A New Perspective on Boosting in Linear Regression via Subgradient Optimization and Relatives

Robert Freund, MIT

Joint with Paul Grigas (MIT) and Rahul Mazumder (MIT)

Princeton University

April 2016
Paper under review:

“A New Perspective on Boosting in Linear Regression via Subgradient Optimization and Relatives”
Outline of Topics

- Optimization Review
- Motivation/Overview/Perspective on Boosting
- Optimization informs Boosting Algorithms and leads to New Boosting Methods
  - Computational/Statistical Guarantees for Boosting Algorithms in Linear Regression
    - Incremental Forward Stagewise Regression
    - Least Squares Boosting
  - A New Boosting Algorithm (dubbed “Regularized Forward Stagewise”) directly connected to the LASSO
    - algorithm is a minor variant of Incremental Forward Stagewise
    - Computational/Statistical Guarantees for the LASSO
Review of Some Optimization Methodology
Review of Some Optimization Methodology

1. Steepest Descent method
2. Subgradient Descent method
3. Steepest Descent in the $\ell_1$-norm
4. (convex) Quadratic Functions
The problem of interest is:

\[ f^* := \min_x f(x) \quad \text{s.t.} \quad x \in \mathbb{R}^n \]

where \( f(x) \) is convex.

Steepest Descent method for minimizing \( f(x) \) when \( f(\cdot) \) is differentiable

Initialize at \( x_0 \in \mathbb{R}^n, k \leftarrow 0 \).

At iteration \( k \):

1. Compute gradient \( g_k = \nabla f(x_k) \).
2. Choose step-size \( \alpha_k \).
3. Set \( x_{k+1} \leftarrow x_k - \alpha_k g_k \).
Subgradients when $f(\cdot)$ is not differentiable

$g$ is a subgradient of $f(\cdot)$ at $x$ if:

$$f(y) \geq f(x) + g^T(y - x) \text{ for all } x, y$$

$\partial f(x)$ is the set of subgradients of $f(\cdot)$ at $x$
Subgradient Descent method

\[ f^* := \min_x f(x) \quad \text{s.t.} \quad x \in \mathbb{R}^n \]

Subgradient Descent method for minimizing \( f(x) \) on \( \mathbb{R}^n \)

Initialize at \( x_0 \in \mathbb{R}^n, k \leftarrow 0 \).

At iteration \( k \):

1. Compute a subgradient \( g_k \) of \( f(x_k) \).
2. Choose step-size \( \alpha_k \).
3. Set \( x_{k+1} \leftarrow x_k - \alpha_k g_k \).
Computational Guarantees for Subgradient Descent

For each $k \geq 0$ and for any $x \in \mathbb{R}^n$, the following inequality holds:

$$
\min_{i \in \{0, \ldots, k\}} f(x^i) - f(x) \leq \frac{\|x - x^0\|_2^2 + L_f^2 \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i}
$$

Here $L_f$ is the Lipschitz constant for the function $f(\cdot)$:

$$
\|f(x) - f(y)\| \leq L_f \|x - y\|_2 \quad \text{for any } x, y
$$
Steepest Descent in the $\ell_1$-norm

\[ f^* := \min_x f(x) \quad \text{s.t.} \quad x \in \mathbb{R}^n \]

Steepest Descent method in the $\ell_1$-norm when $f(\cdot)$ is differentiable

Initialize at $x_0 \in \mathbb{R}^n$, $k \leftarrow 0$.

At iteration $k$:

1. **Compute gradient** $\nabla f(x_k)$.
2. **Compute direction**: $d_k \leftarrow \arg \max_{\|d\|_1 \leq 1} \{ \nabla f(x_k)^T d \}$
3. **Choose step-size** $\alpha_k$.
4. **Set** $x_{k+1} \leftarrow x_k - \alpha_k d_k$.

This is also a **coordinate descent method** since only one coordinate is changed at each iteration.
(convex) Quadratic Functions

Let $f^* := \min_x f(x) := \frac{1}{2}x^T Qx + q^T x + q_0$

$Q$ is symmetric and positive semidefinite

Let $\lambda_{p\text{min}}(Q)$ denote the smallest \textit{positive} eigenvalue of $Q$

Two useful properties of convex quadratic functions

If $f^* > -\infty$, then for any given $x$, there exists an optimal solution $x^*$ for which

$$\|x - x^*\|_2 \leq \sqrt{\frac{2(f(x) - f^*)}{\lambda_{p\text{min}}(Q)}}.$$

Also, it holds that

$$\|\nabla f(x)\|_2 \geq \sqrt{\frac{\lambda_{p\text{min}}(Q) \cdot (f(x) - f^*)}{2}}.$$
Motivation/Overview
High-dimensional Linear Regression

- Linear regression model in the high-dimensional setting

\[ y \approx X \beta \]
Consider the linear regression model

\[ y = X\beta + e \]

- \( y \in \mathbb{R}^n \) is given response data (mean centered)
- \( X \in \mathbb{R}^{n \times p} \) is the given model matrix (each column is a predictor)
- \( \beta \in \mathbb{R}^p \) are the coefficients
- \( e \in \mathbb{R}^n \) is noise

Assume that the columns \( X_j, \ j = 1, \ldots, p \), have been mean centered and standardized to have unit \( \ell_2 \) norm
Example 1: Computational Genomics

Linear regression model:

\[ y = X\beta + e \]

Predict disease susceptibility based on genome sequence

The number of samples \( n \) may be small \((10^2 - 10^3)\) relative to the number of predictors \( p \) \((10^5 - 10^7)\)

It is reasonable to suppose that most genes are irrelevant to predicting susceptibility for a particular disease, i.e., the true model \( \beta^* \) is sparse.
Example 2: Boosting Regression Trees

A regression tree:

\[ x_1 \geq 3? \]

\[ y = 2.7 \quad x_2 < -1? \]

\[ y = 1.1 \quad y = 1.5 \]

Think of enumerating all \( p \) possible regression trees based on some underlying features.

Given \( n \) samples (including output data \( y \) and some features), fill column \( X_j \) with the output of regression tree \( j \) on the \( n \) samples.

We would like to find a “good” linear combination of regression trees \( X\beta \).
The Boosting Approach

In boosting, the goal is to combine “weak” models to form an accurate and predictive model.

A weak model can be a single feature or a more complex model based on a set of underlying features.

Often the number of weak models $p$ is exponentially large (e.g., the number of possible regression trees based on $f$ features yields $p \sim 2^f$).
Linear Regression Aspirations

Linear regression model:

\[ y = X\beta + e \]

Some aspirations in the high-dimensional regime with \( p \gg 0, \ n \gg 0 \) and often \( p > n \):

- Good predictive performance (on out-of-sample data)
- Interpretability/compression via sparsity in the coefficients (\( \|\beta\|_0 := \) number of non-zero coefficients of \( \beta \) is small)

(In this regime, the classical least-squares estimator \( \beta_{LS} \) will overfit the training data)
Least-Squares Optimization

\[ \text{LS} : \quad L_n^* := \min_{\beta} \quad L_n(\beta) := \frac{1}{2n} \| y - X\beta \|^2_2 \]

Any solution of the least-squares problem \( \beta_{LS} \) satisfies:

\[ X^T r_{LS} = 0 \quad \text{where} \quad r_{LS} = y - X\beta_{LS} \]

(If \( X^TX \) is invertible, then \( \beta_{LS} = (X^TX)^{-1}X^Ty \).)
Sparse Least Squares: Best Subset Selection

Aspirations:

- Good predictive performance
- Interpretability/compression via sparsity in the coefficients

One way to (hopefully) achieve these aspirations is to solve the best subset selection problem:

**Best Subset Selection [Miller 2002]**

\[
\begin{align*}
\text{BSS}_k^* &:= \min_{\beta} \quad L_n(\beta) := \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \\
\text{s.t.} & \quad \|\beta\|_0 \leq k
\end{align*}
\]

Recall \(\|\beta\|_0 := \text{number of non-zero coefficients of } \beta\)

This problem is a severe computational challenge (essentially intractable) when \(p\) is large
Sparse Least Squares: the LASSO

Best subset selection is generally computationally intractable when \( p \) is large

Instead, consider using the LASSO:

The LASSO (in constraint-mode) [Tibshirani 1996], [Chen et al. 1998], [Hastie et al. 2015]

\[
\text{LASSO}_\delta^* := \min_{\beta} \quad L_n(\beta) := \frac{1}{2n} \| y - X\beta \|_2^2 \quad \text{s.t.} \quad \| \beta \|_1 \leq \delta
\]

Optimal solutions of the LASSO are often sparse (or sparse enough) and generalize well [Hastie et al. 2015]
Boosting in Linear Regression

Boosting is an alternative to the LASSO:

- Reasonable assurance that our aspirations (good prediction, sparsity) are met
- Scales well when $p \gg 0$, $n \gg 0$ and $p > n$
- Yields well-structured solutions

We consider two prototypical/iconic boosting algorithms in linear regression:

- Incremental Forward Stagewise Regression ($FS_\varepsilon$) [Efron et al. 2004]
- Least Squares Boosting ($LS\text{-}\text{Boost}(\varepsilon)$) [Friedman 2001]
Brief History of Boosting

Boosting has been a topic of significant importance in both the statistics and machine learning communities

- Originally designed for classification [Schapire 1990], [Y. Freund 1995], [Y. Freund and Schapire 1996], [Friedman et al. 2000]
- Later applied to regression [Friedman 2001], [Efron et al. 2004], [Bühlmann 2006], [Hastie et al. 2009]
- Often performs exceptionally well in both instances

However, as compared to the LASSO, there has been much less work understanding why boosting performs well
FS\_\varepsilon Boosting Algorithm
The FS$_{\varepsilon}$ Boosting Algorithm for Linear Regression

To motivate FS$_{\varepsilon}$, consider the unregularized least-squares problem:

$$\text{LS} : \quad L_n^* := \min_{\beta} \quad L_n(\beta) := \frac{1}{2n} \|y - X\beta\|^2$$

Any solution of the least-squares problem $\beta_{LS}$ satisfies:

$$X^T r_{LS} = 0 \quad \text{where} \quad r_{LS} = y - X\beta_{LS}$$

(If $X^TX$ is invertible, then $\beta_{LS} = (X^TX)^{-1}X^T y$)
Least-squares optimality conditions:

$$X^T r_{LS} = 0 \quad \text{where} \quad r_{LS} = y - X\beta_{LS}$$

Incremental Forward Stagewise Regression (FS$_\varepsilon$) is a simple and intuitive boosting algorithm for linear regression:

Start with $\beta^0 \leftarrow 0$, and hence $r^0 \leftarrow y$. Fix $\varepsilon > 0$.

Given $\beta^k$ and $r^k := y - X\beta^k$, determine the weak model $X_j$ most correlated with the current residuals $r^k$:

$$j_k \leftarrow \arg\max_{j \in \{1,...,p\}} |(r^k)^T X_j|$$

(Only) adjust $\beta_{j_k}^k$ by $\pm \varepsilon$ depending on $\text{sgn}((r^k)^T X_{j_k})$
Incremental Forward Stagewise Regression Algorithm ($\text{FS}_\varepsilon$)

**FS$_\varepsilon$ Algorithm**

Initialize at $r^0 = y$, $\beta^0 = 0$, $k = 0$, set $\varepsilon > 0$

At iteration $k \geq 0$:

- Compute:

$$j_k \in \arg \max_{j \in \{1, \ldots, p\}} |(r^k)^T X_j|$$

- Set:

$$\beta^{k+1} \leftarrow \beta^k + \varepsilon \sgn((r^k)^T X_{j_k}) e^{j_k}$$

$$r^{k+1} \leftarrow r^k - \varepsilon \sgn((r^k)^T X_{j_k}) X_{j_k}$$

The parameter $\varepsilon$ is called the **learning rate** or **shrinkage factor**
FS_\varepsilon \text{ in Action}

Leukemia dataset with \( p = 500 \)

Dashed lines; Training set errors

Solid lines: Test set errors

Train/Test Errors for FS_\varepsilon

Number of Iterations

Train/Test Errors (in relative scale)

\begin{itemize}
  \item eps= 0.03
  \item eps= 0.11
  \item eps= 0.17
  \item eps= 0.25
\end{itemize}
The parameter $\varepsilon$ is called the learning rate or shrinkage factor.
Shrinkage and Sparsity Properties of $FS_\varepsilon$

Notice that $FS_\varepsilon$ adds at most one new predictor with weight $\varepsilon$ at every iteration, thus:

$$\|\beta^k\|_1 \leq k\varepsilon \quad \text{and} \quad \|\beta^k\|_0 \leq k.$$ 

Therefore $FS_\varepsilon$ controls shrinkage and sparsity.

Indeed, $FS_\varepsilon$ imparts implicit regularization, dictated by the learning rate $\varepsilon$ and the total number of iterations $k$.

How does $FS_\varepsilon$ control bias?
Regularization and Data-Fidelity Metrics

Shrinkage and Sparsity Properties for FS$_{\varepsilon}$

$$\|\beta^k\|_1 \leq k\varepsilon \quad \text{and} \quad \|\beta^k\|_0 \leq k.$$ 

Other statistical metrics (measures of data-fidelity) we should be interested in:

- Training error: $L_n(\beta^k) - L_\ast$
- Coefficient error: $\|\beta^k - \beta_{LS}\|_2$
- Prediction error: $\|X\beta^k - X\beta_{LS}\|_2$

Let us see what we can say about these metrics . . .
Convergence Properties of $\text{FS}_\varepsilon$

**Theorem: Convergence Properties of $\text{FS}_\varepsilon$**

With the constant shrinkage factor $\varepsilon$, after $k$ iterations there exists $i \leq k$ for which:

(i) (training error): $L_n(\beta^i) - L^*_n \leq \varepsilon$

(ii) (regression coefficients): there exists a solution $\beta^i_{LS}$ for which $\|\beta^i - \beta^i_{LS}\|_2 \leq \varepsilon$

(iii) (predictions): for every least-squares solution solution $\beta_{LS}$ it holds that $\|X\beta^i - X\beta_{LS}\|_2 \leq \varepsilon$

(iv) (correlation values): $\|X^T r^i\|_\infty \leq \varepsilon$

(v) (sparsity and shrinkage): $\|\beta^i\|_0 \leq k$ and $\|\beta^i\|_1 \leq k\varepsilon$

$\lambda_{p\min}(X^T X) > 0$ is the smallest positive eigenvalue of $X^T X$
Theorem: Convergence Properties of $\text{FS}_\varepsilon$

With the constant shrinkage factor $\varepsilon$, after $k$ iterations there exists $i \leq k$ for which:

(i) (training error): $L_n(\beta^i) - L^*_n \leq \frac{p}{2n(\lambda_{pmin}(X^TX))} \left[ \frac{\|X_\beta_{LS}\|_2^2}{\varepsilon(k+1)} + \varepsilon \right]^2$

(ii) (regression coefficients): there exists a solution $\beta^i_{LS}$ for which $\|\beta^i - \beta^i_{LS}\|_2 \leq \varepsilon$

(iii) (predictions): for every least-squares solution solution $\beta_{LS}$ it holds that $\|X\beta^i - X\beta_{LS}\|_2 \leq \varepsilon$

(iv) (correlation values): $\|X^T r^i\|_\infty \leq \varepsilon$

(v) (sparsity and shrinkage): $\|\beta^i\|_0 \leq k$ and $\|\beta^i\|_1 \leq k\varepsilon$

$\lambda_{pmin}(X^TX) > 0$ is the smallest positive eigenvalue of $X^TX$. 
Optimization Review
Motivation/Overview
FS\(_\varepsilon\) Boosting Algorithm
LS-Boost
Regularized FS\(_\varepsilon\) and LASSO
Summary

Convergence Properties of FS\(_\varepsilon\)

**Theorem: Convergence Properties of FS\(_\varepsilon\)**

With the constant shrinkage factor \(\varepsilon\), after \(k\) iterations there exists \(i \leq k\) for which:

(i) (training error): \(L_n(\beta^i) - L^*_n \leq \frac{p}{2n(\lambda_{p_{\text{min}}}(X^TX))} \left[ \frac{\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)^2} + \varepsilon \right]^2\)

(ii) (regression coefficients): there exists a solution \(\beta_{LS}^i\) for which
\[
\|\beta^i - \beta_{LS}^i\|_2 \leq \frac{\sqrt{p}}{(\lambda_{p_{\text{min}}}(X^TX))} \left[ \frac{\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)^2} + \varepsilon \right]
\]

(iii) (predictions): for every least-squares solution \(\beta_{LS}\) it holds that
\[
\|X\beta^i - X\beta_{LS}\|_2 \leq ?
\]

(iv) (correlation values): \(\|X^T r^i\|_\infty \leq ?\)

(v) (sparsity and shrinkage): \(\|\beta^i\|_0 \leq k\) and \(\|\beta^i\|_1 \leq k\varepsilon\)

\(\lambda_{p_{\text{min}}}(X^TX) > 0\) is the smallest positive eigenvalue of \(X^TX\)
Theorem: Convergence Properties of $\text{FS}_\varepsilon$

With the constant shrinkage factor $\varepsilon$, after $k$ iterations there exists $i \leq k$ for which:

(i) (training error): $L_n(\beta^i) - L^*_n \leq \frac{p}{2n(\lambda_{pmin}(X^TX))} \left[ \frac{\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)} + \varepsilon \right]^2$

(ii) (regression coefficients): there exists a solution $\beta^i_{LS}$ for which $\|\beta^i - \beta^i_{LS}\|_2 \leq \frac{\sqrt{p}}{(\lambda_{pmin}(X^TX))} \left[ \frac{\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)} + \varepsilon \right]$

(iii) (predictions): for every least-squares solution solution $\beta_{LS}$ it holds that $\|X\beta^i - X\beta_{LS}\|_2 \leq \frac{\sqrt{p}}{\sqrt{\lambda_{pmin}(X^TX)}} \left[ \frac{\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)} + \varepsilon \right]$

(iv) (correlation values): $\|X^Tr^i\|_\infty \leq \ ?$

(v) (sparsity and shrinkage): $\|\beta^i\|_0 \leq k$ and $\|\beta^i\|_1 \leq k\varepsilon$

$\lambda_{pmin}(X^TX) > 0$ is the smallest positive eigenvalue of $X^TX$
Theorem: Convergence Properties of $\text{FS}_\varepsilon$

With the constant shrinkage factor $\varepsilon$, after $k$ iterations there exists $i \leq k$ for which:

(i) (training error): $L_n(\beta^i) - L^*_n \leq \frac{p}{2n(\lambda_{p_{\min}}(X^TX))} \left[ \frac{\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)} + \varepsilon \right]^2$

(ii) (regression coefficients): there exists a solution $\beta^i_{LS}$ for which
$\|\beta^i - \beta^i_{LS}\|_2 \leq \frac{\sqrt{p}}{\lambda_{p_{\min}}(X^TX)} \left[ \frac{\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)} + \varepsilon \right]$

(iii) (predictions): for every least-squares solution solution $\beta_{LS}$ it holds that
$\|X\beta^i - X\beta_{LS}\|_2 \leq \frac{\sqrt{p}}{\sqrt{\lambda_{p_{\min}}(X^TX)}} \left[ \frac{\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)} + \varepsilon \right]$

(iv) (correlation values): $\|X^T r^i\|_\infty \leq \left[ \frac{\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)} + \varepsilon \right]$

(v) (sparsity and shrinkage): $\|\beta^i\|_0 \leq k$ and $\|\beta^i\|_1 \leq k\varepsilon$

$\lambda_{p_{\min}}(X^TX) > 0$ is the smallest positive eigenvalue of $X^TX$
Theorem: Convergence Properties of FS$_\varepsilon$

With the constant shrinkage factor $\varepsilon$, after $k$ iterations there exists $i \leq k$ for which:

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

(ii) (regression coefficients): there exists a solution $\beta^i_{LS}$ for which

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

(iii) (predictions): for every least-squares solution solution $\beta_{LS}$ it holds that

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

(iv) (correlation values): $\|X^T r^i\|_\infty \leq \left[ \frac{\|X\beta_{LS}\|^2_2}{\varepsilon(k+1)} + \varepsilon \right]$

(v) (sparsity and shrinkage): $\|\beta^i\|_0 \leq k$ and $\|\beta^i\|_1 \leq k\varepsilon$

$\lambda_{\text{pmin}}(X^T X) > 0$ is the smallest positive eigenvalue of $X^T X$
Theorem: Computational Guarantees for FS$_\varepsilon$

With the constant shrinkage factor $\varepsilon$, after $k$ iterations there exists $i \leq k$ for which:

(i) (training error): $L_n(\beta^i) - L^* \leq \frac{p}{2n(\lambda_{pmin}(X^TX))} \left[ \frac{\|X\beta_{LS}\|^2}{\varepsilon(k+1)} + \varepsilon \right]^2$

(ii) (regression coefficients): there exists a solution $\beta_{LS}^i$ for which
$\|\beta^i - \beta_{LS}^i\|_2 \leq \frac{\sqrt{p}}{(\lambda_{pmin}(X^TX))} \left[ \frac{\|X\beta_{LS}\|^2}{\varepsilon(k+1)} + \varepsilon \right]$

(iii) (predictions): for every solution $\beta_{LS}$ it holds that
$\|X\beta^i - X\beta_{LS}\|_2 \leq \frac{\sqrt{p}}{\sqrt{(\lambda_{pmin}(X^TX))}} \left[ \frac{\|X\beta_{LS}\|^2}{\varepsilon(k+1)} + \varepsilon \right]$

(iv) (correlation values): $\|X^Tr^i\|_\infty \leq \left[ \frac{\|X\beta_{LS}\|^2}{\varepsilon(k+1)} + \varepsilon \right]$

(v) (sparsity and shrinkage): $\|\beta^i\|_0 \leq k$ and $\|\beta^i\|_1 \leq k\varepsilon$

$\lambda_{pmin}(X^TX) > 0$ is the smallest positive eigenvalue of $X^TX$
FS$_\varepsilon$ Has Explicit Computational Guarantees

**Theorem: Computational Guarantees for FS$_\varepsilon$**

With the constant shrinkage factor $\varepsilon$, after $k$ iterations there exists $i \leq k$ for which:

1. **(training error):** $L_n(\beta^i) - L^*_n \leq \frac{p}{2n(\lambda_{p_{\text{min}}}(X^TX))} \left[ \frac{\|X\beta_{LS}\_i\|^2}{\varepsilon(k+1)} + \varepsilon \right]^2$

2. **(regression coefficients):** there exists a solution $\beta_{LS}^i$ for which
   \[ \|\beta^i - \beta_{LS}^i\|_2 \leq \frac{\sqrt{p}}{(\lambda_{p_{\text{min}}}(X^TX))} \left[ \frac{\|X\beta_{LS}\_i\|^2}{\varepsilon(k+1)} + \varepsilon \right] \]

3. **(predictions):** for every solution $\beta_{LS}^i$ it holds that
   \[ \|X\beta^i - X\beta_{LS}\|_2 \leq \frac{\sqrt{p}}{\lambda_{p_{\text{min}}}(X^TX)} \left[ \frac{\|X\beta_{LS}\_i\|^2}{\varepsilon(k+1)} + \varepsilon \right] \]

4. **(correlation values):** $\|X^Tr^i\|_\infty \leq \left[ \frac{\|X\beta_{LS}\_i\|^2}{\varepsilon(k+1)} + \varepsilon \right]$

5. **(sparsity and shrinkage):** $\|\beta^i\|_0 \leq k$ and $\|\beta^i\|_1 \leq k\varepsilon$

\[ \lambda_{p_{\text{min}}}(X^TX) > 0 \text{ is the smallest positive eigenvalue of } X^TX \]
FS\(\varepsilon\) Has Explicit Computational Guarantees

**Theorem: Computational Guarantees for FS\(\varepsilon\)**

With the constant shrinkage factor \(\varepsilon\), after \(k\) iterations there exists \(i \leq k\) for which:

(i) (training error): \(L_n(\beta^i) - L^*_n \leq \frac{p}{2n(\lambda_{p_{\min}}(X^T X))} \left[ \|X\beta_{LS}\|_2^2 + \varepsilon \right]^2\)

(ii) (regression coefficients): there exists a solution \(\beta^i_{LS}\) for which
\[
\|\beta^i - \beta^i_{LS}\|_2 \leq \frac{\sqrt{p}}{(\lambda_{p_{\min}}(X^T X))} \left[ \|X\beta_{LS}\|_2^2 + \varepsilon \right]
\]

(iii) (predictions): for every solution \(\beta_{LS}\) it holds that
\[
\|X\beta^i - X\beta_{LS}\|_2 \leq \frac{\sqrt{p}}{\sqrt{\lambda_{p_{\min}}(X^T X)}} \left[ \|X\beta_{LS}\|_2^2 + \varepsilon \right]
\]

(iv) (correlation values): \(\|X^T r^i\|_\infty \leq \left[ \|X\beta_{LS}\|_2^2 + \varepsilon \right]\)

(v) (sparsity and shrinkage): \(\|\beta^i\|_0 \leq k\) and \(\|\beta^i\|_1 \leq S\text{BOUND}\)

Define \(S\text{BOUND} := \varepsilon(k + 1)\)
\[
L_n(\beta^i) - L^*_n \leq \frac{p}{2n(\lambda_{\text{pmin}}(X^T X))} \left[ \|X \beta_{\text{LS}}\|_2^2 + \varepsilon \right]^2
\]

\ell_1 \text{ Shrinkage versus Data-Fidelity Tradeoffs for } F_{\varepsilon}

![Graph showing \ell_1 \text{ shrinkage of coefficients} against training error for different \varepsilon values](image-url)
Interpreting $\lambda_{\text{pmin}}(X^TX)$

Generate $X$ with entries drawn from a standard Gaussian ensemble

It follows from random matrix theory that

$$\lambda_{\text{pmin}}(X^TX) \gtrapprox \frac{1}{n} (\sqrt{p} - \sqrt{n})^2$$

with high probability [Vershynin 2010]
Suppose we generate $\mathbf{X}$ such that covariance $\sigma_{i,j}$ of columns $i$ and $j$ satisfies $\sigma_{i,j} = \rho$ for $i \neq j$, and then normalize the columns. We observe:

![Graph showing the relationship between $\lambda_{p_{\min}}(\mathbf{X}^T \mathbf{X})$ and $\rho$.]
Where do the Computational Guarantees Come From?

Q: Where did the previous computational guarantees come from?

A: Re-interpretation of $FS_\varepsilon$ as an optimization algorithm (of a particular objective function on a particular domain).

- What is the objective function being optimized?
- What is the domain of feasible solutions?
- What is the algorithm?
FS$_\varepsilon$ is Subgradient Descent

Least-squares optimality conditions:

$$\mathbf{X}^T r_{LS} = 0 \quad \text{where} \quad r_{LS} = \mathbf{y} - \mathbf{X}\beta_{LS}$$

**FS$_\varepsilon$ Equivalence Theorem**

The FS$_\varepsilon$ algorithm is an instance of the Subgradient Descent method to solve the following non-smooth convex optimization problem:

$$\min_{r \in P_{\text{res}}} f(r) := \|\mathbf{X}^T r\|_\infty$$

where $P_{\text{res}} := \{r : r = \mathbf{y} - \mathbf{X}\beta \text{ for some } \beta \in \mathbb{R}^p \}$, initialized at $r^0 = \mathbf{y}$ and with a constant step-size $\alpha_i := \varepsilon$ at each iteration.
FS_ε Algorithm

Initialize at \( r^0 = y, \beta^0 = 0, k = 0, \text{ set } \varepsilon > 0 \)

At iteration \( k \geq 0 \):

- Compute:
  \[
  j_k \in \arg \max_{j \in \{1, \ldots, p\}} |(r^k)^T x_j |
  \]
  \[
  g^k \leftarrow \text{sgn}((r^k)^T x_{j_k}) x_{j_k}
  \]

- Set:
  \[
  \beta^{k+1} \leftarrow \beta^k + \varepsilon \cdot \text{sgn}((r^k)^T x_{j_k}) e_{j_k}
  \]
  \[
  r^{k+1} \leftarrow r^k - \varepsilon \cdot g^k
  \]
Incremental Forward Stagewise Regression Algorithm (FS$_\varepsilon$)

**FS$_\varepsilon$ Algorithm**

Initialize at $r^0 = y$, $\beta^0 = 0$, $k = 0$, set $\varepsilon > 0$

At iteration $k \geq 0$:

- Compute:

  $j_k \in \arg \max_{j \in \{1, \ldots, p\}} \left| (r^k)^T X_j \right|$

  $g^k \leftarrow \text{sgn}((r^k)^T X_{j_k}) X_{j_k} \in \partial \| X^T r^k \|_\infty$

- Set:

  $\beta^{k+1} \leftarrow \beta^k + \varepsilon \cdot \text{sgn}((r^k)^T X_{j_k}) e_{j_k}$

  $r^{k+1} \leftarrow r^k - \varepsilon \cdot g^k$
**FS$_\varepsilon$ Algorithm**

Initialize at $r^0 = y$, $\beta^0 = 0$, $k = 0$, set $\varepsilon > 0$

At iteration $k \geq 0$:

- Compute:
  
  $j_k \in \arg \max_{j \in \{1, \ldots, p\}} |(r^k)^T X_j|$

  $g^k \leftarrow \text{sgn}((r^k)^T X_{j_k}) X_{j_k} \in \partial \|X^T r^k\|_\infty$

- Set:

  $\beta^{k+1} \leftarrow \beta^k + \varepsilon \cdot \text{sgn}((r^k)^T X_{j_k}) e_{j_k}$

  $r^{k+1} \leftarrow r^k - \varepsilon \cdot g^k$
LS-Boost(\(\varepsilon\)) Boosting Algorithm
The **LS-Boost(ε)** Boosting Algorithm

Least-squares optimality conditions:

\[ \mathbf{X}^T r_{LS} = 0 \quad \text{where} \quad r_{LS} = \mathbf{y} - \mathbf{X}\beta_{LS} \]

**LS-Boost(ε)** is another simple and intuitive boosting algorithm:

Start with \( \beta^0 \leftarrow 0 \), and hence \( r^0 \leftarrow \mathbf{y} \). Fix \( \varepsilon > 0 \).

Given \( \beta^k \) and \( r^k := \mathbf{y} - \mathbf{X}\beta^k \), determine the weak model \( \mathbf{X}_j \) most correlated with the current residuals \( r^k \):

\[ j_k \leftarrow \arg \max_{j \in \{1, \ldots, p\}} |(r^k)^T \mathbf{X}_j| \]

(Only) adjust \( \beta^k_{j_k} \) by \( \varepsilon \cdot ((r^k)^T \mathbf{X}_{j_k}) \)
**LS-Boost(ε)** Boosting Method

**LS-Boost(ε) Algorithm**

Initialize at $r^0 = y$, $β^0 = 0$, $k = 0$, set $ε > 0$

At iteration $k ≥ 0$:

- Compute:
  
  \[ j_k ∈ \arg \max_{j∈\{1,...,p\}} |(r^k)^T X_j| \]

- Set:
  
  \[ β^{k+1} ← β^k + ε \cdot ((r^k)^T X_{j_k}) e_{j_k} \]
  
  \[ r^{k+1} ← r^k - ε \cdot ((r^k)^T X_{j_k}) X_{j_k} \]

The parameter $ε$ is again called the **learning rate** or **shrinkage factor**
Sparsity Properties of $\text{LS-Boost}(\varepsilon)$

Notice that $\text{LS-Boost}(\varepsilon)$ adds at most one new predictor at every iteration, thus:

$$\|\beta^k\|_0 \leq k.$$ 

Therefore $\text{LS-Boost}(\varepsilon)$ controls sparsity.

How does $\text{LS-Boost}(\varepsilon)$ control bias?
Theorem: Convergence Properties of LS-Boost(ε)

With the shrinkage factor $ε \in (0, 1]$, define the linear convergence rate coefficient $γ$:

$$γ := \left(1 - \frac{ε(2 - ε)λ_{pmin}(X^TX)}{4p}\right).$$

For all $k \geq 0$ the following bounds hold:

(i) (training error): $L_n(β^k) - L^* \leq ?$

(ii) (regression coefficients): there exists a solution $β_{LS}^k$ for which $\|β^k - β_{LS}^k\|_2 \leq ?$

(iii) (predictions): for every least-squares solution solution $β_{LS}$ it holds that $\|Xβ^k - Xβ_{LS}\|_2 \leq ?$

(iv) (gradient/correlation values):

$$\|∇L_n(β^k)\|_∞ = \frac{1}{n}\|X^Tr^k\|_∞ \leq ?$$

(v) (sparsity and shrinkage): $\|β^k\|_0 \leq k$ and $\|β^k\|_1 \leq ?$
Theorem: Convergence Properties of LS-Boost(\(\varepsilon\))

With the shrinkage factor \(\varepsilon \in (0, 1]\), define the linear convergence rate coefficient \(\gamma\):

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

For all \(k \geq 0\) the following bounds hold:

(i) (training error): \(L_n(\beta^k) - L^*_n \leq \frac{1}{2n} \|X\beta_{LS}\|_2^2 \cdot \gamma^k\)

(ii) (regression coefficients): there exists a solution \(\beta_{LS}^k\) for which \(\|\beta^k - \beta_{LS}^k\|_2 \leq \) ?

(iii) (predictions): for every least-squares solution solution \(\beta_{LS}\) it holds that \(\|X\beta^k - X\beta_{LS}\|_2 \leq \) ?

(iv) (gradient/correlation values):
\[
\|\nabla L_n(\beta^k)\|_{\infty} = \frac{1}{n} \|X^T r^k\|_{\infty} \leq \) ?

(v) (sparsity and shrinkage): \(\|\beta^k\|_0 \leq k\) and \(\|\beta^k\|_1 \leq \) ?
Convergence Properties of $\text{LS-BOOST}(\varepsilon)$

**Theorem: Convergence Properties of $\text{LS-BOOST}(\varepsilon)$**

With the shrinkage factor $\varepsilon \in (0, 1]$, define the linear convergence rate coefficient $\gamma$:

$$
\gamma := \left( 1 - \frac{\varepsilon(2 - \varepsilon)\lambda_{p\min}(X^TX)}{4p} \right).
$$

For all $k \geq 0$ the following bounds hold:

(i) (training error): $L_n(\beta^k) - L^*_n \leq \frac{1}{2n} \|X\beta_{LS}\|_2^2 \cdot \gamma^k$

(ii) (regression coefficients): there exists a solution $\beta^k_{LS}$ for which

$$
\|\beta^k - \beta^k_{LS}\|_2 \leq \frac{\|X\beta_{LS}\|_2}{\sqrt{\lambda_{p\min}(X^TX)}} \cdot \gamma^{k/2}
$$

(iii) (predictions): for every least-squares solution solution $\beta_{LS}$ it holds that

$$
\|X\beta^k - X\beta_{LS}\|_2 \leq ?
$$

(iv) (gradient/correlation values):

$$
\|\nabla L_n(\beta^k)\|_\infty = \frac{1}{n} \|X^Tr^k\|_\infty \leq ?
$$

(v) (sparsity and shrinkage): $\|\beta^k\|_0 \leq k$ and $\|\beta^k\|_1 \leq ?$
Convergence Properties of $\text{LS-BOOST}(\varepsilon)$

**Theorem: Convergence Properties of $\text{LS-BOOST}(\varepsilon)$**

With the shrinkage factor $\varepsilon \in (0, 1]$, define the linear convergence rate coefficient $\gamma$:

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

For all $k \geq 0$ the following bounds hold:

1. **(training error):** $L_n(\beta_k) - L_n^* \leq \frac{1}{2n} \|X\beta_{LS}\|_2^2 \cdot \gamma^k$
2. **(regression coefficients):** there exists a solution $\beta_{LS}^k$ for which $\|\beta_k^* - \beta_{LS}^k\|_2 \leq \frac{\|X\beta_{LS}\|_2}{\sqrt{\lambda_{\text{pmin}}(X^TX)}} \cdot \gamma^{k/2}$
3. **(predictions):** for every least-squares solution $\beta_{LS}$ it holds that $\|X\beta_k - X\beta_{LS}\|_2 \leq \|X\beta_{LS}\|_2 \cdot \gamma^{k/2}$
4. **(gradient/correlation values):** $\|\nabla L_n(\beta_k)\|_\infty = \frac{1}{n} \|X^Tr_k\|_\infty \leq \ ?$
5. **(sparsity and shrinkage):** $\|\beta_k\|_0 \leq k$ and $\|\beta_k\|_1 \leq \ ?$
Theorem: Convergence Properties of LS-Boost(ε)

With the shrinkage factor $\varepsilon \in (0, 1]$, define the linear convergence rate coefficient $\gamma$:

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

For all $k \geq 0$ the following bounds hold:

(i) (training error): $L_n(\beta^k) - L_n^* \leq \frac{1}{2n} \|X\beta_{\text{LS}}\|^2 \cdot \gamma^k$

(ii) (regression coefficients): there exists a solution $\beta^k_{\text{LS}}$ for which $\|\beta^k - \beta^k_{\text{LS}}\|_2 \leq \frac{\|X\beta_{\text{LS}}\|_2}{\sqrt{\lambda_{\text{pmin}}(X^TX)}} \cdot \gamma^{k/2}$

(iii) (predictions): for every least-squares solution $\beta_{\text{LS}}$ it holds that $\|X\beta^k - X\beta_{\text{LS}}\|_2 \leq \|X\beta_{\text{LS}}\|_2 \cdot \gamma^{k/2}$

(iv) (gradient/correlation values): 
$$
\|\nabla L_n(\beta^k)\|_\infty = \frac{1}{n} \|X^Tr^k\|_\infty \leq \frac{1}{n} \|X\beta_{\text{LS}}\|_2 \cdot \gamma^{k/2}
$$

(v) (sparsity and shrinkage): $\|\beta^k\|_0 \leq k$ and $\|\beta^k\|_1 \leq \; ?$
Theorem: Convergence Properties of LS-BOOST(\(\varepsilon\))

With the shrinkage factor \(\varepsilon \in (0, 1]\), define the linear convergence rate coefficient \(\gamma\):

\[
\gamma := \left(1 - \frac{\varepsilon(2 - \varepsilon)\lambda_{\text{pmin}}(X^TX)}{4p}\right).
\]

For all \(k \geq 0\) the following bounds hold:

(i) (training error): \(L_n(\beta^k) - L^*_n \leq \frac{1}{2n} \|X\beta_{\text{LS}}\|_2^2 \cdot \gamma^k\)

(ii) (regression coefficients): there exists a solution \(\beta^k_{\text{LS}}\) for which

\[
\|\beta^k - \beta^k_{\text{LS}}\|_2 \leq \frac{\|X\beta_{\text{LS}}\|_2}{\sqrt{\lambda_{\text{pmin}}(X^TX)}} \cdot \gamma^{k/2}
\]

(iii) (predictions): for every least-squares solution solution \(\beta_{\text{LS}}\) it holds that

\[
\|X\beta^k - X\beta_{\text{LS}}\|_2 \leq \|X\beta_{\text{LS}}\|_2 \cdot \gamma^{k/2}
\]

(iv) (gradient/correlation values):

\[
\|\nabla L_n(\beta^k)\|_{\infty} = \frac{1}{n} \|X^T r^k\|_{\infty} \leq \frac{1}{n} \|X\beta_{\text{LS}}\|_2 \cdot \gamma^{k/2}
\]

(v) (sparsity and shrinkage): \(\|\beta^k\|_0 \leq k\) and

\[
\|\beta^k\|_1 \leq \min \left\{ \sqrt{k} \sqrt{\frac{\varepsilon}{2 - \varepsilon}} \sqrt{\|X\beta_{\text{LS}}\|_2^2 - \|X\beta_{\text{LS}} - X\beta^k\|_2^2} , \frac{\varepsilon\|X\beta_{\text{LS}}\|_2}{1 - \sqrt{\gamma}} \left(1 - \gamma^{k/2}\right) \right\}
\]
Theorem: Convergence Properties of LS-Boost(ε)

With the shrinkage factor $\varepsilon \in (0, 1]$, define the linear convergence rate coefficient $\gamma$:

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

For all $k \geq 0$ the following bounds hold:

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

(ii) (regression coefficients): there exists a solution $\beta_{LS}^k$ for which

$$
\|\beta^k - \beta_{LS}^k\|^2_2 \leq \frac{\|X \beta_{LS}\|^2_2}{\sqrt{\lambda_{\text{pmin}}(X^T X)}} \cdot \gamma^{k/2}
$$

(iii) (predictions): for every least-squares solution solution $\beta_{LS}$ it holds that

$$
\|X \beta^k - X \beta_{LS}\|^2_2 \leq \|X \beta_{LS}\|^2_2 \cdot \gamma^{k/2}
$$

(iv) (gradient/correlation values):

$$
\|\nabla L_n(\beta^k)\|_\infty = \frac{1}{n} \|X^T r^k\|_\infty \leq \frac{1}{n} \|X \beta_{LS}\|^2_2 \cdot \gamma^{k/2}
$$

(v) (sparsity and shrinkage): $\|\beta^k\|_0 \leq k$ and $\|\beta^k\|_1 \leq \min \left\{ \sqrt{k} \sqrt{\frac{\varepsilon}{2 - \varepsilon}} \sqrt{\|X \beta_{LS}\|^2_2 - \|X \beta_{LS} - X \beta^k\|^2_2}, \frac{\varepsilon \|X \beta_{LS}\|^2_2}{1 - \sqrt{\gamma}} \left(1 - \gamma^{k/2}\right) \right\}$
**LS-Boost(ε) is Steepest Descent in ℓ₁-norm**

**LS-Boost(ε) is doing Steepest Descent in the ℓ₁-norm to minimize the least-squares loss function**

**LS-Boost(ε) Equivalence Theorem**

The LS-Boost(ε) algorithm is an instance of the Steepest Descent method in the ℓ₁-norm to minimize the least-squares loss function:

$$\min_{\beta \in \mathbb{R}^p} L_n(\beta) := \frac{1}{2n} \| y - X\beta \|^2_2$$

initialized at $\beta^0 = 0$, with step-sizes determined by exact line-search at each iteration.
Regularized $F_{\epsilon}$ and Connections to the LASSO
Recall $\text{FS}_\varepsilon$:

**FS$_\varepsilon$ Algorithm**

Initialize at $r^0 = y$, $\beta^0 = 0$, $k = 0$, set $\varepsilon > 0$

At iteration $k \geq 0$:

- Compute:
  
  $$j_k \in \arg \max_{j \in \{1, \ldots, p\}} |(r^k)^T X_j|$$

  $$g^k \leftarrow \text{sgn}((r^k)^T X_{j_k}) X_{j_k}$$

- Set:
  
  $$\beta^{k+1} \leftarrow \beta^k + \varepsilon \cdot \text{sgn}((r^k)^T X_{j_k}) e_{j_k}$$

  $$r^{k+1} \leftarrow r^k - \varepsilon \cdot g^k$$
The LASSO

The LASSO (in constraint-mode)

\[
\text{LASSO}_\delta^* := \min_{\beta} \quad L_n(\beta) := \frac{1}{2n} \| y - X\beta \|_2^2
\]

\[\text{s.t.} \quad \|\beta\|_1 \leq \delta\]

Compared to boosting (FS_\varepsilon or LS-\textsc{Boost}(\varepsilon)), the dynamic of the data-fidelity/regularization tradeoff in the LASSO is perhaps more explicit.

There are various resemblances between boosting methods and the LASSO . . .
Boosting/LASSO Coefficient Profiles

FS$\_\varepsilon$ Coefficient Profile

Coefficient Values

Iteration
Boosting/LASSO Coefficient Profiles

**FS\_ε Coefficient Profile**

- Coefficient Values vs. Iteration (scaled)

**LASSO Coefficient Profile**

- Coefficient Values vs. Regularization parameter \( \delta \)
Boosting/LASSO Coefficient Profiles

**FS\(\varepsilon\) Coefficient Profile**

**LASSO Coefficient Profile**
Boosting and the LASSO

Why does the LASSO perform well?

- Extensive work studying the statistical properties of the LASSO

Why does boosting perform well (lead to sparse models with good predictive performance)?

- Less work (up until now) for boosting methods

We seek to:

- Improve the understanding of why boosting performs well, and
- Bridge the gap between boosting and the LASSO
Boosting and the LASSO have **structural similarities**

- (that are partially understood)

Can both methodologies be understood as part of a single framework?
- we will see that the answer is **yes**
The LASSO, again

The LASSO (in constraint-mode)

\[ LASSO^*_{\delta} := \min_{\beta} \quad L_n(\beta) := \frac{1}{2n} \| y - X\beta \|_2^2 \]

\[ \text{s.t.} \quad \| \beta \|_1 \leq \delta \]
Regularized Forward Stagewise Regression (R-FS$_{\varepsilon,\delta}$)

Subgradient descent problem for FS$_{\varepsilon}$:

$$\min_{r \in P_{\text{res}}} \ f(r) := \|X^Tr\|_\infty$$

**FS$_{\varepsilon}$ Algorithm**

Initialize at $r^0 = y$, $\beta^0 = 0$, $k = 0$, set $\varepsilon > 0$,

At iteration $k \geq 0$:

- Compute:

  $$j_k \in \arg\max_{j \in \{1, \ldots, p\}} |(r^k)^TX_j|$$

  $$g^k \leftarrow \text{sgn}(r^k)^TX_{j_k})X_{j_k}$$

- Set:

  $$r^{k+1} \leftarrow r^k - \varepsilon g^k$$

  $$\beta^{k+1} \leftarrow \beta^k + \varepsilon \text{sgn}((r^k)^TX_{j_k})e_{j_k}$$
Subgradient descent problem for \( \text{R-FS}_{\varepsilon, \delta} \):

\[
\min_{r \in P_{\text{res}}} f(r) := \|X^T r\|_\infty + \frac{1}{2\delta} \|r - y\|_2^2
\]

**R-FS\(_{\varepsilon, \delta}\) (Regularized FS\(_\varepsilon\)) Algorithm**

Initialize at \( r^0 = y, \beta^0 = 0, k = 0 \), set \( \varepsilon > 0 \), set \( \delta \in (0, \infty] \)

At iteration \( k \geq 0 \):

- Compute:

  \[
  j_k \in \arg \max_{j \in \{1, \ldots, p\}} |(r^k)^T X_j|
  \]

  \[
  g^k \leftarrow \text{sgn}(r^k)^T X_{j_k})X_{j_k} + \frac{1}{\delta}(r^k - y)
  \]

- Set:

  \[
  r^{k+1} \leftarrow r^k - \varepsilon g^k
  \]

  \[
  \beta^{k+1} \leftarrow (1 - \frac{\varepsilon}{\delta}) \beta^k + \varepsilon \text{sgn}((r^k)^T X_{j_k})e_{j_k}
  \]
Two Rationales/Interpretations for $\text{R-FS}_{\varepsilon,\delta}$

1. $\text{R-FS}_{\varepsilon,\delta}$ is Subgradient Descent applied to the “regularized minimum correlation problem”:

$$\min_{r \in P_{\text{res}}} f(r) := \|X^T r\|_\infty + \frac{1}{2\delta} \|r - y\|_2^2$$

2. $\text{R-FS}_{\varepsilon,\delta}$ is tackling the LASSO problem (through duality):

$$\text{LASSO}^*_\delta := \min_{\beta} L_n(\beta) := \frac{1}{2n} \|y - X\beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_1 \leq \delta$$

We will show computational guarantees for $\text{R-FS}_{\varepsilon,\delta}$ that are analogous to $\text{FS}_\varepsilon$

- Except that now the boosting profile “terminates” at a LASSO solution
Regularized Forward Stagewise Regression (R-FS$\varepsilon$)

R-FS$\varepsilon$ is a trivial re-scaling of FS$\varepsilon$:

**R-FS$\varepsilon$ Algorithm**

Initialize at $r^0 = y$, $\beta^0 = 0$, $k = 0$, set $\varepsilon > 0$, set $\delta \in (0, \infty]$

At iteration $k \geq 0$:

- Compute:
  $$j_k \in \text{arg max}_{j \in \{1, \ldots, p\}} |(r^k)^T X_j|$$

- Set:
  $$\beta^{k+1} \leftarrow (1 - \frac{\varepsilon}{\delta}) \beta^k + \varepsilon \cdot \text{sgn}((r^k)^T X_j^* e_{j_k})$$
  $$r^{k+1} \leftarrow r^k - \varepsilon \cdot \text{sgn}((r^k)^T X_j^*) X_{j_k} + \frac{1}{\delta} (r^k - y)$$
Motivation for $R$-$FS_\varepsilon$: the LASSO

The LASSO (in constraint-mode)

$$\text{LASSO}_\delta := \min_{\beta} \quad L_n(\beta) := \frac{1}{2n} \|y - X\beta\|_2^2$$

s.t. \quad \|\beta\|_1 \leq \delta

We show that $R$-$FS_\varepsilon$ optimizes the $\text{LASSO}_\delta$ problem ...
Convergence Properties of R-FS$_{\varepsilon}$

Theorem: Convergence Properties of R-FS$_{\varepsilon}$

With the shrinkage factor $\varepsilon \leq \delta$, after $k$ iterations there exists $i \leq k$ for which:

(i) (training error): $L_n(\beta^i) - \text{LASSO}_\delta^* \leq ?$

(ii) (predictions): for every LASSO solution $\beta_{LS}^*$ it holds that $\|X\beta^i - X\beta_{LS}^*\|_2 \leq ?$

(iii) (shrinkage): $\|\beta^i\|_1 \leq ?$

(iv) (sparsity): $\|\beta^i\|_0 \leq k$
Convergence Properties of $R\text{-FS}_\varepsilon$

**Theorem: Convergence Properties of $R\text{-FS}_\varepsilon$**

With the shrinkage factor $\varepsilon \leq \delta$, after $k$ iterations there exists $i \leq k$ for which:

(i) (training error): $L_n(\beta^i) - \text{LASSO}_\delta^* \leq \frac{\delta}{n} \left[ \frac{\|X\beta_{LS}\|_2^2}{2\varepsilon(k+1)} + 2\varepsilon \right]$

(ii) (predictions): for every LASSO solution $\beta_{LS}^*$ it holds that $\|X\beta^i - X\beta_{LS}^*\|_2 \leq ?$

(iii) (shrinkage): $\|\beta^i\|_1 \leq ?$

(iv) (sparsity): $\|\beta^i\|_0 \leq k$
Theorem: Convergence Properties of R-FS\(_{\varepsilon}\)

With the shrinkage factor \(\varepsilon \leq \delta\), after \(k\) iterations there exists \(i \leq k\) for which:

(i) (training error): \(L_n(\beta^i) - \text{LASSO}^*_\delta \leq \frac{\delta}{n} \left[ \frac{\|X\beta_{LS}\|_2^2}{2\varepsilon(k+1)} + 2\varepsilon \right]\)

(ii) (predictions): for every LASSO solution \(\beta^*_{LS}\) it holds that
\[
\|X\beta^i - X\beta_{LS}\|_2 \leq \sqrt{\frac{\delta\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)}} + 4\delta\varepsilon
\]

(iii) (shrinkage): \(\|\beta^i\|_1 \leq \frac{\delta}{\varepsilon(k+1)} + 2\varepsilon\)

(iv) (sparsity): \(\|\beta^i\|_0 \leq k\)
Convergence Properties of R-FS$_{\varepsilon}$

**Theorem: Convergence Properties of R-FS$_{\varepsilon}$**

With the shrinkage factor $\varepsilon \leq \delta$, after $k$ iterations there exists $i \leq k$ for which:

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

(ii) (predictions): for every LASSO solution $\beta^*_{LS}$ it holds that
$$\|X\beta^i - X\beta^*_{LS}\|_2 \leq \sqrt{\frac{\delta\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)}} + 4\delta\varepsilon$$

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

(iv) (sparsity): $\|\beta^i\|_0 \leq k$
**Theorem: Convergence Properties of R-FS\(_{\varepsilon}\)**

With the shrinkage factor \(\varepsilon \leq \delta\), after \(k\) iterations there exists \(i \leq k\) for which:

(i) (training error): \(L_n(\beta^i) - \text{LASSO}^*_\delta \leq \frac{\delta}{n} \left[ \frac{\|X\beta_{LS}\|^2}{2\varepsilon(k+1)} + 2\varepsilon \right]\)

(ii) (predictions): for every LASSO solution \(\beta_{LS}^*\) it holds that
\[
\|X\beta^i - X\beta_{LS}\|_2 \leq \sqrt{\frac{\delta\|X\beta_{LS}\|^2}{\varepsilon(k+1)}} + 4\delta\varepsilon
\]

(iii) (shrinkage): \(\|\beta^i\|_1 \leq \delta \left[ 1 - (1 - \frac{\varepsilon}{\delta})^k \right] \leq \delta\)

(iv) (sparsity): \(\|\beta^i\|_0 \leq k\)
Convergence Properties of R-FS$_{\varepsilon}$

**Theorem: Convergence Properties of R-FS$_{\varepsilon}$**

With the shrinkage factor $\varepsilon \leq \delta$, after $k$ iterations there exists $i \leq k$ for which:

1. *(training error):* $L_n(\beta^i) - \text{LASSO}_\delta^* \leq \frac{\delta}{n} \left[ \frac{\|X\beta_{LS}\|_2^2}{2\varepsilon(k+1)} + 2\varepsilon \right]$

2. *(predictions):* for every LASSO solution $\beta_{LS}^*$ it holds that
   \[
   \|X\beta^i - X\beta_{LS}\|_2 \leq \sqrt{\frac{\delta\|X\beta_{LS}\|_2^2}{\varepsilon(k+1)}} + 4\delta \varepsilon
   \]

3. *(shrinkage):* $\|\beta^i\|_1 \leq \delta \left[ 1 - \left(1 - \frac{\varepsilon}{\delta}\right)^k \right] \leq \delta$

4. *(sparsity):* $\|\beta^i\|_0 \leq k$

For finite $\delta$, R-FS$_{\varepsilon}$ provides computational guarantees for the LASSO

Hence a trivial modification of FS$_{\varepsilon}$ produces LASSO solutions

This connects FS$_{\varepsilon}$ directly to the LASSO
Some Preliminary Computation

Leukemia dataset: publicly available microarray classification dataset

- $p = 500$, $n = 72$, $\varepsilon = 10^{-3}$ (fixed)
- Continuous responses were generated artificially
- The true model has 10 non-zero coefficients

<table>
<thead>
<tr>
<th>Method</th>
<th>hyper-parameters</th>
<th>SNR</th>
<th>Test Error</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{FS}_\varepsilon$</td>
<td>$k$</td>
<td>1</td>
<td>0.3431 (0.0087)</td>
<td>28</td>
</tr>
<tr>
<td>R-$\text{FS}_{\varepsilon,\delta}$</td>
<td>$(\delta, k)$</td>
<td>1</td>
<td>0.3411 (0.0086)</td>
<td>25</td>
</tr>
<tr>
<td>LASSO</td>
<td>$\delta$</td>
<td>1</td>
<td>0.3460 (0.0086)</td>
<td>30</td>
</tr>
<tr>
<td>$\text{FS}_\varepsilon$</td>
<td>$k$</td>
<td>10</td>
<td>0.0681 (0.0014)</td>
<td>67</td>
</tr>
<tr>
<td>R-$\text{FS}_{\varepsilon,\delta}$</td>
<td>$(\delta, k)$</td>
<td>10</td>
<td>0.0659 (0.0014)</td>
<td>60</td>
</tr>
<tr>
<td>LASSO</td>
<td>$\delta$</td>
<td>10</td>
<td>0.0677 (0.0015)</td>
<td>61</td>
</tr>
</tbody>
</table>

In this instance, R-$\text{FS}_{\varepsilon,\delta}$ tends to deliver sparser models with better test error
Summary

- $\text{FS}_\varepsilon$ is doing Subgradient Descent to minimize the maximum correlation between the residuals and the predictors
  - computational guarantees for $\text{FS}_\varepsilon$ for least-squares regression

- $\text{LS-BOOST}(\varepsilon)$ is doing $\ell_1$-norm Steepest Descent method to minimize least-squares loss function
  - computational guarantees for $\text{LS-BOOST}(\varepsilon)$ for least-squares regression

- $\text{R-FS}_\varepsilon$ is a trivial modification of $\text{FS}_\varepsilon$ that yields direct computational guarantees for the LASSO