_1-norm of the minimum ℓ_1-norm least squares solution. Curiously enough, Figure 8 shows that in some cases, the profile of R-FS$_{\varepsilon,\delta}$ bears a lot of similarities with that of the LASSO (as presented in Figure 2). However, the profiles are in general different. Indeed, R-FS$_{\varepsilon,\delta}$ imposes a uniform bound of δ on the ℓ_1-shrinkage, and so for values larger than $\delta_{\max}$ we cannot possibly expect R-FS$_{\varepsilon,\delta}$ to approximate the LASSO path. However, even if δ is taken to be sufficiently large (but finite) the profiles may be different. In this connection, it is helpful to draw the analogy between the curious similarities between the FS$_\varepsilon$ (i.e., R-FS$_{\varepsilon,\delta}$ with $\delta = \infty$) and LASSO coefficient profiles, even though the profiles are different in general.

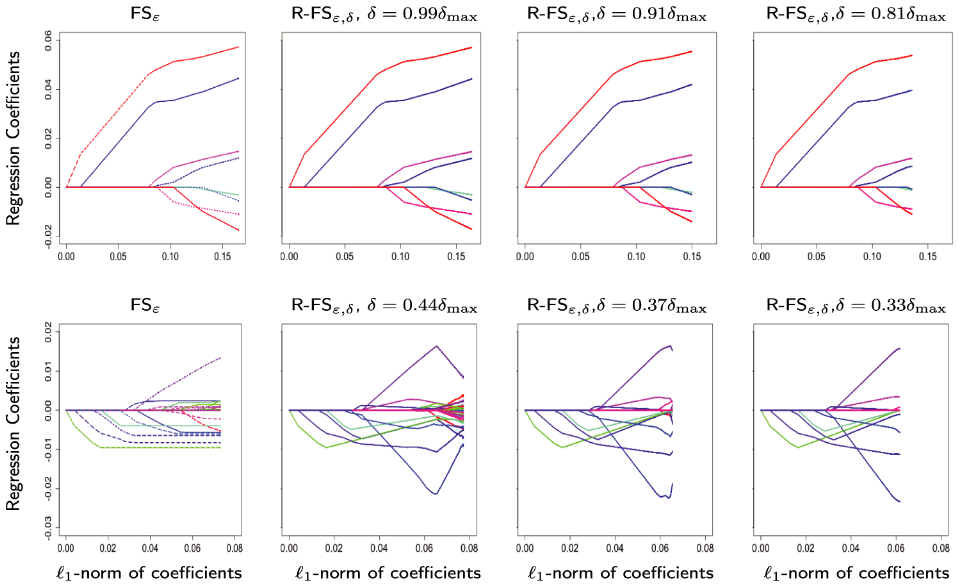

FIG. 8. *Coefficient profiles for* R-FS$_{\varepsilon,\delta}$ *as a function of the* ℓ_1*-norm of the regression coefficients, for the same datasets appearing in Figure* 2. *For each example, different values of* δ *have been considered. The left panel corresponds to the choice* $\delta = \infty$, *that is,* FS$_\varepsilon$. *In all the above cases, the algorithms were run for a maximum of* 100,000 *boosting iterations with* $\varepsilon = 10^{-4}$. [*Top Panel*] *Corresponds to the Prostate cancer dataset with* $n = 98$ *and* $p = 8$. *All the coefficient profiles look similar, and they all seem to coincide with the* LASSO *profile* (*see also Figure* 2). [*Bottom Panel*] *Shows the Prostate cancer dataset with a subset of samples* $n = 10$ *with all interactions included with* $p = 44$. *The coefficient profiles in this example are sensitive to the choice of* δ *and are seen to be more constrained towards the end of the path, for decreasing* δ *values. The profiles are different than the* LASSO *profiles, as seen in Figure* 2. *The regression coefficients at the end of the path correspond to approximate* LASSO *solutions, for the respective values of* δ.

Readers familiar with convex optimization methods will notice that R-FS$_{\varepsilon,\delta}$ bears a striking resemblance to another notable optimization algorithm: the Frank–Wolfe method [13, 15, 33]. Indeed, the update for the coefficients in R-FS$_{\varepsilon,\delta}$ is of the form

$$\hat{\beta}^{k+1} \leftarrow (1-\bar{\alpha})\hat{\beta}^k + \bar{\alpha}\tilde{\beta}^k \qquad \text{where } \tilde{\beta}^k := \delta\,\mathrm{sgn}\big((\hat{r}^k)^T \mathbf{X}_{j_k}\big)e_{j_k} \tag{4.3}$$

and $\bar{\alpha} := \frac{\varepsilon}{\delta} \in (0, 1]$, where $\tilde{\beta}^k \in \arg\min_{\beta:\|\beta\|_1 \le \delta}\{\nabla L_n(\hat{\beta}^k)^T \beta\}$, which is *exactly* of the form of the Frank–Wolfe method applied to the LASSO, with constant step-sizes $\bar{\alpha} := \frac{\varepsilon}{\delta}$. The fact that R-FS$_{\varepsilon,\delta}$ is equivalent to both the subgradient descent and Frank–Wolfe methods is no coincidence; indeed, this is a special case of a more general primal-dual equivalence between certain subgradient algorithms and the Frank–Wolfe method developed in [1]. It should be noted, however, that despite this additional equivalence, the use of constant step-sizes in Frank–Wolfe is typically not very sensible—yet turns out to be relevant in the boosting context.

We refer the reader to the Supplementary Material [14] Section A.3.5 for further discussion regarding this connection with the Frank–Wolfe method.

5. A modified forward stagewise algorithm for computing the LASSO path.

In Section 4, we introduced the boosting algorithm R-FS$_{\varepsilon,\delta}$ (which is a very close cousin of FS$_\varepsilon$) that delivers solutions to the LASSO problem (1.1) for a fixed but arbitrary δ, in the limit as $k \to \infty$ with $\varepsilon \approx 0$. Furthermore, our experiments in Section 6 suggest that R-FS$_{\varepsilon,\delta}$ may lead to estimators with good statistical properties for a wide range of values of δ, provided that the value of δ is not too small. While R-FS$_{\varepsilon,\delta}$ by itself may be considered as a regularization scheme with excellent statistical properties, the boosting profile delivered by R-FS$_{\varepsilon,\delta}$ might in some cases be different from the LASSO coefficient profile, as we saw in Figure 8. Therefore, in this section we investigate the following question: is it possible to modify the R-FS$_{\varepsilon,\delta}$ algorithm, while still retaining its basic algorithmic characteristics, so that it delivers an approximate LASSO coefficient profile for any dataset? We answer this question in the affirmative herein.

To fix ideas, let us consider producing the (approximate) LASSO path by producing a sequence of (approximate) LASSO solutions on a predefined grid of regularization parameter values δ in the interval $(0, \bar{\delta}]$ given by $0 < \bar{\delta}_0 < \bar{\delta}_1 < \cdots < \bar{\delta}_K = \bar{\delta}$. [A standard method for generating the grid points is to use a geometric sequence such as $\bar{\delta}_i = \eta^{-i} \cdot \bar{\delta}_0$ for $i = 0, \dots, K$, for some $\eta \in (0, 1)$.] Motivated by the notion of warm-starts popularly used in the statistical computing literature in the context of computing a path of LASSO solutions via coordinate descent methods [20], we propose here a slight modification of the R-FS$_{\varepsilon,\delta}$ algorithm that sequentially updates the value of δ according to the predefined grid values $\bar{\delta}_0, \bar{\delta}_1, \dots, \bar{\delta}_K = \bar{\delta}$, and does so prior to each update of $\hat{r}^i$ and $\hat{\beta}^i$. We call this method PATH-R-FS$_\varepsilon$, whose complete description is as follows:

Algorithm: PATH-R-FS$_\varepsilon$

Fix the learning rate $\varepsilon > 0$, choose values $\bar{\delta}_i$, $i = 0, \dots, K$, satisfying $0 < \bar{\delta}_0 \leq \bar{\delta}_1 \leq \cdots \leq \bar{\delta}_K \leq \bar{\delta}$ such that $\varepsilon \leq \bar{\delta}_0$. Initialize at $\hat{\beta}^0 = 0$.

For $0 \leq k \leq K$ select coefficient index $j_k \in \arg\max_{j \in \{1,\dots,p\}} |(\hat{r}^k)^T \mathbf{X}_j|$ and perform the update

$$\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon\big[\operatorname{sgn}\big((\hat{r}^k)^T \mathbf{X}_{j_k}\big)\mathbf{X}_{j_k} + (\hat{r}^k - \mathbf{y})/\bar{\delta}_k\big],$$

$$\hat{\beta}^{k+1}_{j_k} \leftarrow (1 - \varepsilon/\bar{\delta}_k)\hat{\beta}^k_{j_k} + \varepsilon \operatorname{sgn}\big((\hat{r}^k)^T \mathbf{X}_{j_k}\big) \quad \text{and}$$

$$\hat{\beta}^{k+1}_j \leftarrow (1 - \varepsilon/\bar{\delta}_k)\hat{\beta}^k_j, \qquad j \neq j_k.$$

Notice that PATH-R-FS$_\varepsilon$ retains the identical structure of a forward stagewise regression type of method, and uses the same essential update structure (3.1) of R-FS$_{\varepsilon,\delta}$. Indeed, the updates of $\hat{r}^{k+1}$ and $\hat{\beta}^{k+1}$ in PATH-R-FS$_\varepsilon$ are identical to those in (3.1) of R-FS$_{\varepsilon,\delta}$ except that they use the regularization value $\bar{\delta}_k$ at iteration k instead of the constant value of δ as in R-FS$_{\varepsilon,\delta}$.

Theoretical guarantees for PATH-R-FS$_\varepsilon$. Analogous to Theorem 4.1 for R-FS$_{\varepsilon,\delta}$, the following theorem describes properties of the PATH-R-FS$_\varepsilon$ algorithm. In particular, the theorem provides rigorous guarantees about the distance between the PATH-R-FS$_\varepsilon$ algorithm and the LASSO coefficient profiles—which apply to any general dataset.

THEOREM 5.1 (Computational Guarantees for PATH-R-FS$_\varepsilon$). *Consider the* PATH-R-FS$_\varepsilon$ *algorithm with the given learning rate* ε *and regularization parameter sequence* $\{\bar{\delta}_k\}$. *Let* $k \geq 0$ *denote the total number of iterations. Then the following holds*:

(i) (LASSO *feasibility and average training error*): *for each* $i = 0, \dots, k$, $\hat{\beta}^i$ *provides an approximate solution to the* LASSO *problem for* $\delta = \bar{\delta}_i$. *More specifically,* $\hat{\beta}^i$ *is feasible for the* LASSO *problem for* $\delta = \bar{\delta}_i$, *and satisfies the following suboptimality bound with respect to the entire boosting profile*:

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

(ii) (ℓ_1-*shrinkage of coefficients*): $\|\hat{\beta}^i\|_1 \leq \bar{\delta}_i$ *for* $i = 0, \dots, k$;

(iii) (*sparsity of coefficients*): $\|\hat{\beta}^i\|_0 \leq i$ *for* $i = 0, \dots, k$.

COROLLARY 5.1 (PATH-R-FS$_\varepsilon$ approximates the LASSO path). *For every fixed* $\varepsilon > 0$ *and* $k \to \infty$, *it holds that*

$$\limsup_{k\to\infty}\frac{1}{k+1}\sum_{i=0}^{k}\bigl(L_n(\hat{\beta}^i) - L^*_{n,\bar{\delta}_i}\bigr) \leq \frac{2\bar{\delta}\varepsilon}{n}$$

(*and the quantity on the right-hand side of the above bound goes to zero as* $\varepsilon \to 0$).

The proof of Theorem 5.1 is presented in the Supplementary Material [14] Section A.4.1.

Interpreting the computational guarantees. Let us now provide some interpretation of the results stated in Theorem 5.1. Recall that Theorem 4.1 presented bounds on the distance between the training errors achieved by the boosting algorithm R-FS$_{\varepsilon,\delta}$ and LASSO training errors for a *fixed* but arbitrary δ that is specified a priori. The message in Theorem 5.1 generalizes this notion to a *family* of LASSO solutions corresponding to a *grid* of δ values. The theorem thus quantifies how the boosting algorithm PATH-R-FS$_\varepsilon$ *simultaneously* approximates a path of LASSO solutions.

Part (i) of Theorem 5.1 first implies that the sequence of regression coefficient vectors $\{\hat{\beta}^i\}$ is feasible along the LASSO path, for the LASSO problem (1.1) for the sequence of regularization parameter values $\{\bar{\delta}_i\}$. In considering guarantees with respect to the training error, we would ideally like guarantees that hold across the entire spectrum of $\{\bar{\delta}_i\}$ values. While part (i) does not provide such strong guarantees, part (i) states that these quantities will be sufficiently small *on average*. Indeed, for a fixed ε and as $k \to \infty$, part (i) states that the average of the differences between the training errors produced by the algorithm and the optimal training errors is at most $\frac{2\bar{\delta}\varepsilon}{n}$. This nonvanishing bound (for $\varepsilon > 0$) is a consequence of the fixed learning rate ε used in PATH-R-FS$_\varepsilon$—such bounds were also observed for R-FS$_{\varepsilon,\delta}$ and FS$_\varepsilon$.

On average, the training error of the PATH-R-FS$_\varepsilon$ solutions will be sufficiently close (as controlled by the learning rate ε) to the LASSO training error for the corresponding regularization parameter grid values $\{\bar{\delta}_i\}$. And while PATH-R-FS$_\varepsilon$ provides the most amount of flexibility in terms of controlling for model complexity since it allows for *any* (monotone) sequence of regularization parameter values in the range $(0, \bar{\delta}]$, this freedom comes at the cost of weaker training error guarantees with respect to any particular $\bar{\delta}_i$ value (as opposed to R-FS$_{\varepsilon,\delta}$ which provides strong guarantees with respect to the fixed value δ). Nevertheless, part (i) of Theorem 5.1 guarantees that the training errors will be sufficiently small on average across the entire path of regularization parameter values explored by the algorithm.

It is interesting that PATH-R-FS$_\varepsilon$ approximates the LASSO path training errors, with associated shrinkage and sparsity bounds—all the while performing only boosting steps. In a sense, the price it pays for being a boosting method is that the approximation to the LASSO path is only on average over the chosen grid points (as opposed to holding simultaneously over all grid points). In contrast, classic algorithms that exactly track the piecewise-linear LASSO path require an exponential number of iterations in the worst case; see [26] for a very general result in this regard. Tibshirani [47] proposes “shrunken stagewise” which is shown to deliver LASSO solutions in a limiting sense as certain parameters go to zero and under technical assumptions. On the other hand, there are several efficient LASSO path algorithms with computational guarantees that hold approximately over the *entire* LASSO path; such methods typically choose the $\{\bar{\delta}_i\}$ values adaptively; see [27] and [28] for some general results in this regard, and in particular [29] for optimal general complexity results in this context. Similar adaptive choices of $\{\bar{\delta}_i\}$ values are also studied by [47] for computing an approximate LASSO path.

6. Some computational experiments.

We consider an array of examples exploring statistical properties of the different boosting algorithms studied herein. We consider different types of synthetic and real datasets, which are briefly described here.

Synthetic datasets. We considered synthetically generated datasets of the following types:

- **Eg-A.** Here, the data matrix $\mathbf{X}$ is generated from a multivariate normal distribution, that is, for each $i = 1, \dots, n$, $\mathbf{x}_i \sim \mathrm{MVN}(0, \Sigma)$. Here, $\mathbf{x}_i$ denotes the ith row of $\mathbf{X}$ and $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ has all off-diagonal entries equal to ρ and all diagonal entries equal to one. The response $\mathbf{y} \in \mathbb{R}^n$ is generated as $\mathbf{y} = \mathbf{X}\beta^{\mathrm{pop}} + \epsilon$, where $\epsilon_i \overset{\mathrm{iid}}{\sim} N(0, \sigma^2)$. The underlying regression coefficient was taken to be sparse with $\beta_i^{\mathrm{pop}} = 1$ for all $i \le 5$ and $\beta_i^{\mathrm{pop}} = 0$ otherwise. σ^2 is chosen so as to control the signal to noise ratio $\mathrm{SNR} := \mathrm{Var}(\mathbf{x}'\beta)/\sigma^2$.

  Different values of SNR, n, p and ρ were taken and they have been specified in our results when and where appropriate.
- **Eg-B.** Here, the datasets are generated similar to above, with $\beta_i^{\mathrm{pop}} = 1$ for $i \le 10$ and $\beta_i^{\mathrm{pop}} = 0$ otherwise. We took $\mathrm{SNR} = 1$ in this example.

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

- *Leukemia dataset.* This dataset, taken from [10], was processed to have $n = 72$ and $p = 500$. $\mathbf{y}$ was created as $\mathbf{y} = \mathbf{X}\beta^{\mathrm{pop}} + \epsilon$; with $\beta_i^{\mathrm{pop}} = 1$ for all $i \le 10$ and zero otherwise.
- *Golub dataset.* This dataset, taken from the R package `mpm`, was processed to have $n = 73$ and $p = 500$, with artificial responses generated as above.
- *Khan dataset.* This dataset, taken from the website of [31], was processed to have $n = 73$ and $p = 500$, with artificial responses generated as above.
- *Prostate dataset.* This dataset, analyzed in [12], was processed to create three types of different datasets: (a) the original dataset with $n = 97$ and $p = 8$, (b) a dataset with $n = 97$ and $p = 44$, formed by extending the covariate space to include second-order interactions, and (c) a third dataset with $n = 10$ and $p = 44$, formed by subsampling the previous dataset.

For more detail on the above datasets, we refer the reader to the Supplementary Material [14] Section A.5.

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

6.1. *Statistical properties of boosting algorithms*: *An empirical study.* We performed some experiments to better understand the statistical behavior of the different boosting methods described in this paper. We summarize our findings here; for details (including tables, figures and discussions), we refer the reader to the Supplementary Material Section A.5 [14].

Sensitivity of the learning rate in LS-BOOST(ε) *and* FS_ε. We explored how the training and test errors for LS-BOOST(ε) and FS_ε change as a function of the number of boosting iterations and the learning rate. We observed that the best predictive models were sensitive to the choice of ε—the best models were obtained at values larger than zero and smaller than one. When compared to LASSO solutions, stepwise regression [12] and FS_0 [12]; FS_ε and LS-BOOST(ε) were found to be as good as the others, and in some cases were better than the rest.

Statistical properties of R-$\text{FS}_{\varepsilon,\delta}$, LASSO *solutions, and* FS_ε: *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 (solving the) LASSO. We found that when δ was 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 and the predictive performance was as good as, and in some cases better than, the LASSO solution. We observed 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 play a more important role. The best models delivered by R-$\text{FS}_{\varepsilon,\delta}$ were more sparse than FS_ε.

Acknowledgements. The authors would like to thank Alexandre Belloni, Jerome Friedman, Trevor Hastie, Arian Maleki and Tomaso Poggio for helpful discussions and encouragement. A preliminary unpublished version of some of the results herein was posted on the ArXiv [16]. The authors would like to thank the Editor, Associate Editor and anonymous referees for helpful comments that have improved the paper.

SUPPLEMENTARY MATERIAL

Supplement to "A new perspective on boosting in linear regression via subgradient optimization and relatives" (DOI: 10.1214/16-AOS1505SUPP; .pdf). Additional proofs, technical details, figures and tables are provided in the Supplementary Section.

REFERENCES

[1] BACH, F. (2015). Duality between subgradient and conditional gradient methods. *SIAM J. Optim.* **25** 115–129. MR3296634

[2] BERTSEKAS, D. (1999). *Nonlinear Programming*. Athena Scientific, Belmont, MA.

[3] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

[4] BREIMAN, L. (1998). Arcing classifiers. *Ann. Statist.* **26** 801–849. MR1635406

[5] BREIMAN, L. (1999). Prediction games and arcing algorithms. *Neural Comput.* **11** 1493–1517.

[6] BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* 559–583.

[7] BÜHLMANN, P. and HOTHORN, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statist. Sci.* **22** 477–505. MR2420454

[8] BÜHLMANN, P. and YU, B. (2003). Boosting with the L2 loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339.

[9] BÜHLMANN, P. and YU, B. (2006). Sparse boosting. *J. Mach. Learn. Res.* **7** 1001–1024.

[10] DETTLING, M. and BÜHLMANN, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19** 1061–1069.

[11] DUCHI, J. and SINGER, Y. (2009). Boosting with structural sparsity. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 297–304. ACM, New York.

[12] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407–499. MR2060166

[13] FRANK, M. and WOLFE, P. (1956). An algorithm for quadratic programming. *Nav. Res. Logist. Q.* **3** 95–110. MR0089102

[14] FREUND, R., GRIGAS, P. and MAZUMDER, R. (2017). Supplement to "A new perspective on boosting in linear regression via subgradient optimization and relatives." DOI:10.1214/16-AOS1505SUPP.

[15] FREUND, R. M. and GRIGAS, P. (2014). New analysis and results for the Frank–Wolfe method. *Math. Program.* To appear.

[16] FREUND, R. M., GRIGAS, P. and MAZUMDER, R. (2013). AdaBoost and forward stagewise regression are first-order convex optimization methods. Preprint. Available at arXiv:1307.1192.

[17] FREUND, Y. (1995). Boosting a weak learning algorithm by majority. *Inform. and Comput.* **121** 256–285. MR1348530

[18] FREUND, Y. and SCHAPIRE, R. (1996). Experiments with a new boosting algorithm. In *Machine Learning*: *Proceedings of the Thirteenth International Conference* 148–156. Morgan Kauffman, San Francisco.

[19] FRIEDMAN, J. (2008). Fast sparse regression and classification. Technical Report, Dept. Statistics, Stanford Univ.

[20] FRIEDMAN, J., HASTIE, T., HOEFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **2** 302–332.

[21] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting. *Ann. Statist.* **28** 337–407. MR1790002

[22] FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328

[23] FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328

[24] FRIEDMAN, J. H. (2012). Fast sparse regression and classification. *Int. J. Forecast.* **28** 722–738.

[25] FRIEDMAN, J. H. and POPESCU, B. E. (2003). Importance sampled learning ensembles. *J. Mach. Learn. Res.* **94305**.

[26] GÄRTNER, B., JAGGI, M. and MARIA, C. (2012). An exponential lower bound on the complexity of regularization paths. *J. Comput. Geom.* **3** 168–195. MR3030324

[27] GIESEN, J., JAGGI, M. and LAUE, S. (2012). Optimizing over the Growing Spectrahedron. In *Algorithms—ESA* 2012: 20*th Annual European Symposium*, *Ljubljana*, *Slovenia*, *September* 10–12, 2012. *Proceedings* 503–514. Springer, Berlin.

[28] GIESEN, J., JAGGI, M. and LAUE, S. (2012). Approximating parameterized convex optimization problems. *ACM Trans. Algorithms* **9** Art. 10, 17. MR3008305

[29] GIESEN, J., MUELLER, J., LAUE, S. and SWIERCY, S. (2012). Approximating concavely parameterized optimization problems. In *Advances in Neural Information Processing Systems* **25** (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 2105–2113. Curran Associates, Red Hook, NY.

[30] HASTIE, T., TAYLOR, J., TIBSHIRANI, R. and WALTHER, G. (2007). Forward stagewise regression and the monotone lasso. *Electron. J. Stat.* **1** 1–29.

[31] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer, New York.

[32] HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. Chapman & Hall, London. MR1082147

[33] JAGGI, M. (2013). Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *Proceedings of the* 30*th International Conference on Machine Learning* (*ICML*-13) 427–435.

[34] MALLAT, S. G. and ZHANG, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans*. *Signal Process*. **41** 3397–3415.

[35] MASON, L., BAXTER, J., BARTLETT, P. and FREAN, M. (2000). Boosting algorithms as gradient descent **12** 512–518.

[36] MILLER, A. (2002). *Subset Selection in Regression*. CRC Press, Boca Raton, FL.

[37] NESTEROV, Y. E. (2003). *Introductory Lectures on Convex Optimization*: *A Basic Course*. *Applied Optimization* **87**. Kluwer Academic, Boston, MA.

[38] PATI, Y. C., REZAIIFAR, R. and KRISHNAPRASAD, P. S. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conference Record of the Twenty-Seventh Asilomar Conference on Signals*, *Systems and Computers*, 1993 40–44. IEEE, New York.

[39] POLYAK, B. (1987). *Introduction to Optimization*. Optimization Software, New York.

[40] RÄTSCH, G., ONODA, T. and MÜLLER, K.-R. (2001). Soft margins for AdaBoost. *Mach*. *Learn*. **42** 287–320.

[41] ROSSET, S., SWIRSZCZ, G., SREBRO, N. and ZHU, J. (2007). ℓ_1 regularization in infinite dimensional feature spaces. In *Conference on Learning Theory* 544–558. Springer, Berlin.

[42] ROSSET, S., ZHU, J. and HASTIE, T. (2003/2004). Boosting as a regularized path to a maximum margin classifier. *J*. *Mach*. *Learn*. *Res*. **5** 941–973. MR2248005

[43] SCHAPIRE, R. (1990). The strength of weak learnability. *Mach*. *Learn*. **5** 197–227.

[44] SCHAPIRE, R. E. and FREUND, Y. (2012). *Boosting*: *Foundations and Algorithms*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2920188

[45] SHOR, N. Z. (1985). *Minimization Methods for Nondifferentiable Functions*. *Springer Series in Computational Mathematics* **3**. Springer, Berlin. MR0775136

[46] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **58** 267–288. MR1379242

[47] TIBSHIRANI, R. J. (2015). A general framework for fast stagewise algorithms. *J*. *Mach*. *Learn*. *Res*. **16** 2543–2588.

[48] TUKEY, J. (1977). *Exploratory Data Analysis*. Addison-Wesley, Boston, MA.

[49] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[50] WEISBERG, S. (1980). *Applied Linear Regression*. Wiley, New York. MR0591462

[51] ZHAO, P. and YU, B. (2007). Stagewise lasso. *J*. *Mach*. *Learn*. *Res*. **8** 2701–2726.

R. M. FREUND
R. MAZUMDER
MIT SLOAN SCHOOL OF MANAGEMENT
AND OPERATIONS RESEARCH CENTER
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS
USA
E-MAIL: rfreund@mit.edu
rahulmaz@mit.edu

P. GRIGAS
DEPARTMENT OF INDUSTRIAL ENGINEERING
AND OPERATIONS RESEARCH
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA
USA
E-MAIL: pgrigas@berkeley.edu