# Understanding Randomization in Greedy Model Search

**Xin Chen**
Department of Operations Research
and Financial Engineering
Princeton University
xc5557@princeton.edu

**Jason M. Klusowski**
Department of Operations Research
and Financial Engineering
Princeton University
jason.klusowski@princeton.edu

**Yan Shuo Tan**
Department of Statistics
and Data Science
National University of Singapore
yanshuo@nus.edu.sg

**Chang Yu**
Department of Operations Research
and Financial Engineering
Princeton University
cy7438@princeton.edu

## Abstract

We study feature subsampling in greedy model search, a randomization mechanism motivated by random forests and other ensemble methods. Although recent theory draws an analogy between this randomization and ridge regularization, these results largely rely on base learners optimized via ordinary least squares (OLS). We therefore investigate the effects of feature subsampling on greedy forward selection, a tractable abstraction of the adaptive split search used by decision trees. Assuming an orthogonal design, we prove that, under suitable conditions, ensembling with a fixed feature subsampling size can reduce expected training error without increasing degrees of freedom, thereby reducing fixed design prediction error. We also show that both the training error and degrees of freedom need not be monotone in the subsampling rate, breaking the analogy with standard shrinkage methods like the lasso or ridge regression under orthogonality. Furthermore, we characterize the exact asymptotic behavior of the estimator, showing that it adaptively reweights OLS coefficients based on their rank, with weights that are bounded by explicit logistic functions.

## 1 Introduction

Introduced by Breiman [2001], random forests (RFs) are ensembles of decision trees, each grown via greedy, recursive partitioning of the input space [Breiman et al., 1984]. Each tree is independently randomized in two ways. First, it is trained on a bootstrap sample of the data. Second, at every internal node it considers only a randomly drawn subset of $m$ (out of $p$ total) features when searching for the next split.

There has been significant theoretical work on the consistency of RFs, especially in high dimensions [Scornet et al., 2015, Klusowski, 2020, Syrgkanis and Zampetakis, 2020, Chi et al., 2022, Klusowski and Tian, 2024, Mazumder and Wang, 2023], but there is not yet a complete consensus on *how randomization improves prediction performance*. Breiman [2001] originally explained the benefits of randomization as increasing diversity among the trees, thereby reducing the variance in their predictions. Interestingly, in his concluding remarks, he speculated that randomization may "act to reduce bias" but that "the mechanism for this is not obvious". These two explanations have had

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

divergent fates, with Hastie et al. [2009] arguing in their influential textbook that "the improvements in prediction obtained by bagging or random forests are solely a result of variance reduction".

Seminal work by Mentch and Zhou [2020] offered a quantitative analysis of the role played by the feature subsampling ratio $\gamma = m/p$. In their numerical experiments, they found that decreasing $\gamma$ decreases the degrees of freedom in RFs. To further explore the generality of this phenomenon, they applied ensembling with feature subsampling to the greedy forward selection (FS) algorithm [Draper and Smith, 1966], finding that decreasing $\gamma$ also decreases the degrees of freedom of the resulting family of estimators. Finally, in work concurrent with that of LeJeune et al. [2020], they analyzed ensembling with feature subsampling on ordinary least squares (OLS) estimators. Such an ensemble is linear in the response, whereas ensembles of FS generally are not because the selected features depend on the response. Assuming orthonormal features, they found that the resulting regression coefficients were shrunk from the OLS values by a factor of $\gamma$, making the algorithm equivalent to ridge regression. They hence concluded that feature subsampling "serves as a form of implicit regularization" and that $\gamma$ "serves much the same purpose as the shrinkage penalty in explicitly regularized regression procedures like lasso and ridge regression". Under orthogonality and without bootstrap resampling, they also noted that averaging scales each OLS coefficient by its selection frequency.

Recent work building on Mentch and Zhou [2020] has proceeded in two directions. In one direction, Patil and LeJeune [2024] relaxed the orthonormal feature assumption to one they called "asymptotic freeness" (relative to the feature subsampling operator). Under suitable asymptotic freeness assumptions, they obtained the risk of an equivalent ridge estimator plus a sketching variance term proportional to inverse ensemble size, which vanishes for an infinite ensemble. Meanwhile, other works studying rowwise sampling ("data subsampling") established a broad theoretical equivalence between this form of randomized ensembling and ridge penalization [Du et al., 2023, Patil and Du, 2023]. Together, these works seem to strengthen the analogy between various forms of subsampling randomization and ridge style shrinkage.

In the other direction, studies by Curth et al. [2024], Liu and Mazumder [2025], Mei et al. [2026], and Revelas et al. [2025] suggest that feature subsampling can also reduce bias in RFs via two mechanisms, although their notions of bias differ. The first is enlarging the effective model class [Curth et al., 2024]. The second is overcoming failures of greedy optimization [Liu and Mazumder, 2025]. These results challenge the "pure variance reduction" hypothesis of Hastie et al. [2009] and suggest that *although the shrinkage analogy in Mentch and Zhou [2020] seems appropriate when describing ensembles of base learners fit with convex optimization, this analogy becomes strained when subsampling is interleaved with greedy optimization, as is the case for RFs*. Existing theory on subsampling based ensembling primarily concerns base learners that are not adaptive or that are analytically tractable because they arise from convex optimization, as well as oracle versions of RFs [Mei et al., 2026]. For settings in which feature subsampling is interleaved with greedy, adaptive fitting, evidence for bias reduction mechanisms, however, remains largely empirical [Curth et al., 2024, Liu and Mazumder, 2025, Revelas et al., 2025].

The goal of our paper is therefore to isolate the interaction between feature subsampling and greedy, response dependent selection. Inspired by Mentch and Zhou [2020], we study their ensembling strategy as applied to FS rather than decision trees as in RFs. The connection between FS and regression trees is closer than one might first expect. A regression tree can be written as a linear combination of orthogonal stump functions [Cattaneo et al., 2024, Lemma 2.1] and splitting a terminal node selects the available stump most correlated with the residual. This is precisely the same logic as FS, except that the dictionary of stumps is not fixed but depends on the current partition of the input space.

We assume a fixed design setting and orthogonal features, in which case FS is equivalent to best subset selection (BSS, Hastie et al. 2020). Letting FS($k$) and EFS($k, m$) refer to forward selection after $k$ steps and its ensemble, respectively, we obtain the following theoretical results.

- Under a condition on the expected squared coefficients and for $p$ large enough, a fixed $m < p$ improves expected training error over FS($k$).
- If the degrees of freedom of FS($k$) form a concave sequence in $k$, then the degrees of freedom of EFS($k, m$) are nondecreasing in $m$.

- If the concavity condition does not hold, then the degrees of freedom of EFS$(k, m)$ need not be monotone in $m$, and can exceed those of FS$(k)$ at intermediate values.

Together, the training error and concavity conditions imply that a single deterministic choice of $m < p$ gives smaller expected training error and no larger degrees of freedom than FS$(k)$, and hence smaller fixed design prediction error. On the other hand, degrees of freedom reduction is not guaranteed but depends on certain structural assumptions. Furthermore, in both ridge and lasso regression under orthogonality, the expected training error (resp. degrees of freedom) is nondecreasing (resp. nonincreasing) in the regularization parameter $\lambda$ [Zou et al., 2007, Hastie et al., 2009], yielding a monotone tradeoff between the two quantities. Since this monotonic relationship does not hold for $m$ in EFS$(k, m)$, we see that the analogy with ridge and lasso does not fully describe the role of randomization in this setting.

To prove these results, we characterize how feature subsampling smooths the hard thresholding of FS$(k)$ through adaptive reweighting of OLS coefficients. These weights depend only on the ranks of the observed OLS coefficients and not their values. We establish an exact finite recurrence for the selection probabilities, which, in the asymptotic regime $m, p \to \infty$ with $m/p \to \gamma \in (0, 1)$, yields limiting weights bounded by explicit logistic functions, with growth rate depending on $\gamma$ and a transition near the selection cutoff $k$ when $k$ is large. The same weight representation lets us compare training errors coordinate by coordinate and use majorization to compare degrees of freedom under the concavity condition.

## 2 Preliminaries

### 2.1 Setup and notation

Assume a fixed design regression model $y_i = f(\mathbf{x}_i) + \varepsilon_i$, where $f$ is an arbitrary regression function and $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (IID) noise variables with mean zero and variance $\sigma^2 > 0$. Noiseless constructions and limits are stated explicitly. We assume $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times p}$ is a fixed design matrix and write $\mathbf{y} = (y_1, \ldots, y_n)^\top$. For vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, we define the inner product as $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b}/n$, with associated norm $\|\mathbf{a}\|^2 = \langle \mathbf{a}, \mathbf{a} \rangle$. The notation $\|\cdot\|_{\ell_2}$ instead denotes the ordinary Euclidean norm. For each $j \in [p] = \{1, \ldots, p\}$, let $\mathbf{x}_{\cdot j,p} = (x_{1j}, x_{2j}, \ldots, x_{nj})^\top$ denote the $j$-th feature vector, which we normalize in the theoretical results so that $\|\mathbf{x}_{\cdot j,p}\| = 1$. We use regular font to denote functions such as the coordinate functions $x_{j,p}$.

### 2.2 Algorithms

**Forward selection.** Forward selection (FS) adds one feature at each step, choosing the feature that most reduces training error after refitting all selected coefficients by least squares. Starting with an empty active set $\mathcal{M}_0 = \emptyset$ and $\hat{\mathbf{f}}_{0,p} = \mathbf{0}$, it proceeds for $l = 1, 2, \ldots, k$, where $0 \le k \le p$, by updating $\mathcal{M}_{l-1}$ with the index

$$j_l = \operatorname*{arg\,min}_{j \in \mathcal{V}_l} \|\mathbf{y} - \mathbf{P}_{\mathcal{M}_{l-1} \cup \{j\}} \mathbf{y}\| = \operatorname*{arg\,max}_{j \in \mathcal{V}_l} \frac{|\langle \mathbf{r}_{l-1}, \mathbf{x}_{\cdot j,p} \rangle|}{\|\mathbf{P}^\perp_{\mathcal{M}_{l-1}} \mathbf{x}_{\cdot j,p}\|}, \tag{1}$$

where $\mathcal{V}_l = [p] \setminus \mathcal{M}_{l-1}$ is the set of eligible indices. Here, $\mathbf{P}_{\mathcal{M}_l}$ is the projection matrix onto the column span of the selected features, and $\mathbf{P}^\perp_{\mathcal{M}_l}$ projects onto its orthogonal complement. We set

$$\hat{\mathbf{f}}_{l,p} = \mathbf{P}_{\mathcal{M}_l} \mathbf{y}.$$

The residual vector at step $l$ is $\mathbf{r}_l = \mathbf{y} - \hat{\mathbf{f}}_{l,p}$. A zero denominator in (1) is assigned score zero, since that feature adds no direction to the selected span. Ties are resolved by a fixed ordering of the original feature indices. After $k$ steps, the resulting predictive function $\hat{f}_{k,p}$ is referred to as the FS$(k)$ estimator.

**Ensemble forward selection.** Following Mentch and Zhou [2020], we introduce randomization into each FS$(k)$ step by modifying (1) so that the maximization is taken over a randomly drawn subset. Only candidate features are randomized, and every base estimator uses the same observations without

bootstrap sampling. Specifically, at step $l$, we sample a fresh candidate set $\mathcal{V}_l \subseteq [p] \setminus \mathcal{M}_{l-1}$ of size $m$ uniformly without replacement, where $1 \le m \le p$. If $|[p] \setminus \mathcal{M}_{l-1}| < m$, we use all remaining features without randomization. The update index $j_l$ is then chosen as in (1), but restricted to $j \in \mathcal{V}_l$. Stopping after $k$ steps yields a randomized FS($k$) estimator $\check{f}_{k,m,p,b}$ for replicate $b$. Repeating this process independently for $b = 1, \dots, B$ and aggregating via simple averaging yields the ensemble estimator

$$\hat{f}_{k,m,p,B} = \frac{1}{B} \sum_{b=1}^{B} \check{f}_{k,m,p,b}.$$

By the strong law of large numbers, conditional on the data, the estimator $\hat{f}_{k,m,p,B}$ converges almost surely to its expected value as $B \to \infty$, with the expectation taken over the randomness of the subset choices. For the rest of this paper, we refer to this limit as the EFS($k, m$) estimator, denoting it as $\hat{f}_{k,m,p}$. Appendix B.1 bounds the conditional mean squared error of finite ensemble averaging at order $1/B$, without requiring orthogonality.

### 2.3 Prediction error decomposition

In a fixed design setting, the complexity of an estimator $\hat{f}$ is often quantified via its *degrees of freedom* [Hastie et al., 2009],

$$\mathrm{df}(\hat{f}) = \sum_{i=1}^{n} \frac{\mathrm{Cov}(y_i, \hat{f}(\mathbf{x}_i))}{\sigma^2}.$$

Letting $\mathbf{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top$ and $\hat{\mathbf{f}} = (\hat{f}(\mathbf{x}_1), \dots, \hat{f}(\mathbf{x}_n))^\top$ denote the true and fitted function values respectively, the expected mean squared prediction error for estimating $\mathbf{f}$ can be expressed in terms of expected training error and degrees of freedom through the optimism identity[1]

$$\mathbb{E}\big[\|\mathbf{f} - \hat{\mathbf{f}}\|^2\big] + \sigma^2 = \mathbb{E}\big[\|\mathbf{y} - \hat{\mathbf{f}}\|^2\big] + \frac{2\sigma^2}{n} \mathrm{df}(\hat{f}). \tag{2}$$

We will therefore analyze how EFS($k, m$) compares to FS($k$) on each term on the right hand side of (2). Our comparisons concern fixed values of $k$ and $m$, but if either is selected using the response, the degrees of freedom must also account for this selection.

## 3 Properties of EFS($k, m$) weights

### 3.1 EFS($k, m$) as reweighted OLS

We assume orthogonal features for the remaining results, except in Section 6, its supplementary analyses, and the simulations.

**Assumption 3.1** (Orthogonality)**.** *The design matrix satisfies* $\widehat{\mathbf{\Sigma}} = \mathbf{X}^\top \mathbf{X}/n = \mathbf{I}_p$. *In particular,* $p \le n$.

Extending the results to general correlated designs is an important direction, but would require controlling how correlations perturb the adaptive ordering induced by the greedy procedure. With correlated features, each selection can change the remaining residual correlations and projection denominators, making selection probabilities depend on the Gram matrix and preceding selections. Following established practice, we therefore focus on the orthogonal case, which allows us to derive exact asymptotic expressions for interpretable coefficient behavior and precisely characterize how the randomization from feature subsampling propagates through the FS($k$) algorithm. This also enables direct comparisons with the broader regularization landscape, including ridge, lasso [Tibshirani, 1996], elastic net [Zou and Hastie, 2005], LARS [Efron et al., 2004], SCAD [Fan and Li, 2001], MCP [Zhang, 2010], and SureShrink [Donoho and Johnstone, 1995], since all of them have been studied in the orthogonal setting.

Let $\hat{f}_{\mathrm{OLS}} = \sum_{j=1}^{p} \hat{\beta}_{j,p} x_{j,p}$ denote the ordinary least squares (OLS) projection of the response onto the full dictionary. Under Assumption 3.1, its coefficients can be expressed as inner products

[1] Although this is similar in spirit to a bias-variance decomposition, the expected training error is not classical squared bias, and degrees of freedom is not estimator variance.

$\hat{\beta}_{j,p} = \langle \mathbf{y}, \mathbf{x}_{\cdot j,p} \rangle$. The signal projection coefficients are $\beta_{j,p} = \langle \mathbf{f}, \mathbf{x}_{\cdot j,p} \rangle$. It turns out that both the FS($k$) and EFS($k, m$) estimators can be written in terms of the OLS coefficients. For ease of notation, we establish an ordering convention for the features $x_{1,p}, x_{2,p}, \ldots, x_{p,p}$ based on these coefficients.

**Definition 3.1** (Feature ordering)**.** Let $\pi$ be the permutation that sorts the OLS coefficients by magnitude, using the same tie rule as FS. Write $\hat{\beta}_{(j),p} = \hat{\beta}_{\pi(j),p}$, $\beta_{(j),p} = \beta_{\pi(j),p}$, and $x_{(j),p} = x_{\pi(j),p}$, so that

$$|\hat{\beta}_{(1),p}| \geq |\hat{\beta}_{(2),p}| \geq \cdots \geq |\hat{\beta}_{(p),p}|. \tag{3}$$

Note that this ordering is random as it depends on the response noise. For coefficients and features, subscripts without parentheses denote original feature indices, whereas parenthesized subscripts denote ranks in the ordering (3). Under orthogonality, ties have probability zero whenever the noise has a joint density. The fixed tie rule also defines the estimators otherwise.

Using Definition 3.1 together with the equivalence of FS and BSS, we can write the FS($k$) estimator as

$$\hat{f}_{k,p} = \sum_{j=1}^{p} \mathbf{1}(j \leq k) \hat{\beta}_{(j),p} x_{(j),p}, \tag{4}$$

which reflects the familiar fact that, in the orthogonal setting, FS($k$) (and BSS) selects the top $k$ features most correlated with the response. The indicators $\mathbf{1}(j \leq k)$ can be thought of as weights that modulate the OLS solution. Similarly, the ensemble estimator after $k$ steps has the form

$$\hat{f}_{k,m,p} = \sum_{j=1}^{p} w_j^{k,m,p} \hat{\beta}_{(j),p} x_{(j),p}, \tag{5}$$

where $w_j^{k,m,p} = \mathbb{P}(\pi(j) \in \mathcal{M}_k \mid \mathbf{y})$ is the probability that the feature ranked $j$ is selected, with randomness over candidate sampling. The weight at each rank depends only on $(j, k, m, p)$, but its assignment to a particular feature depends on the observed coefficient ordering. The asymptotic analysis in Section 3.3 shows how these weights smooth the hard thresholding of FS($k$).

This smoothing is related to the analysis of bagging and subagging by Bühlmann and Yu [2002], where observation resampling and averaging smooth unstable predictors such as decision stumps. In EFS, feature subsampling generates random inclusion indicators, and ensemble averaging replaces them by their selection probabilities. When $m = p$, every base learner equals FS($k$), so averaging alone leaves its hard threshold unchanged.

### 3.2 Basic properties

The weight sequence takes values in $[0, 1]$ and sums to $k$, since each base learner selects exactly $k$ features. Unfortunately, the exact values of the weights involve complicated combinatorics and do not have simple closed form expressions in general. Nonetheless, one can show that the sequence is weakly decreasing and, more importantly, derive a recurrence relation in terms of $j$, $k$, and $p$, which serves as our main technical tool for the subsequent results. We state and prove these technical facts as Lemmas B.1 and B.2.

Comparisons at the same $k$ match the number of greedy steps per base learner, but averaging can give EFS more than $k$ nonzero coefficients. For example, when $m = 1$ and $1 \leq k \leq p$ is fixed,

$$\hat{f}_{k,1,p} = \frac{k}{p} \hat{f}_{\mathrm{OLS}}, \qquad \mathrm{df}(\hat{f}_{k,1,p}) = k.$$

Thus the ensemble can be dense even when its degrees of freedom equal $k$. Our comparisons allow this larger support and do not establish an advantage under a constraint on the total number of selected features.

### 3.3 Logistic approximation

The finite sample recurrence is three dimensional (in $j$, $k$, and $p$) and inhomogeneous, making it challenging to find a simple closed form solution. However, in the scaling limit as $m, p \to \infty$ with $m/p \to \gamma \in (0, 1)$, the recurrence converges to a two dimensional system that decouples into a one dimensional subsystem with an explicit analytical solution.

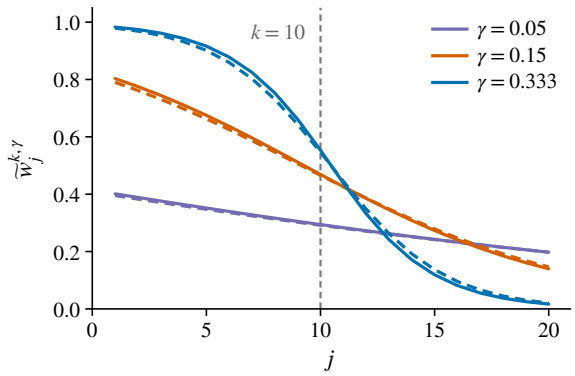



Figure 1: Comparison of $\tilde{w}_j^{k,\gamma}$ (solid) and the logistic approximation (8) (dashed) for $\gamma = 0.05, 0.15,$ and $0.333$, with $k = 10$.

**Theorem 3.1** (Asymptotic weights for fixed $k$)**.** *Fix positive integers $j, k$ and $\gamma \in (0, 1)$, and set $\alpha = -\log(1-\gamma)$. The limiting weights $\tilde{w}_j^{k,\gamma} = \lim_{\substack{m\to\infty,\, p\to\infty,\\ m/p\to\gamma}} w_j^{k,m,p}$*

*(a) exist and equal*

$$\sum_{i=1}^{k}(-1)^{k-i}\prod_{l=i}^{k}\left(e^{-\alpha(j-l)} - e^{-\alpha j}\right), \tag{6}$$

*(b) satisfy the bounds*

$$\frac{1}{1+e^{-\alpha(h(\alpha,k)-j)}} \le \tilde{w}_j^{k,\gamma} \le \frac{1}{1+e^{-\alpha(h(\alpha,k)+1-j)}}, \tag{7}$$

*where $h(\alpha, k) = \frac{1}{\alpha}\log(e^{\alpha k} - 1)$ and $\alpha \in (0, \infty)$ is defined via $\gamma = 1 - e^{-\alpha}$.*

The bounds in (7) reveal that the weights are sandwiched between two decreasing logistic functions with midpoints $h(\alpha, k)$ and $h(\alpha, k) + 1$ and common growth rate $\alpha$.[2] For $k \ge 1$ and large $m, p$ with $m/p \approx \gamma$, these bounds motivate the approximation

$$w_j^{k,m,p} \approx \frac{1}{1+e^{-\alpha(h(\alpha,k)+1/2-j)}}. \tag{8}$$

When $\gamma$ is small, the growth rate satisfies $\alpha \approx \gamma$.

These approximations clarify how the EFS$(k, m)$ parameters affect the shape of the weight function. As $k$ increases, the logistic curve translates rightward. As $\gamma$ increases (larger $m$), the curve becomes steeper, more closely approximating the sharp thresholding of FS$(k)$. Figure 1 illustrates this behavior for $\gamma \in \{0.05, 0.15, 0.333\}$.

For fixed $\alpha > 0$, $h(\alpha, k) - k \to 0$ as $k \to \infty$. When the selection threshold $k$ also tends to infinity, the logistic bounds tighten to curves separated by half an index. We capture this by introducing the "offset" parameter $d = k - j$, measuring distance from the selection threshold.

**Theorem 3.2** (Asymptotic weights for large $k$)**.** *Fix $d \in \mathbb{Z}$ and $\gamma \in (0, 1)$, and set $\alpha = -\log(1-\gamma)$. Along any integer sequence with $1 \le k, m \le p$, the limiting weights $\check{w}_d^{\gamma} = \lim_{\substack{k,p\to\infty,\, k/p\to 0,\\ m/p\to\gamma}} w_{k-d}^{k,m,p}$*

*(a) exist and equal*

$$\sum_{i=0}^{\infty}(-1)^i e^{-\alpha i(d+i/2+1/2)}, \tag{9}$$

[2] If the $m$ candidates are drawn independently with replacement from the remaining features, the same limiting weight formulas hold with $\alpha$ replaced by $\gamma$. The first step tail probability $(1 - j/p)^m$ tends to $e^{-\gamma j}$, giving the limiting rank recurrence with $e^{-\gamma}$ in place of $1 - \gamma = e^{-\alpha}$.

*(b) satisfy the symmetry property* $\breve{w}_d^\gamma = 1 - \breve{w}_{-(d+1)}^\gamma$,

*(c) satisfy the bounds*

*(i)* $\frac{1}{1+e^{-\alpha d}} \le \breve{w}_d^\gamma \le \frac{1}{1+e^{-\alpha(d+1/2)}}, \quad d < 0,$

*(ii)* $\frac{1}{1+e^{-\alpha(d+1/2)}} \le \breve{w}_d^\gamma \le \frac{1}{1+e^{-\alpha(d+1)}}, \quad d \ge 0,$

*(d) form a nondecreasing sigmoidal sequence.*[3]

*The convergence is uniform in rank, with* $\max_{1\le j\le p} \left| w_j^{k,m,p} - \breve{w}_{k-j}^\gamma \right| \to 0$ *along the same sequences.*

# 4 Main results

We now present our main results comparing EFS($k, m$) to FS($k$) in terms of the training error and degrees of freedom decomposition (2).

## 4.1 Expected training error

Since FS($k$) coincides with EFS($k, m$) when $m = p$, it is natural to ask whether a fixed choice of $m < p$ can give smaller expected training error. We show that such a deterministic choice exists and *quantify* the gap in expected training error.

**Theorem 4.1** (Training error improvement)**.** *Grant Assumption 3.1 and fix* $k \ge 1$. *Assume* $p \ge 2k$, $\sum_{j=k+1}^{2k} \mathbb{E}[\hat{\beta}^2_{(j),p}] > 0$, *and* $p$ *is large enough that*

$$p \ge 10 \frac{\sum_{j=1}^{k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big]}{\sum_{j=k+1}^{2k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big]}.$$

*Then there exists a deterministic* $m_p^\star < p$, *independent of the observed response, such that*

$$\mathbb{E}\big[\|\mathbf{y} - \hat{\mathbf{f}}_{k,m_p^\star,p}\|^2\big] \le \mathbb{E}\big[\|\mathbf{y} - \hat{\mathbf{f}}_{k,p}\|^2\big] - \frac{1}{4} \frac{\big(\frac{1}{k}\sum_{j=k+1}^{2k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big]\big)^2}{\frac{1}{k}\sum_{j=1}^{k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big]}. \tag{10}$$

Because FS orders variables by observed correlations, the condition and bound use the response induced coefficients $\hat{\beta}_{(j),p}$. To interpret the condition, write

$$A_{k,p} = \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big], \qquad B_{k,p} = \frac{1}{k}\sum_{j=k+1}^{2k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big].$$

When the ratio $A_{k,p}/B_{k,p}$ is near one, spreading weight to the next $k$ coefficients can improve the fit at little cost to the first $k$ coefficients, whereas a sharp drop is less favorable. The proof constructs $m_p^\star$ from these averages, independently of the realized response. See (31) in Appendix B.4.

**Example 4.1** (Constant regression function plus Gaussian noise)**.** Suppose $y_i = \mu + \varepsilon_i$, with constant $\mu$, IID $\mathcal{N}(0, \sigma^2)$ noise, and centered features. For fixed $k$ and $\sigma > 0$, the expected squared OLS coefficients near the selection cutoff have comparable magnitudes as $p \to \infty$ with $n = n_p \ge p + 1$, so $A_{k,p}/B_{k,p} \to 1$. Thus Theorem 4.1 gives smaller expected training error for all sufficiently large $p$. Appendix B.5 derives the resulting bound.

Improvement over $m = p$ alone, however, need not imply an interior optimum. Appendix B.8 gives a separate example where expected training error has a strict minimum at $1 < m < p$.

## 4.2 Degrees of freedom

We now show that, under a natural assumption, decreasing the feature subsampling ratio $m/p$ does not increase the degrees of freedom. The key is a decomposition of the degrees of freedom of EFS($k, m$) as a weighted sum of successive increments in the degrees of freedom of FS($j$).

[3] With $\Delta^2 \breve{w}_d^\gamma = \breve{w}_{d-1}^\gamma - 2\breve{w}_d^\gamma + \breve{w}_{d+1}^\gamma$, the centered second differences are nonnegative at $d \le -1$ and nonpositive at $d \ge 0$.

**Theorem 4.2** (Degrees of freedom improvement)**.** *Grant Assumption 3.1. For $0 \le k \le p$ and $1 \le m \le p$, the degrees of freedom satisfy*

$$\mathrm{df}(\hat{f}_{k,m,p}) = \sum_{j=1}^{p} w_j^{k,m,p} \Big( \mathrm{df}(\hat{f}_{j,p}) - \mathrm{df}(\hat{f}_{j-1,p}) \Big), \tag{11}$$

*where $\hat{f}_{0,p} = 0$ and $\mathrm{df}(\hat{f}_{0,p}) = 0$. In particular, if $j \mapsto \mathrm{df}(\hat{f}_{j,p})$ is concave on $\{0, \dots, p\}$, then $m \mapsto \mathrm{df}(\hat{f}_{k,m,p})$ is nondecreasing.*

**Corollary 4.1** (Prediction error improvement)**.** *Under the conditions of Theorem 4.1, suppose also that $j \mapsto \mathrm{df}(\hat{f}_{j,p})$ is concave. Then the same deterministic $m_p^\star$ in Theorem 4.1 satisfies*

$$\mathrm{df}(\hat{f}_{k,m_p^\star,p}) \le \mathrm{df}(\hat{f}_{k,p}), \qquad \mathbb{E}\big[\|\mathbf{f} - \hat{\mathbf{f}}_{k,m_p^\star,p}\|^2\big] < \mathbb{E}\big[\|\mathbf{f} - \hat{\mathbf{f}}_{k,p}\|^2\big].$$

The concavity condition in Theorem 4.2 means that each additional FS step adds less covariance complexity than the previous one. It is plausible at low signal to noise ratio (SNR), when $|\hat{\beta}_{(k),p}| \gg |\beta_{(k),p}|$. Indeed, in such cases,

$$\frac{\sigma^2}{n}\Big( \mathrm{df}(\hat{f}_{k,p}) - \mathrm{df}(\hat{f}_{k-1,p}) \Big) = \frac{\sigma^2}{n}\,\mathrm{df}\left(\hat{\beta}_{(k),p} x_{(k),p}\right) = \mathbb{E}\big[\hat{\beta}_{(k),p}\big(\hat{\beta}_{(k),p} - \beta_{(k),p}\big)\big] \approx \mathbb{E}\big[\hat{\beta}^2_{(k),p}\big], \tag{12}$$

where the final expression, $\mathbb{E}[\hat{\beta}^2_{(k),p}]$, is weakly decreasing in $k$ by Definition 3.1, suggesting concavity when the approximation is accurate. Empirical support for the concavity of FS($k$) degrees of freedom, even in settings that are not orthogonal, appears in Figures 1 and 4 of Hastie et al. [2020] and Figure 7 of Mentch and Zhou [2020].

*Remark.* In the constant regression setting of Example 4.1, $\beta_{(j),p} = 0$ for every rank $j$, so the approximation in (12) becomes exact and the degrees of freedom form a concave sequence.

The following lemma makes this intuition precise by giving a sufficient condition for concavity with a nonzero signal.

**Lemma 4.1** (Concavity at low SNR)**.** *Grant Assumption 3.1 and suppose the noise is Gaussian with $\sigma > 0$. Let $s = |\{j : \beta_{j,p} \ne 0\}|$ be the number of nonzero projected signal coefficients. For every $p \ge 3$ and $1 \le s \le p$, the sequence $k \mapsto \mathrm{df}(\hat{f}_{k,p})$ is concave on $\{0, 1, \dots, p\}$ whenever*

$$\max_{1 \le j \le p} |\beta_{j,p}| \le \frac{\sigma}{\sqrt{n}} \max\left\{1, \sqrt{2}\big(\sqrt{\log p} - \sqrt{\log s} - \sqrt{\log\log p}\big)\right\}. \tag{13}$$

*The sequence is also concave when $s = 0$ or $p = 2$.*

The same coefficient bound also gives $A_{k,p}/B_{k,p} \to 1$ for fixed $\sigma > 0$ as $p \to \infty$, with $n = n_p \ge p$ and $k = o(p)$, as shown in Lemma B.11. Thus these weak signals satisfy both conditions of Corollary 4.1, giving smaller expected training and prediction errors for the same deterministic $m_p^\star$ when $p$ is sufficiently large.

When concavity fails, the degrees of freedom need not be monotone in $m$. This can happen when a sharp boundary separates signal from noise, as in the sparse setting below, where the coefficient signal to noise ratios grow with $n$.

**Proposition 4.1** (Nonmonotonic degrees of freedom)**.** *Fix an integer $k \ge 1$, $\sigma > 0$, and $\gamma \in (0, 1)$. Grant Assumption 3.1 along any integer sequence with $p \ge k$, $p \to \infty$, $n = n_p \ge p$, and $1 \le m = m_p \le p$ satisfying $m_p/p \to \gamma$. Suppose $\mathbf{f} = \sum_{j=1}^{k} \mathbf{x}_{\cdot j,p}$ in the fixed original dictionary and the noise is Gaussian. Then*

$$\lim_{p \to \infty} \frac{\mathrm{df}(\hat{f}_{k,m_p,p}) - \mathrm{df}(\hat{f}_{k,p})}{\log p} = \begin{cases} 2(1-\gamma), & k = 1, \\ +\infty, & k \ge 2. \end{cases}$$

*For $k \ge 2$, more precisely, the gap multiplied by $\sigma/\sqrt{n}$ has a finite, strictly positive limit depending only on $k, \gamma$.*

## 5 Further comparisons

**Comparison with elastic net.** Elastic net regression [Zou and Hastie, 2005] combines the $\ell_1$ and $\ell_2$ penalties of lasso and ridge. Under Assumption 3.1, the "naïve elastic net" before rescaling has coefficients, for $\lambda_1, \lambda_2 \geq 0$,

$$\hat{\beta}^{\mathrm{EN}}_{(j),p} = \frac{(|\hat{\beta}_{(j),p}| - \lambda_1/2)_+}{1+\lambda_2} \operatorname{sgn}(\hat{\beta}_{(j),p}). \tag{14}$$

For comparison, the logistic approximation (8) gives the EFS$(k, m)$ coefficient, for large $m, p$ with $m/p \approx \gamma$, as

$$\hat{\beta}^{\mathrm{EFS}}_{(j),p} \approx \frac{\hat{\beta}_{(j),p}}{1 + e^{-\alpha(h(\alpha,k)+1/2-j)}}.$$

From (14), we see that even in the orthogonal setting, the ridge penalty shrinks all coefficients retained by lasso by the common factor $1/(1+\lambda_2)$, without changing their support at fixed $\lambda_1$. Feature subsampling instead redistributes a fixed total inclusion weight $k$ across ranks, smoothing the hard threshold of FS. The two families of attainable coefficient profiles therefore differ, although they can coincide for particular data and parameter choices.

The comparisons of BSS, FS, and lasso in Hastie et al. [2020] likewise show that different solution paths perform best in different SNR regimes. Ghosh et al. [2025] give related SNR dependent minimax results for parameter risk under random isotropic Gaussian designs. Under the conditions of Corollary 4.1, the EFS path improves expected training error without increasing degrees of freedom relative to BSS.

**Revisiting shrinkage.** Decreasing $m$ does not shrink every coefficient (some weights increase and others decrease), but it does shrink the norm of the fitted values.

**Proposition 5.1.** *Grant Assumption 3.1. For $0 \leq k \leq p$ and $1 \leq m < m' \leq p$,*

$$\|\hat{\mathbf{f}}_{k,m,p}\| \leq \|\hat{\mathbf{f}}_{k,m',p}\|.$$

The extremes are as follows. Maximal shrinkage occurs at $m = 1$, where $\|\hat{\mathbf{f}}_{k,1,p}\| = (k/p)\|\hat{\mathbf{f}}_{\mathrm{OLS}}\|$. No additional shrinkage from randomization occurs at $m = p$, where EFS$(k, m)$ reduces to FS$(k)$.

## 6 Another mechanism for training error reduction

Theorem 4.1 establishes training error reduction via the "enlarged model class" mechanism of Curth et al. [2024]. The orthogonality assumption makes FS equivalent to BSS, so randomization helps only by searching over a richer family of coefficient profiles. However, Liu and Mazumder [2025] identified a second mechanism. Feature subsampling can help *individual base estimators* avoid poor greedy choices.

We now construct examples demonstrating this second mechanism. In the orthogonal setting, FS$(k)$ always selects the $k$ features most correlated with the response, which means that individual EFS base estimators cannot improve its training error. We therefore introduce correlated features for which random candidate sets can prevent an initially attractive feature from being selected, allowing a different sequence of choices to reach a better fit. In these examples, FS ends at a strict local minimum. Replacing any one selected feature by any unselected feature increases the least squares training error.

**Proposition 6.1** (Escaping local minima)**.** *For every integer tuple $(n, p, k, m)$ with $n \geq p$, $2 \leq k \leq p-2$, and $1 \leq m < p$, there exist deterministic normalized, linearly independent feature columns and a noiseless response for which FS$(k)$ selects a strict local minimum under exchanges of a single feature, but*

$$\|\mathbf{y} - \hat{\mathbf{f}}_{k,m,p}\|^2 \leq \mathbb{E}_\theta\big[\|\mathbf{y} - \check{\mathbf{f}}_{k,m,p}\|^2\big] < \|\mathbf{y} - \hat{\mathbf{f}}_{k,p}\|^2,$$

*where $\check{f}_{k,m,p}$ is one randomized FS base estimator and $\theta$ denotes randomness in candidate sampling. The design and response may depend on $(n, p, k, m)$.*

Appendix B.13 also gives a finite sample guarantee for a fixed family of correlated designs, including improvement for every $2 \leq m < p$ when $\max\{3, p-m+1\} \leq k \leq p-1$.

## 7 Simulations

To assess the generality of Theorems 4.1 and 4.2 beyond orthogonal designs, we conducted a simulation study. We generated IID training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $n = 1000$, $p = 100$, where $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with first order autoregressive (AR(1)) correlation $\Sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0.5$, $y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and exact sparsity $\beta_j = \mathbf{1}(j \leq 10)$. The noise variance $\sigma^2$ was chosen to achieve the desired signal to noise ratio SNR $= \boldsymbol{\beta}^\top \mathbf{\Sigma} \boldsymbol{\beta} / \sigma^2$.

For EFS, we used $B = 500$ independent base learners and selected $(k, m)$ via ten fold cross validation, then evaluated the full path for $k \in [20]$ at the selected $m$. Both methods fitted an intercept. We compared covariance degrees of freedom and training error for EFS$(k, m)$ against FS$(k)$ at the same number of FS steps $k$ per base learner, averaging over 1,000 independent datasets at each SNR. We plot these two quantities directly because (2) decomposes fixed design prediction error into expected training error and degrees of freedom. These simulations concern a tuned procedure under correlated designs, whereas the theorems compare estimators at a fixed $m$ under orthogonality. Appendix A gives the simulation details and uncertainty estimates for paired differences. Code and synthetic results are available at `https://github.com/jasonklusowski/randomization-greedy-search`.

Figure 2 shows results for SNR $\in \{0.05, 0.25, 1.0\}$. At low SNR (0.05 and 0.25), EFS reduces degrees of freedom and prediction error relative to FS at every $k$. Training error decreases at every $k$ for SNR 0.25. At SNR 0.05, it increases for $k \leq 3$ and decreases for $k \geq 7$, with the differences at $k = 4, 5, 6$ within Monte Carlo uncertainty. At high SNR (1.0), degrees of freedom instead increase near the true sparsity level $k = 10$, consistent with Proposition 4.1, despite the violation of Assumption 3.1. Prediction error increases at $k = 10$ and decreases at every other $k$.

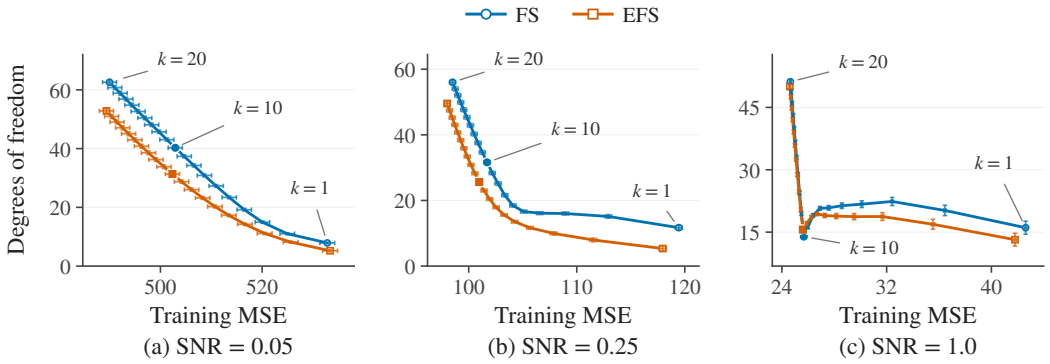



Figure 2: Degrees of freedom vs. training MSE for EFS$(k, m)$ and FS$(k)$ under a nonorthogonal design with $p/n = 0.1$, AR(1) correlation and exact sparsity, and $k \in [20]$ (true sparsity is 10). Means are over 1,000 independent datasets at each SNR, with pointwise 95% Monte Carlo intervals in each coordinate. Markers identify $k = 1, 10, 20$, with filled markers at $k = 10$. Appendix A reports uncertainty for paired differences.

## 8 Discussion

Under an orthogonal fixed design and the conditions of Theorems 4.1 and 4.2, a fixed EFS$(k, m)$ estimator can have smaller expected training error and no larger degrees of freedom than FS$(k)$, and hence smaller prediction error.

Mentch and Zhou [2020] concluded that feature subsampling is mainly useful in low SNR regimes, but our results show it can also help in the absence of noise (see Proposition 6.1). Mental models derived from ridge and lasso regression may not transfer cleanly to algorithmic regularization via feature subsampling, dropout, or gradient descent [Wager et al., 2013, Neyshabur et al., 2015, Arora et al., 2019, Bartlett et al., 2020].

The main theoretical limitation is orthogonality, partially addressed by simulations with a correlated design. Our exact results concern EFS; comparable guarantees for full RFs with tree structured split availability and bootstrap sampling remain open.

## Acknowledgments and Disclosure of Funding

Klusowski gratefully acknowledges support from the National Science Foundation through NSF CAREER grant DMS-2239448 and from the Alfred P. Sloan Foundation through a Sloan Research Fellowship. Tan was supported by NUS Startup Grant A-8000448-00-00 and the Singapore Ministry of Education (MOE) AcRF Tier 1 Grants A-8002498-00-00 and A-8004458-00-00.

During revision, we used OpenAI Codex (GPT-6 Astra) to assist with checking mathematical arguments, and developing and testing the replication code. The authors reviewed the resulting material and take responsibility for the final manuscript and code.

## References

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1907378117.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome Friedman, R. A. Olshen, and Charles J. Stone. *Classification and regression trees*. Chapman and Hall/CRC, 1984.

Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927 – 961, 2002. doi: 10.1214/aos/1031689014. URL https://doi.org/10.1214/aos/1031689014.

Matias D. Cattaneo, Rajita Chandak, and Jason M. Klusowski. Convergence rates of oblique regression trees for flexible function libraries. *The Annals of Statistics*, 52(2):466–490, 2024.

Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415 – 3438, 2022. doi: 10.1214/22-AOS2234. URL https://doi.org/10.1214/22-AOS2234.

Gwynneth H. Coogan and Ken Ono. A q-series identity and the arithmetic of Hurwitz zeta functions. *Proceedings of the American Mathematical Society*, 131(3):719–724, 2003. ISSN 00029939, 10886826. URL http://www.jstor.org/stable/1194472.

Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. Why do random forests work? Understanding tree ensembles as self-regularizing adaptive smoothers. *arXiv preprint arXiv:2402.01502*, 2024.

David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995. doi: 10.1080/01621459.1995.10476626. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476626.

N. R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, 1966.

Jin-Hong Du, Pratik Patil, and Arun Kumar Kuchibhotla. Subsample ridge ensembles: equivalences and generalized cross-validation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8585–8631, 2023.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. doi: 10.1214/009053604000000067. URL https://doi.org/10.1214/009053604000000067.

S. Elaydi. *An Introduction to Difference Equations*. Undergraduate Texts in Mathematics. Springer New York, third edition, 2005. ISBN 9780387276021. URL https://books.google.com/books?id=GWZKAAAAQBAJ.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Shubhangi Ghosh, Yilin Guo, Haolei Weng, and Arian Maleki. Signal-to-noise ratio aware minimax analysis of sparse linear regression. *arXiv preprint arXiv:2501.13323*, 2025.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 1988.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 11 2020. doi: 10.1214/19-STS733. URL https://doi.org/10.1214/19-STS733.

Jason M. Klusowski. Sparse learning with CART. In *Advances in Neural Information Processing Systems*, volume 33, pages 11612–11622, 2020.

Jason M. Klusowski and Peter M. Tian. Large Scale Prediction with Decision Trees. *Journal of the American Statistical Association*, 119(545):525–537, 2024.

Daniel LeJeune, Hamid Javadi, and Richard Baraniuk. The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 3525–3535. PMLR, 2020.

Brian Liu and Rahul Mazumder. Randomization can reduce both bias and variance: A case study in random forests. *Journal of Machine Learning Research*, 26(150):1–49, 2025.

Rahul Mazumder and Haoyue Wang. On the convergence of CART under sufficient impurity decrease condition. In *Advances in Neural Information Processing Systems*, volume 36, pages 57754–57782, 2023.

Tianxing Mei, Yingying Fan, and Jinchi Lv. Exogenous randomness empowering random forests. *Journal of Machine Learning Research*, 27(98):1–95, 2026. URL https://jmlr.org/papers/v27/25-0247.html.

Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171):1–36, 2020.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning, 2015. URL https://arxiv.org/abs/1412.6614.

Pratik Patil and Jin-Hong Du. Generalized equivalences between subsampling and ridge regularization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78926–78963. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f976982cd1c1b9e076c096787ef6652e-Paper-Conference.pdf.

Pratik Patil and Daniel LeJeune. Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Christos Revelas, Otilia Boldea, and Bas J. M. Werker. When do random forests work? *arXiv preprint arXiv:2504.12860*, 2025.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4): 1716–1741, August 2015.

Vasilis Syrgkanis and Manolis Zampetakis. Estimation and inference with trees and forests in high dimensions. In *Conference on Learning Theory*, pages 3453–3454. PMLR, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Stefan Wager, Sida Wang, and Percy S. Liang. Dropout training as adaptive regularization. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/38db3aed920cf82ab059bfccbd02be6a-Paper.pdf.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010. doi: 10.1214/09-AOS729. URL https://doi.org/10.1214/09-AOS729.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173 – 2192, 2007. doi: 10.1214/009053607000000127. URL https://doi.org/10.1214/009053607000000127.

# A Additional simulation details

## A.1 Correlated designs and Monte Carlo uncertainty

For the correlated design experiment, EFS varied $m$ over the grid $\{2, 3, 5, 8, 12, 19, 29, 44, 66, 100\}$. Ten fold cross validation jointly selected $(k, m)$ over this grid and $k \in [20]$. We then refitted the complete path at the selected $m$, so $m$ was the same for all displayed values of $k$ within a replicate. Each EFS fit averaged $B = 500$ independent randomized FS paths, with uniform candidate sampling without replacement at each step. Both methods fitted an unpenalized intercept, and each training fold was centered separately.

We generated 1,000 independent datasets at each SNR and repeated the entire tuning and fitting procedure on each dataset. FS and EFS used the same data within each replicate for a paired comparison. For a fitted vector $\hat{\mathbf{f}}$, we recorded training error $\|\mathbf{y} - \hat{\mathbf{f}}\|^2$, prediction error $\|\mathbf{f} - \hat{\mathbf{f}}\|^2$, and

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))(\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)),$$

where $\mathbf{f}$ is the known mean vector at the training design. Averaging the last quantity estimates covariance degrees of freedom for the full fitted procedure, including tuning and intercept fitting, averaged over the simulated designs.

For each quantity and $k$, we formed the EFS minus FS difference within each replicate before estimating its mean and Monte Carlo standard error. The standard error is the sample standard deviation of these paired differences divided by $\sqrt{1000}$. We obtained approximate pointwise 95% Monte Carlo intervals using the $t_{999}$ critical value. These intervals include variation from the data, tuning, and finite ensemble, and do not provide simultaneous coverage across $k$ or SNR. The coordinate error bars in Figure 2 use the same calculation separately for each method.

At SNR $0.05$, the paired training error differences at $k = 1, 2, 3$ have pointwise 95% intervals $[0.329, 0.767]$, $[0.429, 0.937]$, and $[0.177, 0.686]$, respectively. The paired prediction error intervals instead lie below zero for every $k \in [20]$ at both SNR $0.05$ and $0.25$. At SNR $1.0$, the corresponding intervals at $k = 10, 11$ are $[0.0058, 0.0325]$ and $[-0.0194, -0.0003]$, respectively.

# B Proofs

## B.1 Finite ensemble bound

For finite $B \geq 1$, the fitted vectors satisfy the conditional mean squared error bound

$$\mathbb{E}_\theta\big[\|\hat{\mathbf{f}}_{k,m,p,B} - \hat{\mathbf{f}}_{k,m,p}\|^2 \mid \mathbf{y}\big] \leq \frac{\|\mathbf{y}\|^2}{B}, \tag{15}$$

where $\theta$ denotes randomness in candidate sampling with the design fixed. The bound holds without orthogonality.

*Proof of* (15). Fix the design $\mathbf{X}$, the response $\mathbf{y}$, and integers $0 \leq k \leq p$, $1 \leq m \leq p$, and $B \geq 1$. Let $\check{\mathbf{f}}_{k,m,p,b}$ be the vector of fitted values from the $b$th randomized FS path, using independent candidate sampling across paths. Then

$$\hat{\mathbf{f}}_{k,m,p,B} = \frac{1}{B}\sum_{b=1}^{B} \check{\mathbf{f}}_{k,m,p,b}, \qquad \hat{\mathbf{f}}_{k,m,p} = \mathbb{E}_\theta[\check{\mathbf{f}}_{k,m,p,1} \mid \mathbf{y}].$$

Here, $\theta$ denotes the randomness in candidate sampling, with $\mathbf{X}$ fixed. All norms use the normalized inner product on response vectors, so $\|\mathbf{v}\|^2 = \mathbf{v}^\top\mathbf{v}/n$. Each path fits by least squares. Thus, if $\mathcal{M}_{k,b}$ is its selected set,

$$\check{\mathbf{f}}_{k,m,p,b} = \mathbf{P}_{\mathcal{M}_{k,b}}\mathbf{y}, \qquad \|\check{\mathbf{f}}_{k,m,p,b}\|^2 \leq \|\mathbf{y}\|^2.$$

The projection onto the selected column span is well defined even when those columns are linearly dependent.

Conditional on $\mathbf{y}$, the centered fitted vectors are independent and have mean zero. Expanding the squared norm of their average, the cross terms therefore have expectation zero, giving

$$\begin{aligned}
\mathbb{E}_\theta\big[\|\hat{\mathbf{f}}_{k,m,p,B} - \hat{\mathbf{f}}_{k,m,p}\|^2 \mid \mathbf{y}\big] &= \frac{1}{B^2}\sum_{b=1}^{B} \mathbb{E}_\theta\big[\|\check{\mathbf{f}}_{k,m,p,b} - \hat{\mathbf{f}}_{k,m,p}\|^2 \mid \mathbf{y}\big] \\
&= \frac{1}{B}\left(\mathbb{E}_\theta\big[\|\check{\mathbf{f}}_{k,m,p,1}\|^2 \mid \mathbf{y}\big] - \|\hat{\mathbf{f}}_{k,m,p}\|^2\right) \\
&\leq \frac{\|\mathbf{y}\|^2}{B}.
\end{aligned}$$

This proves the bound without requiring orthogonality or full column rank of the design. □

## B.2 Proofs for finite EFS$(k, m)$ weights

Throughout these proofs, $j$ denotes an observed rank, not an original feature label. Conditional on the response and its ordering permutation $\pi$, write $\mathcal{R}_l = \{j : \pi(j) \in \mathcal{M}_l\}$ and $\mathcal{C}_l = \{j : \pi(j) \in \mathcal{V}_l\}$ for the selected and candidate ranks. Orthogonality implies that the algorithm selects the smallest rank in $\mathcal{C}_l$. For recursive calculations we allow any integer $m \geq 1$, always sampling $\min\{m, p - l + 1\}$ candidates at step $l$ in a population of size $p$. We set $w_j^{0,m,p} = 0$ and assign weight zero to ranks outside $\{1, \ldots, p\}$, including the empty population.

**Lemma B.1** (Monotonicity)**.** *For integers $p, m \geq 1$, $0 \leq k \leq p$, and $1 \leq j < j' \leq p$, we have $w_j^{k,m,p} \geq w_{j'}^{k,m,p}$.*

**Lemma B.2** (Recurrence relation)**.** *For integers $p, m \geq 1$, $1 \leq k \leq p$, and $1 \leq j \leq p$, put $\bar{m} = \min\{m, p\}$. The weights satisfy the recurrence relation*

$$\begin{aligned}
w_j^{k,m,p} = \frac{\binom{p-j}{\bar{m}-1}}{\binom{p}{\bar{m}}} + \frac{\binom{p-j}{\bar{m}}}{\binom{p}{\bar{m}}} w_j^{k-1,m,p-1} \\
+ \left(1 - \frac{\binom{p-j+1}{\bar{m}}}{\binom{p}{\bar{m}}}\right) w_{j-1}^{k-1,m,p-1},
\end{aligned} \tag{16}$$

*where $\binom{a}{b} = 0$ for integers $a \geq 0$ and $b > a$ or $b < 0$. In particular, $w_j^{k,m,p} = \mathbf{1}(j \leq k)$ when $m \geq p$.*

*Proof of Lemma B.1.* To prove monotonicity of the weights with respect to $j$, recall that $w_j^{k,m,p} = \mathbb{P}(j \in \mathcal{R}_k)$. Here, the randomness is over the choices of the candidate sets $\mathcal{C}_1, \ldots, \mathcal{C}_k$. It suffices to compare adjacent ranks $a = j-1$ and $b = j$, with $2 \le j \le p$. Define the events

$$\mathcal{E}_b = \{b \in \mathcal{R}_k,\ a \notin \mathcal{R}_k\}, \qquad \mathcal{E}_a = \{a \in \mathcal{R}_k,\ b \notin \mathcal{R}_k\}.$$

On a path of candidate sets in $\mathcal{E}_b$, exchange $a$ and $b$ in every candidate set. Every selected rank other than $b$, on the original path, lies outside $\{a, b\}$. If its candidate set contains $a$ or $b$, it is smaller than both, since they are adjacent and $a$ is never selected. Thus the exchange preserves every other selection and replaces the selection of $b$ by $a$. The eligible sets are exchanged in the same way, so the transformed path is valid and belongs to $\mathcal{E}_a$. The exchange is its own inverse on sequences of candidate sets, hence is injective. Every valid path has the same probability

$$\prod_{l=1}^{k} \binom{p-l+1}{\min\{m, p-l+1\}}^{-1}.$$

Therefore $\mathbb{P}(\mathcal{E}_b) \le \mathbb{P}(\mathcal{E}_a)$, which proves the inequality for adjacent ranks and hence the lemma. □

*Proof of Lemma B.2.* Let $J$ denote the rank selected at the first step. The event $J > r$ means that all $\bar{m}$ candidates exceed $r$, so

$$\mathbb{P}(J > r) = \frac{\binom{p-r}{\bar{m}}}{\binom{p}{\bar{m}}}, \qquad r \in \{0, \ldots, p\}.$$

If $J = j$, rank $j$ is selected. If $J > j$, deleting the first selected rank leaves $j$ at rank $j$ in a population of size $p-1$. If $J < j$, it moves to rank $j-1$. The remaining $k-1$ steps use the same candidate size $m$, using all remaining features if fewer are available. Therefore, by the law of total probability,

$$w_j^{k,m,p} = \mathbb{P}(J = j) + \mathbb{P}(J > j) w_j^{k-1,m,p-1} + \mathbb{P}(J < j) w_{j-1}^{k-1,m,p-1}.$$

Substituting the survival probabilities above and using $\mathbb{P}(J = j) = \mathbb{P}(J > j-1) - \mathbb{P}(J > j)$ gives (16) by the Pascal identity. The zero conventions make the formula valid also at $j = 1$, $j = p$, and $p = 1$. When $m \ge p$, every step uses all remaining ranks and selects them in order. □

### B.3 Proofs for asymptotic EFS($k, m$) weights

To prove Theorem 3.1, we give several useful lemmas. Throughout this subsection, fix $0 < \gamma < 1$ and write $q = 1 - \gamma = e^{-\alpha} \in (0, 1)$. Unless another limit is specified, $j \ge 1$ and $k \ge 0$ are fixed integers. The first lemma establishes existence of the limiting weights and gives their recurrence.

**Lemma B.3.** *The limit* $\tilde{w}_j^{k,\gamma} = \lim_{\substack{m \to \infty,\ p \to \infty,\\ m/p \to \gamma}} w_j^{k,m,p}$ *exists. For* $k \ge 1$ *it satisfies*

$$\tilde{w}_j^{k,\gamma} = (1-q)q^{j-1} + q^j \tilde{w}_j^{k-1,\gamma} + (1 - q^{j-1}) \tilde{w}_{j-1}^{k-1,\gamma}, \tag{17}$$

*with* $\tilde{w}_j^{0,\gamma} = \tilde{w}_0^{k,\gamma} = 0$.

*Proof of Lemma B.3.* To show the limit form of the coefficients in recurrence (16), we derive the following result using basic asymptotic facts. For a fixed $j$, for $m, p \to \infty$ and $m/p \to \gamma$, we have

$$\frac{\binom{p-j}{m}}{\binom{p}{m}} = \prod_{i=0}^{j-1} \frac{p-m-i}{p-i} \longrightarrow q^j,$$

$$\frac{\binom{p-j}{m-1}}{\binom{p}{m}} = \frac{m}{p} \prod_{i=0}^{j-2} \frac{p-m-i}{p-1-i} \longrightarrow (1-q)q^{j-1}.$$

The empty product for $j = 1$ equals one. The third coefficient tends to $1 - q^{j-1}$ by the first formula with $j-1$ in place of $j$.

Now, we show that $\tilde{w}_j^{k,\gamma}$ is well defined via induction on $k$, simultaneously for all fixed ranks. For the base case, note that $w_j^{0,m,p} = 0$ for all $j$ and all $p$, so $\tilde{w}_j^{0,\gamma} = 0$ is well defined for all $j$. Now suppose

that $\tilde{w}_j^{k,\gamma}$ is well defined for all steps up to $k$ and for all $j$. Consider step $(k+1)$. In recurrence (16), the two weights on the right hand side are $w_j^{k,m,p-1}$ and $w_{j-1}^{k,m,p-1}$. Since $m/(p-1) \to \gamma$, they converge by the induction hypothesis. At $j = 1$ the term at rank zero is zero by convention. Also, $m < p - k$ eventually, so the sampling cap causes no difficulty at any of these states with a fixed number of steps. Therefore, taking limits in (16), we obtain the limit at step $k+1$ and the recurrence in Lemma B.3, completing the induction. □

**Lemma B.4** (Monotonicity of limiting weights)**.** *For all $j \geq 1$ and $k \geq 0$,*

$$\tilde{w}_j^{k,\gamma} \leq \tilde{w}_j^{k+1,\gamma}, \qquad \tilde{w}_j^{k,\gamma} \geq \tilde{w}_{j+1}^{k,\gamma}.$$

*Proof of Lemma B.4.* For each finite population, couple the fits at steps $k$ and $k+1$ along the same randomized path. Its selected rank sets are nested, so the inclusion probabilities are nondecreasing in $k$. Their rank monotonicity follows from Lemma B.1. Taking the limits at fixed ranks and steps gives both inequalities. □

However, recurrence (17) contains dependencies on both $\tilde{w}_j^{k-1,\gamma}$ and $\tilde{w}_{j-1}^{k-1,\gamma}$. These dependencies make it difficult to construct a closed form and uncover properties of $\tilde{w}_j^{k,\gamma}$. Ideally, we would like recurrences involving a change in $j$ to be separate from recurrences involving a change in $k$. Fortunately, recurrence (17) can be separated into the following pair of recurrences.

**Lemma B.5.** *From the recurrence in Lemma B.3, we can construct the following recurrence pair for $j, k \geq 1$*

$$\tilde{w}_j^{k,\gamma} = (q^{j-k} - q^j)(1 - \tilde{w}_j^{k-1,\gamma}), \tag{18}$$

$$\tilde{w}_j^{k,\gamma} = 1 - q^k - (q^{k-j+1} - q^k)\tilde{w}_{j-1}^{k,\gamma}. \tag{19}$$

*Proof of Lemma B.5.* For this proof, write $u_j^k = \tilde{w}_j^{k,\gamma}$. At rank one, (17) gives $u_1^k = (1-q) + qu_1^{k-1} = 1 - q^k$, which satisfies both recurrences directly. We will induct on $k$, simultaneously for all $j \geq 2$. First, we verify that both recurrences satisfy the base case of $k = 1$. Substituting $u_j^1 = (1-q)q^{j-1}$ and $u_j^0 = 0$ verifies both formulas.

Now, suppose that both recurrences are valid for steps up to $k$ and all $j$. Consider step $(k+1)$. Multiplying (19) by $q^{j-1-k}$ gives

$$(1 - q^{j-1})u_{j-1}^k = q^{j-1-k}(1 - q^k - u_j^k).$$

Substituting this identity into the recurrence in Lemma B.3, we obtain the following result

$$u_j^{k+1} = (q^{j-k-1} - q^j)(1 - u_j^k),$$

which proves (18) at step $k+1$ for every rank. Now, we prove that recurrence (19) is satisfied for step $(k+1)$ in a similar fashion. Put $A_j = q^{j-k-1}(1 - q^{k+1}) > 0$. We plug in recurrence (18) at ranks $j$ and $j-1$ into Lemma B.3, substituting $u_j^k = 1 - u_j^{k+1}/A_j$ and $u_{j-1}^k = 1 - u_{j-1}^{k+1}/A_{j-1}$. Since

$$\frac{q^j}{A_j} = \frac{q^{k+1}}{1 - q^{k+1}}, \qquad \frac{1 - q^{j-1}}{A_{j-1}} = \frac{q^{k-j+2} - q^{k+1}}{1 - q^{k+1}},$$

rearranging terms, we obtain the following recurrence

$$u_j^{k+1} = 1 - q^{k+1} - (q^{k-j+2} - q^{k+1})u_{j-1}^{k+1}.$$

Thus, recurrence (19) is satisfied for step $(k+1)$. Therefore, by induction, both recurrences are satisfied for all $j, k \geq 1$. □

*Proof of Theorem 3.1.* To prove (a), note that from Lemma B.3, we have that $\tilde{w}_j^{k,\gamma}$ exists. From (18), a first order inhomogeneous recurrence relation with variable coefficients, starting from the weight at step zero, we directly obtain the following closed form for $\tilde{w}_j^{k,\gamma}$ [Elaydi, 2005, Equation 1.2.4]

$$\tilde{w}_j^{k,\gamma} = \sum_{i=1}^{k} (-1)^{k-i} \prod_{l=i}^{k} \left(q^{j-l} - q^j\right),$$

which is (6).

To prove (b), we first show the lower bound. According to Lemma B.4, we have $\tilde{w}_j^{k,\gamma} \geq \tilde{w}_j^{k-1,\gamma}$. From (18), we know

$$\tilde{w}_j^{k,\gamma} \geq (q^{j-k} - q^j)(1 - \tilde{w}_j^{k,\gamma}).$$

Therefore, solving for $\tilde{w}_j^{k,\gamma}$, we have

$$\tilde{w}_j^{k,\gamma} \geq \frac{1 - q^k}{1 - q^k + q^{k-j}}.$$

For the upper bound, take $j \geq 2$. The multiplier $q^{k-j+1} - q^k$ is nonnegative, and rank monotonicity in Lemma B.4, together with (19), gives

$$\tilde{w}_j^{k,\gamma} \leq 1 - q^k - (q^{k-j+1} - q^k)\tilde{w}_j^{k,\gamma}.$$

Solving for $\tilde{w}_j^{k,\gamma}$, we have

$$\tilde{w}_j^{k,\gamma} \leq \frac{1 - q^k}{1 - q^k + q^{k-j+1}}.$$

For $j = 1$ this upper bound is equality, since $\tilde{w}_1^{k,\gamma} = 1 - q^k$. Finally, note that we can write

$$\frac{1 - q^k}{1 - q^k + q^{k-j}} = \frac{1}{1 + e^{\alpha j}/(e^{\alpha k} - 1)} = \frac{1}{1 + e^{-\alpha(h(\alpha,k)-j)}},$$

and the upper bound has the same form with $h(\alpha, k) + 1$ in place of $h(\alpha, k)$. □

*Remark.* The limiting weights are weakly decreasing in rank by Lemma B.4, and retain total mass $k$

$$\sum_{j=1}^{\infty} \tilde{w}_j^{k,\gamma} = k.$$

To prove the second property, note that the upper bound in (7) gives a summable geometric bound for each fixed $k$. We may therefore sum and reindex (17) to obtain

$$\begin{aligned}\sum_{j=1}^{\infty} \tilde{w}_j^{k,\gamma} &= \sum_{j=1}^{\infty}(1-q)q^{j-1} + \sum_{j=1}^{\infty} q^j \tilde{w}_j^{k-1,\gamma} + \sum_{j=1}^{\infty}(1 - q^{j-1})\tilde{w}_{j-1}^{k-1,\gamma} \\ &= 1 + \sum_{j=1}^{\infty} \tilde{w}_j^{k-1,\gamma} = k,\end{aligned}$$

by induction from $k = 0$. The restriction $0 < \gamma < 1$ matters here. When $m/p \to 0$, every limiting weight at a fixed rank and step is zero and mass escapes to increasing ranks. When $m/p \to 1$, however, the limits at fixed ranks are the weights from hard thresholding $\mathbf{1}(j \leq k)$.

We now turn to the proof of Theorem 3.2. Before that, we present a useful lemma regarding an alternative form of $\tilde{w}_j^{k,\gamma}$.

**Lemma B.6.** *For $j, k \geq 1$, the asymptotic weights $\tilde{w}_j^{k,\gamma}$ can be rewritten in the following closed form*

$$\tilde{w}_j^{k,\gamma} = \sum_{i=1}^{k}(-1)^{i+1} e^{\alpha i(k-j-i/2+1/2)} \prod_{l=k-i+1}^{k}(1 - e^{-\alpha l}).$$

*Proof of Lemma B.6.* Using the closed form from (a), we reindex each product by its length $i$. Factoring $e^{-\alpha(j-l)}$ out of its $l$th factor gives

$$\begin{aligned}\prod_{l=k-i+1}^{k}\left(e^{-\alpha(j-l)} - e^{-\alpha j}\right) &= e^{\alpha \sum_{l=k-i+1}^{k}(l-j)} \prod_{l=k-i+1}^{k}(1 - e^{-\alpha l}) \\ &= e^{\alpha i(k-j-i/2+1/2)} \prod_{l=k-i+1}^{k}(1 - e^{-\alpha l}),\end{aligned}$$

and the sign becomes $(-1)^{i+1}$. □

For the following proofs, we perform the substitution $d = k - j$. We first prove the existence and derive a closed form for $\check{w}_d^\gamma$ slightly different from (9).

**Lemma B.7.** *For every integer d, the limit* $\check{w}_d^\gamma = \lim_{k\to\infty} \tilde{w}_{k-d}^{k,\gamma}$ *exists and equals*

$$\check{w}_d^\gamma = \sum_{i=1}^{\infty} (-1)^{i+1} e^{\alpha i(d-i/2+1/2)}. \tag{20}$$

*Proof of Lemma B.7.* Fix $d$, let $j = k - d$, and take $k \geq \max\{1, d+1\}$ in Lemma B.6. Extend its sum by setting the summands to zero for $i > k$. For each fixed $i$, the product tends to one as $k \to \infty$. The absolute value of the $i$th summand is bounded by

$$e^{\alpha i(d+1/2)-\alpha i^2/2},$$

which is summable over $i \geq 1$ for fixed $d$ and $\alpha > 0$. Therefore, by dominated convergence, the limit $\check{w}_d^\gamma$ exists and is given by the absolutely convergent series (20). □

*Proof of Theorem 3.2.* We first prove the closed form and the stated properties for the iterated limit in Lemma B.7. The joint convergence is proved below. To obtain the closed form, we show that (20) is the same as (9). Using the difference of squares formula,

$$\check{w}_d^\gamma = e^{\alpha(d+1/2)^2/2} \sum_{i=1}^{\infty} (-1)^{i+1} e^{-\alpha(i-d-1/2)^2/2}.$$

Since $d$ is an integer, pairing $i$ with $2d + 1 - i$ cancels the terms with $1 \leq i \leq 2d$ when $d \geq 0$. For $d < 0$, the same pairing shows that the finite sum over $2d + 1 \leq i \leq 0$ vanishes, so those terms may be inserted. In either case the series may start at $i = 2d + 1$. Reindexing $i = r + 2d + 1$ gives

$$\check{w}_d^\gamma = \sum_{r=0}^{\infty} (-1)^r e^{-\alpha r(d+r/2+1/2)},$$

which proves (9). To prove (b), using the closed form from (20), note that

$$1 - \check{w}_{-(d+1)}^\gamma = 1 + \sum_{i=1}^{\infty} (-1)^i e^{-\alpha i(d+i/2+1/2)} = \check{w}_d^\gamma,$$

proving (b).

To prove (c), by taking $k \to \infty$ with $j = k - d$ in the bounds from (7), we have

$$\frac{1}{1 + e^{-\alpha d}} \leq \check{w}_d^\gamma \leq \frac{1}{1 + e^{-\alpha(d+1)}}. \tag{21}$$

Additionally, using the bounds from (21), we have

$$\check{w}_d^\gamma \leq \frac{1}{1 + e^{-\alpha(d+1)}} \leq \check{w}_{d+1}^\gamma, \tag{22}$$

demonstrating that $\check{w}_d^\gamma$ is nondecreasing for $d \in \mathbb{Z}$.

To complete the proof of (c) and prove (d), define the auxiliary real function

$$F(t) = \sum_{i=0}^{\infty} (-1)^i e^{-\alpha i(t+i/2+1/2)}, \qquad t \geq -1/2.$$

It agrees with $\check{w}_d^\gamma$ at every integer $d \geq 0$. Its series and its derivatives converge locally uniformly by Gaussian decay in $i$. In particular, $F$ is continuous at $-1/2$. Now, consider the following identity from Coogan and Ono [2003, Proposition 1.1], written for $0 < z, r < 1$

$$A(z, r) \coloneqq \sum_{i=0}^{\infty} (-1)^i z^{2i} r^{i^2} = \frac{1}{1+z} \sum_{i=0}^{\infty} z^i \prod_{l=0}^{i-1} \frac{1 - zr^l}{1 + zr^{l+1}}. \tag{23}$$

The formal identity applies analytically here. For fixed $r \in (0,1)$, both sides converge locally uniformly in $|z| < 1$, since the finite products are uniformly bounded on compact subsets of that disk. They therefore represent the same analytic function.

For $0 < z, r < 1$, each product factor lies between $(1-z)/(1+z)$ and one. Thus

$$A(z,r) \geq \frac{1}{1+z} \sum_{i=0}^{\infty} \left( \frac{z(1-z)}{1+z} \right)^i = \frac{1}{1+z^2}.$$

Set $r = e^{-\alpha/2}$ and $z = e^{-\alpha(t+1/2)/2}$ for $t > -1/2$. Then $F(t) = A(z,r)$ and

$$F(t) \geq \frac{1}{1+e^{-\alpha(t+1/2)}}.$$

For integer $d \geq 0$, this is the improved lower bound in (ii). Reflection using (b) gives the improved upper bound for $d < 0$. Together with (21), this proves (c). Taking $t \downarrow -1/2$ in the last inequality gives $F(-1/2) \geq 1/2$. Here we use continuity of the theta series at $t = -1/2$.

For fixed $z \in (0,1)$, each factor $(1-zr^l)/(1+zr^{l+1})$ is nonnegative and nonincreasing in $r \in (0,1)$. Hence every summand on the right of (23), and therefore their sum, is nonincreasing in $r$. The theta series on the left is differentiable in $r$ by Gaussian decay, so $\partial_r A(z,r) \leq 0$. At the values of $z, r$ above,

$$F''(t) = \alpha^2 \sum_{i=0}^{\infty} (-1)^i i^2 z^{2i} r^{i^2} = \alpha^2 r\, \partial_r A(z,r) \leq 0.$$

Thus $F$ is concave on $[-1/2, \infty)$, by continuity at the endpoint. It follows that the centered second difference of $\check{w}_d^\gamma$ is nonpositive for every integer $d \geq 1$. At $d = 0$, concavity between $-1/2$ and 1 gives

$$\check{w}_0^\gamma = F(0) \geq \frac{2}{3} F(-1/2) + \frac{1}{3} F(1) \geq \frac{1}{3} + \frac{1}{3} \check{w}_1^\gamma.$$

Since $\check{w}_{-1}^\gamma = 1 - \check{w}_0^\gamma$, this is exactly $\check{w}_{-1}^\gamma - 2\check{w}_0^\gamma + \check{w}_1^\gamma \leq 0$. The symmetry in (b) reverses the sign of the centered second difference under $d \mapsto -d-1$, proving nonnegative second differences for $d \leq -1$. Together with (22), this proves (d). □

To pass from the iterated limit to the joint limit in Theorem 3.2, we next control the finite population error, the dependence on the subsampling ratio, and the finite step error.

**Lemma B.8** (Finite population approximation)**.** *For integers $p \geq 2$, $1 \leq m < p$, and $1 \leq k \leq p$, put $\gamma = m/p$ and $q = 1-\gamma$. Then*

$$\max_{1 \leq j \leq p} \left| w_j^{k,m,p} - \tilde{w}_j^{k,\gamma} \right| \leq \min \left\{ 1, \frac{2}{p} \left( \frac{k-1}{\gamma^3} + \frac{q}{\gamma^4} \right) \right\}.$$

*Proof of Lemma B.8.* Consider an infinite list of ranks for a probabilistic representation of the limiting weights. At each step, remove the $G$th remaining rank, where the successive choices are independent and $\mathbb{P}(G = r) = \gamma q^{r-1}$ for $r \geq 1$. Conditioning on the first choice gives recurrence (17), so its inclusion probabilities after $k$ steps are $\tilde{w}_j^{k,\gamma}$.

Track the current eligible rank $r$ of one fixed original feature until it is selected. At finite time $t$, let $N = p - t$, let $J_N$ be the minimum rank in a uniform subset of size $\min\{m, N\}$, and put $F_t(r) = \mathbb{P}(J_N > r)$, with $F_t(0) = 1$. We claim that, for $1 \leq r \leq N$,

$$0 \leq q^r - F_t(r) \leq \frac{q^{r-1}}{p} \left( rt + \frac{r(r-1)}{2} \right). \tag{24}$$

For $0 \leq i < N$, put $a_i := \max\{0, 1 - m/(N-i)\}$. Then $F_t(r) = \prod_{i=0}^{r-1} a_i$ for every $1 \leq r \leq N$, including when the candidate size is capped by $N$. Each factor lies in $[0, q]$. If $N - i \geq m$, then

$$q - a_i = \frac{m(t+i)}{p(N-i)} \leq \frac{t+i}{p}.$$

If $N - i < m$, then $a_i = 0$ and $t + i > p - m = pq$, so the same upper bound holds. Expanding the difference of the two products gives

$$0 \le q^r - F_t(r) \le q^{r-1} \sum_{i=0}^{r-1} (q - a_i) \le \frac{q^{r-1}}{p} \left( rt + \frac{r(r-1)}{2} \right),$$

which proves (24) for every eligible rank.

From state $r$, the probabilities of selection, a decrease to rank $r-1$, and remaining at rank $r$ are respectively

$$F_t(r-1) - F_t(r), \qquad 1 - F_t(r-1), \qquad F_t(r).$$

The infinite process replaces $F_t(s)$ by $q^s$. Writing $\Delta_t(s) = q^s - F_t(s)$, the total variation distance between these distributions on three points is $\max\{\Delta_t(r-1), \Delta_t(r)\}$. For $0 \le t \le k-1$, this is bounded by $b_{r-1} + b_r$, where $b_0 = 0$ and

$$b_r = \frac{q^{r-1}}{p} \left( r(k-1) + \frac{r(r-1)}{2} \right), \qquad r \ge 1.$$

Until their first disagreement, couple the two rank processes so that their conditional probability of disagreement is this total variation distance. Continue the infinite process after any disagreement. In the infinite process, a rank cannot be revisited after it is left, and the probability of remaining at rank $r$ is $q^r$. Its expected number of visits to that rank is therefore at most $(1 - q^r)^{-1} \le \gamma^{-1}$, whether or not the rank is reached. The probability of a first disagreement is bounded by the expected accumulated errors along the infinite process. Hence it is at most

$$\begin{aligned} \frac{1}{\gamma} \sum_{r=1}^{\infty} (b_{r-1} + b_r) &= \frac{2}{p\gamma} \sum_{r=1}^{\infty} q^{r-1} \left( r(k-1) + \frac{r(r-1)}{2} \right) \\ &= \frac{2}{p} \left( \frac{k-1}{\gamma^3} + \frac{q}{\gamma^4} \right). \end{aligned}$$

If no disagreement occurs, the two inclusion indicators agree. The bound is independent of its original rank $j$, and the difference between two inclusion probabilities is also at most one. □

**Lemma B.9** (Dependence on the subsampling ratio)**.** *For $\gamma, \gamma' \in (0,1)$, put $\gamma_0 = \min\{\gamma, \gamma'\}$. For every integer $k \ge 0$,*

$$\sup_{j \ge 1} \left| \tilde{w}_j^{k,\gamma} - \tilde{w}_j^{k,\gamma'} \right| \le \min \left\{ 1, \frac{2|\gamma - \gamma'|}{\gamma_0^3} \right\}.$$

*The same bound holds for $\sup_{d \in \mathbb{Z}} \left| \check{w}_d^{\gamma} - \check{w}_d^{\gamma'} \right|$.*

*Proof of Lemma B.9.* The two geometric survival probabilities at rank $r$ differ by at most

$$r(1 - \gamma_0)^{r-1} |\gamma - \gamma'|.$$

As in the proof of Lemma B.8, the total variation distance between the transition distributions is at most the sum of the two adjacent survival differences, and the expected number of visits to any rank is at most $1/\gamma_0$. Summing these errors gives

$$\frac{2|\gamma - \gamma'|}{\gamma_0} \sum_{r=1}^{\infty} r(1 - \gamma_0)^{r-1} = \frac{2|\gamma - \gamma'|}{\gamma_0^3}.$$

The claim is immediate when $k = 0$. For each fixed integer $d$, let $k \to \infty$ with $j = k - d$ and use Lemma B.7. The bound does not depend on $d$, which proves the assertion for the offset profile. □

**Lemma B.10** (Finite step approximation)**.** *For every integer $k \ge 1$ and $\gamma \in (0,1)$, put $q = 1 - \gamma$. Then*

$$\sup_{j \ge 1} \left| \tilde{w}_j^{k,\gamma} - \check{w}_{k-j}^{\gamma} \right| \le \min \left\{ 1, q^k (1 + \gamma^{-2}) \right\}. \tag{25}$$

*Proof of Lemma B.10.* For this proof, write $u_j^t = \tilde{w}_j^{t,\gamma}$ and $W_d = \check{w}_d^\gamma$, and put $e_j^t \coloneqq u_j^t - W_{t-j}$. Reindexing the absolutely convergent series in (9) gives

$$W_d = 1 - q^{d+1} W_{d+1}.$$

First take $j > k$. By (18) and the profile identity,

$$e_j^t = -q^{j-t} e_j^{t-1} - q^j (1 - u_j^{t-1}), \qquad 1 \le t \le k.$$

The alternating series in (20) has decreasing positive terms starting at $q^j$ when $d = -j$, so $|e_j^0| = W_{-j} \le q^j$. Taking absolute values and iterating, using $0 \le u_j^t \le 1$ and $q^{j-t} \le q^{j-k}$, gives

$$|e_j^k| \le q^j \sum_{i=0}^{k} q^{i(j-k)} \le \frac{q^j}{1 - q^{j-k}} \le \frac{q^k}{\gamma}.$$

Now take $1 \le j \le k$. At rank one, $u_1^k = 1 - q^k$ and the profile identity give $e_1^k = q^k (W_k - 1)$, so $|e_1^k| \le q^k$. For $j \ge 2$, (19) gives

$$e_j^k = -q^{k-j+1} e_{j-1}^k - q^k (1 - u_{j-1}^k).$$

Taking absolute values and iterating over ranks up to $j$, whose coefficients are at most $q^{k-j+1}$, gives

$$|e_j^k| \le q^k \sum_{i=0}^{j-1} q^{i(k-j+1)} \le \frac{q^k}{1 - q^{k-j+1}} \le \frac{q^k}{\gamma}.$$

Since both weights lie in $[0, 1]$ and $\gamma^{-1} \le 1 + \gamma^{-2}$, the two cases prove (25). □

*Joint convergence in Theorem 3.2.* Fix $\gamma \in (0, 1)$ and let $a = m/p \in (0, 1)$. Lemmas B.8, B.9, and B.10 give

$$\max_{1 \le j \le p} \left| w_j^{k,m,p} - \check{w}_{k-j}^\gamma \right| \le \frac{2}{p} \left( \frac{k-1}{a^3} + \frac{1-a}{a^4} \right) + (1-a)^k (1 + a^{-2}) + \frac{2|a - \gamma|}{\min\{a, \gamma\}^3}.$$

Along the stated sequences, $a \to \gamma$, $k/p \to 0$, and $k \to \infty$, so every term on the right tends to zero. This proves uniform convergence over the ranks $1 \le j \le p$. For any fixed integer $d$, the rank $j = k - d$ belongs to this range eventually. Its limit is therefore the profile already computed above, which proves the assertion about the joint limit and completes the theorem. □

### B.4 Proof of Theorem 4.1

*Proof.* From the decompositions (5) and (4), we have the following identity

$$\|\mathbf{y} - \hat{\mathbf{f}}_{k,p}\|^2 - \|\mathbf{y} - \hat{\mathbf{f}}_{k,m,p}\|^2 = \sum_{j=k+1}^{p} \hat{\beta}_{(j),p}^2 \Big[ 1 - \big(1 - w_j^{k,m,p}\big)^2 \Big] - \sum_{j=1}^{k} \hat{\beta}_{(j),p}^2 \big(1 - w_j^{k,m,p}\big)^2. \tag{26}$$

Let $m_p$ be deterministic and chosen below. For $p \ge 2k$, taking expectations in (26) and discarding the nonnegative terms with $j > 2k$ gives

$$\mathbb{E}\Big[ \|\mathbf{y} - \hat{\mathbf{f}}_{k,p}\|^2 - \|\mathbf{y} - \hat{\mathbf{f}}_{k,m_p,p}\|^2 \Big] \ge \sum_{j=k+1}^{2k} \mathbb{E}\big[\hat{\beta}_{(j),p}^2\big] \Big[ 1 - \big(1 - w_j^{k,m_p,p}\big)^2 \Big] - \sum_{j=1}^{k} \mathbb{E}\big[\hat{\beta}_{(j),p}^2\big] \big(1 - w_j^{k,m_p,p}\big)^2. \tag{27}$$

To simplify further, set

$$A_{k,p} = \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big], \qquad B_{k,p} = \frac{1}{k}\sum_{j=k+1}^{2k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big],$$

and note that $A_{k,p} \geq B_{k,p} > 0$. By Definition 3.1 and Lemma B.1, both $\mathbb{E}[\hat{\beta}^2_{(j),p}]$ and $w_j^{k,m_p,p}$ are weakly decreasing in $j$. The Chebyshev sum inequality [Hardy et al., 1988] therefore gives

$$\sum_{j=1}^{k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big]\big(1 - w_j^{k,m_p,p}\big)^2 \leq A_{k,p}\sum_{j=1}^{k}\big(1 - w_j^{k,m_p,p}\big)^2,$$

and

$$\sum_{j=k+1}^{2k} \mathbb{E}\big[\hat{\beta}^2_{(j),p}\big]\Big[1 - \big(1 - w_j^{k,m_p,p}\big)^2\Big] \geq B_{k,p}\sum_{j=k+1}^{2k}\Big[1 - \big(1 - w_j^{k,m_p,p}\big)^2\Big]. \tag{28}$$

We first bound the finite weights for $2 \leq m_p \leq p - k$. Put

$$N = p - k + 1, \qquad b = N - m_p, \qquad q = \frac{b}{N}, \qquad w_j = w_j^{k,m_p,p},$$

and define

$$d_j = 1 - w_j \quad (1 \leq j \leq k), \qquad D = \sum_{j=1}^{k} d_j = \sum_{j=k+1}^{p} w_j, \qquad L = \sum_{j=2k+1}^{p} w_j.$$

The equality defining $D$ follows because the weights sum to $k$. Thus $D$ is the total weight assigned beyond the first $k$ ranks, and $L$ is the part assigned beyond rank $2k$. For $h \geq 1$, let

$$a_h = \prod_{r=0}^{h-1} \frac{b + r}{N + r}, \qquad a_0 = 1.$$

We claim that

$$d_j \leq a_{k-j+1} \leq q, \qquad q \leq D \leq \frac{b}{m_p - 1}, \qquad L \leq q^k \frac{b}{m_p - 1}. \tag{29}$$

For the first bound, induct on the number of selections. Let $J$ be the rank selected at the first step. Conditional on $J < j$, the remaining problem has population $p - 1$, $k - 1$ selections and target rank $j - 1$. Conditional on $J > j$, the target rank remains $j$. Both problems have the same $N$ and $b$. With $h = k - j + 1$, induction therefore gives

$$d_j \leq \mathbb{P}(J < j)a_h + \mathbb{P}(J > j)a_{h-1}.$$

When $j = k$, the second bound uses only $a_0 = 1$. When $j = 1$, the first event has probability zero. Conditional on $J \geq j$, all candidates lie among $p - j + 1$ ranks, so

$$\frac{\mathbb{P}(J > j)}{\mathbb{P}(J \geq j)} = \frac{p - j + 1 - m_p}{p - j + 1} = \frac{b + h - 1}{N + h - 1} = \frac{a_h}{a_{h-1}}.$$

Substitution proves $d_j \leq a_h$. The case $k = 1$ initializes the induction exactly. Every factor in $a_h$ is at most one, so $a_h \leq q$.

After $s$ selections, at least $k - s$ ranks among the first $k$ remain, which we use to bound $D$. Conditional on any history, the probability that the next selection has rank greater than $k$ is therefore at most

$$\frac{\binom{p-k}{m_p}}{\binom{p-s}{m_p}} = a_{k-s}.$$

Summing over the $k$ steps gives $D \leq \sum_{h=1}^{k} a_h$. The recurrence for $a_h$ gives $(b + h - 1)a_{h-1} - (b + h)a_h = (m_p - 1)a_h$. Telescoping therefore yields

$$\sum_{h=1}^{k} a_h = \frac{b - (b + k)a_k}{m_p - 1} \leq \frac{b}{m_p - 1}.$$

If no rank greater than $k$ has been selected before the final step, exactly one of the first $k$ remains, and it is missed with probability $q$. Thus the probability of selecting at least one rank greater than $k$ is at least $q$, which proves $D \geq q$.

The same argument for the first $2k$ ranks gives

$$L \leq \sum_{s=0}^{k-1} \frac{\binom{p-2k}{m_p}}{\binom{p-s}{m_p}}.$$

When $b \leq k$, the numerator is zero and $L = 0$. Otherwise,

$$\frac{\binom{p-2k}{m_p}}{\binom{p-k}{m_p}} = \prod_{i=1}^{k} \frac{b-i}{N-i} \leq q^k.$$

All denominators are positive because $p \geq 2k$. Combining this with the bound on $\sum_{h=1}^{k} a_h$ proves the last inequality in (29).

Write $t = B_{k,p}/A_{k,p}$ and let $G$ denote the expected training error gap on the left side of (27). The first bound in (29) gives

$$\sum_{j=1}^{k} d_j^2 \leq qD.$$

Also, the weights are nonnegative, so

$$\sum_{j=k+1}^{2k} w_j(2 - w_j) \geq 2(D - L) - D^2.$$

It follows from (27)–(28) that

$$\frac{G}{A_{k,p}} \geq D(2t - q - tD) - 2tL. \tag{30}$$

We now choose

$$N = p - k + 1, \qquad b = \left\lfloor \frac{NB_{k,p}}{3A_{k,p}} \right\rfloor, \qquad m_p^\star = N - b = \left\lceil N\left(1 - \frac{B_{k,p}}{3A_{k,p}}\right) \right\rceil. \tag{31}$$

For fixed $p, k$, these quantities depend only on the averages $A_{k,p}, B_{k,p}$ of expected squared coefficients and are fixed before observing the response. Since $p \geq 10/t$ and $p \geq 2k$, we have $N \geq p/2 + 1$, $N \geq 6$ and $1/N \leq t/5$. Also $Nt \geq 5 + t > 5$, so $1 \leq b < N$ and $m_p^\star \geq 2N/3 > 1$. Thus $2 \leq m_p^\star \leq p - k$, and all the preceding bounds apply with $m_p = m_p^\star$.

Put $e = 1/N$ and $x = q/t$. Since $Nt/3 > 1$ and $\lfloor y \rfloor \geq y/2$ for $y \geq 1$, the floor in (31) gives

$$\frac{1}{6} \leq x \leq \frac{1}{3}.$$

Moreover,

$$D \leq \frac{q}{1 - q - e} \leq \frac{xt}{1 - (x + 1/5)t} \leq \frac{5}{7}.$$

The function $D \mapsto D(2t - q - tD)$ is increasing for $D \leq 1 - x/2$, and

$$1 - \frac{x}{2} \geq \frac{5}{6} > \frac{5}{7}.$$

Using $D \geq q$ and the bound on $L$ in (30) therefore gives

$$\frac{G}{A_{k,p}} \geq 2tq - (1 + t)q^2 - \frac{2tq^{k+1}}{1 - q - e}.$$

When $k \geq 2$, we have $q^{k+1} \leq q^3$, and hence

$$\frac{G}{B_{k,p}^2/A_{k,p}} \geq H_t(x), \qquad H_t(x) = 2x - (1 + t)x^2 - \frac{2x^3t^2}{1 - (x + 1/5)t}.$$

For fixed $x \in [1/6, 1/3]$, this expression is decreasing in $t \in (0, 1]$. Thus $H_t(x) \geq H_1(x)$. The function

$$H_1(x) = 2x - 2x^2 - \frac{2x^3}{4/5 - x}$$

is concave on $(0, 4/5)$, since $H_1''(x) = -256/[125(4/5 - x)^3] < 0$. Its minimum on $[1/6, 1/3]$ is therefore attained at an endpoint, where

$$H_1(1/6) = \frac{5}{19} > \frac{1}{4}, \qquad H_1(1/3) = \frac{2}{7} > \frac{1}{4}.$$

This proves the claimed bound when $k \geq 2$.

For $k = 1$, the same choice has $N = p$ and $pt \geq 10$. Since $\lfloor y \rfloor \geq 3y/4$ for $y \geq 3$, we have $x \in [1/4, 1/3]$. The exact weights are

$$w_1^{1,m_p^\star,p} = 1 - q, \qquad w_2^{1,m_p^\star,p} = \frac{p}{p-1} q(1-q) \geq q(1-q).$$

Keeping only the reward at rank two and using that $u(2 - u)$ is increasing on $[0, 1]$, the expected training error gap satisfies

$$\frac{G}{B_{k,p}^2 / A_{k,p}} \geq 2x - x^2 - 3x^2 t + 2x^3 t^2 - x^4 t^3.$$

The right side is decreasing in $t$, with derivative $-x^2(3 - 4xt + 3x^2t^2) < 0$. At $t = 1$, it equals $g(x) = 2x - 4x^2 + 2x^3 - x^4$. This function is concave, and

$$g(1/4) = \frac{71}{256} > \frac{1}{4}, \qquad g(1/3) = \frac{23}{81} > \frac{1}{4}.$$

The claimed bound follows for the single deterministic estimator EFS$(k, m_p^\star)$. □

### B.5 Constant regression function plus Gaussian noise

Suppose $y_i = \mu + \varepsilon_i$, where $\mu$ is constant and $\varepsilon_1, \ldots, \varepsilon_n$ are IID $\mathcal{N}(0, \sigma^2)$ noise variables. Assume the constant component is orthogonal to the feature dictionary, $\langle \mathbf{1}_n, \mathbf{x}_{\cdot j,p} \rangle = 0$ for all $j$, as happens after fitting an intercept or centering the feature columns. Consequently, the fixed design mean vector $\mathbf{f} = \mu \mathbf{1}_n$ has no projection onto the feature directions. Under Assumption 3.1, the unsorted OLS coefficients $\langle \boldsymbol{\varepsilon}, \mathbf{x}_{\cdot j,p} \rangle$ are independent $\mathcal{N}(0, \sigma^2/n)$ random variables. After sorting them by decreasing absolute value as in Definition 3.1, standard Gaussian extreme value asymptotics imply that, for each fixed $r$,

$$\mathbb{E}\big[\hat{\beta}_{(r),p}^2\big] = (1 + o(1)) \frac{2\sigma^2 \log p}{n}, \qquad p \to \infty.$$

Thus, for fixed $k$ and $\sigma > 0$, along any sequence $n = n_p \geq p + 1$, we have $A_{k,p}/B_{k,p} \to 1$ and the condition in Theorem 4.1 holds for all sufficiently large $p$. The theorem gives the more interpretable bound

$$\mathbb{E}\big[\|\mathbf{y} - \hat{\mathbf{f}}_{k,m_p^\star,p}\|^2\big] \leq \mathbb{E}\big[\|\mathbf{y} - \hat{\mathbf{f}}_{k,p}\|^2\big] - (1 + o(1)) \frac{\sigma^2 \log p}{2n}.$$

The comparable expected squared coefficients near the selection cutoff allow EFS to improve the fit by spreading the weight budget across more coefficients.

### B.6 Dependence on the expected squared coefficients

**Proposition B.1.** *For every $c > 2$, there is a sequence of centered, normalized orthogonal designs with $n = p + 1$, deterministic mean vectors, and IID $\mathcal{N}(0, 1)$ noise such that, writing*

$$A_{1,p} = \mathbb{E}[\hat{\beta}_{(1),p}^2], \qquad B_{1,p} = \mathbb{E}[\hat{\beta}_{(2),p}^2],$$

*we have $A_{1,p}/(pB_{1,p}) \to c$, but*

$$\mathbb{E}\big[\|\mathbf{y} - \hat{\mathbf{f}}_{1,m,p}\|^2\big] > \mathbb{E}\big[\|\mathbf{y} - \hat{\mathbf{f}}_{1,p}\|^2\big]$$

*for every $p \geq 2$ and $1 \leq m < p$.*

*Proof.* First suppose a fixed response has squared OLS coefficients $A_0 \geq B_0 > 0$ and zero otherwise. For $m < p$, put $r = (p-m)/p$. The weights after one step satisfy

$$w_1^{1,m,p} = 1 - r, \qquad w_2^{1,m,p} = \frac{m}{p-1} r \leq r.$$

Since $u(2-u)$ increases on $[0,1]$, (26) gives the training error improvement

$$\begin{aligned}\Delta_m &= B_0 w_2^{1,m,p}\big(2 - w_2^{1,m,p}\big) - A_0 r^2 \\ &\leq B_0 r(2-r) - A_0 r^2 \\ &\leq -r^2\{A_0 - (2p-1)B_0\}.\end{aligned}$$

Here we used $r \geq 1/p$. Thus every $m < p$ has $\Delta_m \leq -[A_0 - (2p-1)B_0]/p^2 < 0$ when $A_0/B_0 > 2p-1$.

Fix $c > 2$ and $p \geq 2$ for the noisy construction. Take $n = p+1$ and normalized orthogonal features in the orthogonal complement of the constant vector. Let $\boldsymbol{\beta}^{(0)} = (\sqrt{cp}, 1, 0, \ldots, 0)$ and use mean vector $\mathbf{f} = \lambda \sum_{j=1}^p \beta_j^{(0)} \mathbf{x}_{\cdot,j,p}$ with IID $\mathcal{N}(0,1)$ noise. The empirical coefficients satisfy

$$\frac{\hat{\beta}_{j,p}}{\lambda} = \beta_j^{(0)} + \frac{Z_j}{\lambda\sqrt{n}},$$

where $Z_1, \ldots, Z_p$ are IID standard normal variables. Sorting squared coordinates is continuous. For $\lambda \geq 1$, their sum is bounded by $2(cp+1) + (2/n)\sum_{j=1}^p Z_j^2$, which is integrable. The absolute value of the training error difference in (26), divided by $\lambda^2$, is bounded by this same sum. Dominated convergence therefore makes each expected training error difference, divided by $\lambda^2$, tend to the strictly negative value above with $A_0 = cp$, $B_0 = 1$. It also gives $A_{1,p}/B_{1,p} \to cp$ as $\lambda \to \infty$ for this fixed $p$. There are only $p-1$ candidate sizes. Thus we may choose a deterministic $\lambda_p$ large enough that every expected difference is negative and $|A_{1,p}/B_{1,p} - cp| \leq 1/p$. This proves the proposition. Since $A_{1,p}/B_{1,p} \sim cp$, a condition $p \geq Ch(A_{1,p}/B_{1,p})$ with $h(t) = o(t)$ would hold eventually for these models, despite the absence of any improvement. □

### B.7 Weak signals

**Lemma B.11** (Weak signals)**.** *Fix $\sigma > 0$. Grant Assumption 3.1 along any integer sequence with $p \to \infty$ and $n = n_p \geq p$. Let $k = k_p \geq 1$ be integers with $k = o(p)$. For each $p$, let the mean vector be deterministic, suppose the noise is Gaussian, and write $s_p = |\{j : \beta_{j,p} \neq 0\}|$ and $\tau_p = \sigma/\sqrt{n_p}$. Assume that, for all sufficiently large $p$ with $s_p \geq 1$,*

$$\max_{1 \leq j \leq p} |\beta_{j,p}| \leq \tau_p \max\left\{1, \sqrt{2}\left(\sqrt{\log p} - \sqrt{\log s_p} - \sqrt{\log\log p}\right)\right\}.$$

*Then, for every fixed integer $r \geq 1$,*

$$\frac{\mathbb{E}[\hat{\beta}^2_{(r),p}]}{2\tau_p^2 \log p} \longrightarrow 1. \tag{32}$$

*Moreover, writing*

$$A_{k,p} = \frac{1}{k}\sum_{r=1}^{k} \mathbb{E}[\hat{\beta}^2_{(r),p}], \qquad B_{k,p} = \frac{1}{k}\sum_{r=k+1}^{2k} \mathbb{E}[\hat{\beta}^2_{(r),p}],$$

*for $p \geq 2k$, we have $A_{k,p} \sim B_{k,p} \sim 2\tau_p^2 \log(p/k)$, and hence $A_{k,p}/B_{k,p} \to 1$.*

*Proof.* By orthogonality and Gaussian noise, we can write $\hat{\beta}_{j,p} = \tau_p(Z_j + \mu_{j,p})$, where $Z_1, \ldots, Z_p$ are independent standard normal variables and $\mu_{j,p} = \beta_{j,p}/\tau_p$. Let $T_j = |Z_j + \mu_{j,p}|$, and let $T_{(1)} \geq \cdots \geq T_{(p)}$ denote their decreasing order. Write $s = s_p$, $M = \max_j |\mu_{j,p}|$, $L = \log(p/k)$, and $a = \sqrt{2L} + 1$.

We first bound the sum of the largest $k$ squared coefficients. When $s \geq 1$, the coefficient bound gives $M \leq 1 + \sqrt{2\log p} - \sqrt{2\log s}$ for all sufficiently large $p$. If $s \geq k$, then

$$\sqrt{\log p} - \sqrt{\log s} \leq \sqrt{\log(p/k)} - \sqrt{\log(s/k)},$$

so $M + \sqrt{2\log(s/k)} \le a$. Thus, for $x \ge 0$, a Gaussian tail bound gives

$$\sum_{j:\mu_{j,p}\neq 0} \mathbb{P}(T_j > a + x) \le 2se^{-(a+x-M)^2/2} \le 2ke^{-x^2/2}.$$

If $1 \le s < k$, put $d = \sqrt{2\log(k/s)}$. Since $\sqrt{\log p} - \sqrt{\log s} \le \sqrt{\log(p/s)}$, we have $M \le a + d$. Using $u_+ = \max\{u, 0\}$, we obtain

$$\begin{aligned}\sum_{j:\mu_{j,p}\neq 0} \mathbb{P}(T_j > a + x) &\le 2se^{-(x-d)_+^2/2} \\ &= 2ke^{-(d^2+(x-d)_+^2)/2} \le 2ke^{-x^2/4},\end{aligned}$$

where the last inequality follows from $d^2 + (x - d)_+^2 \ge x^2/2$. The zero coefficients contribute at most $2pe^{-(a+x)^2/2} \le 2ke^{-x^2/2}$. Consequently, in all cases, including $s = 0$,

$$\Lambda(a + x) = \sum_{j=1}^{p} \mathbb{P}(T_j > a + x) \le 4ke^{-x^2/4}, \qquad x \ge 0.$$

Since

$$\sum_{r=1}^{k} T_{(r)}^2 \le ka^2 + \sum_{j=1}^{p} (T_j^2 - a^2)_+,$$

integrating the tail bound gives

$$\begin{aligned}\frac{A_{k,p}}{\tau_p^2} &\le a^2 + \frac{1}{k}\int_0^\infty 2(a + x)\Lambda(a + x)\, dx \\ &\le a^2 + 8a\sqrt{\pi} + 16 = 2L + O(\sqrt{L} + 1),\end{aligned}$$

uniformly over $s$ and the signal coefficients.

For the lower bound, let $V_{1,p} \ge \cdots \ge V_{p,p}$ denote the decreasing order of $Z_1^2, \ldots, Z_p^2$. Let $Z$ be standard normal with distribution function $\Phi$. For every real $\mu$, the magnitude $|Z + \mu|$ stochastically dominates $|Z|$. Indeed, for $t, \mu \ge 0$, the derivative in $\mu$ of $\Phi(t - \mu) + \Phi(t + \mu) - 1$ is nonpositive. Independence and a coordinatewise quantile coupling therefore give

$$\mathbb{E}[T_{(r)}^2] \ge \mathbb{E}[V_{r,p}], \qquad 1 \le r \le p.$$

Fix $0 < \epsilon < 1$ and put $b = \sqrt{2(1 - \epsilon)L}$. The number $N_b$ of indices with $|Z_j| > b$ is binomial with mean $\lambda = p\mathbb{P}(|Z| > b)$. For $b \ge 1$, integrating the normal density over $[b, b + 1/b]$ gives $\mathbb{P}(|Z| > b) \ge cb^{-1}e^{-b^2/2}$ for an absolute constant $c > 0$. Therefore

$$\frac{\lambda}{k} \ge \frac{ce^{\epsilon L}}{\sqrt{2(1 - \epsilon)L}} \longrightarrow \infty.$$

As $k \ge 1$, we also have $\lambda \to \infty$. The Chebyshev inequality and $\mathrm{Var}(N_b) \le \lambda$ now give $\mathbb{P}(N_b \ge 2k) \to 1$. Since $2k \le p$ eventually,

$$\frac{B_{k,p}}{\tau_p^2} \ge \mathbb{E}[T_{(2k)}^2] \ge \mathbb{E}[V_{2k,p}] \ge b^2\mathbb{P}(N_b \ge 2k) = 2(1 - \epsilon)L(1 - o(1)).$$

Combining the bounds with $B_{k,p} \le A_{k,p}$, using $L \to \infty$, and then letting $\epsilon \downarrow 0$ gives $A_{k,p} \sim B_{k,p} \sim 2\tau_p^2 \log(p/k)$, and hence $A_{k,p}/B_{k,p} \to 1$. For each fixed $r$, taking $k = r$ in the bounds and using

$$T_{(2r)}^2 \le T_{(r)}^2 \le \frac{1}{r}\sum_{j=1}^{r} T_{(j)}^2$$

also proves (32), since $\log(p/r)/\log p \to 1$.

For $p \ge 2k$, Gaussian noise with $\tau_p > 0$ gives $B_{k,p} > 0$. Since $A_{k,p}/B_{k,p} \to 1$, the condition $p \ge 10A_{k,p}/B_{k,p}$ in Theorem 4.1 holds for all sufficiently large $p$. Lemma 4.1 gives concavity under the same coefficient bound, including when every projected signal coefficient is zero. Corollary 4.1 therefore gives strictly smaller expected prediction error for the same deterministic choice $m_p^\star$. $\square$

### B.8 An interior minimum of expected training error

Theorem 4.1 establishes improvement over $m = p$, but does not by itself exclude a minimum at $m = 1$. Set $p = n = 3$, $k = 1$, and take an orthogonal design with mean coefficients $(2, 1, 1/\sqrt{10})$. At zero noise, the rank weights for $m = 1, 2, 3$ are respectively $(1/3, 1/3, 1/3)$, $(2/3, 1/3, 0)$, and $(1, 0, 0)$. Thus the respective training errors are

$$\frac{34}{15}, \qquad \frac{89}{90}, \qquad \frac{11}{10},$$

so $m = 2$ is a strict interior minimum. This example also holds with positive Gaussian noise variance sufficiently small. Indeed, couple the responses as $\mathbf{y}_\sigma = \mathbf{f} + \sigma\mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3)$. The strict ordering of the mean coefficient magnitudes makes each fitted vector continuous at $\sigma = 0$. Moreover, for $0 \le \sigma \le 1$, orthogonality and weights in $[0, 1]$ give $\|\mathbf{y}_\sigma - \hat{\mathbf{f}}_{1,m,3}\|^2 \le \|\mathbf{y}_\sigma\|^2 \le 2\|\mathbf{f}\|^2 + 2\|\mathbf{z}\|^2$. Dominated convergence therefore preserves both strict inequalities in expectation for all sufficiently small $\sigma > 0$.

### B.9 Proof of Lemma 4.1

*Proof.* Write $\tau = \sigma/\sqrt{n}$. In this proof, feature indices refer to the fixed, original dictionary, before the ordering determined by the response. Set $\mu_j = \beta_{j,p}/\tau$ and $X_j = \hat{\beta}_{j,p}/\tau$. Orthogonality and Gaussian noise imply that the $X_j$ are independent $\mathcal{N}(\mu_j, 1)$ variables. Let $T_j = |X_j|$, and write $I_k = \pi(k)$ for the original index of the $k$th largest magnitude, so that $T_{I_1} > \cdots > T_{I_p}$ almost surely. The covariance definition of degrees of freedom gives

$$\Delta_k = \mathrm{df}(\hat{f}_{k,p}) - \mathrm{df}(\hat{f}_{k-1,p}) = \mathbb{E}[X_{I_k}(X_{I_k} - \mu_{I_k})].$$

The sign of $X_j$, conditionally on $T_j = t$, satisfies $\mathbb{E}[X_j \mid T_j = t] = t\tanh(\mu_j t)$. Consequently,

$$\Delta_k = \sum_{j=1}^{p} \mathbb{E}[h_j(T_j)\,\mathbf{1}(I_k = j)], \qquad h_j(t) = t^2 - \mu_j t\tanh(\mu_j t).$$

Here and below $h_j$ depends only on $|\mu_j|$. The density of $T_j$ on $(0, \infty)$ is

$$f_j(t) = \sqrt{\frac{2}{\pi}}\exp\left(-\frac{t^2 + \mu_j^2}{2}\right)\cosh(\mu_j t), \qquad h_j(t) = -t\,\frac{f_j'(t)}{f_j(t)}.$$

When $p = 2$, integration by parts gives, for $t \ge 0$,

$$\mathbb{E}[X_j(X_j - \mu_j)\mathbf{1}(T_j > t)] = -\int_t^\infty u f_j'(u)\,du = \mathbb{P}(T_j > t) + t f_j(t).$$

Conditioning on the other magnitude and summing over $j = 1, 2$ therefore gives

$$\mathrm{df}(\hat{f}_{1,2}) = 1 + 2\int_0^\infty t f_1(t) f_2(t)\,dt > 1.$$

Since $\mathrm{df}(\hat{f}_{0,2}) = 0$ and $\mathrm{df}(\hat{f}_{2,2}) = 2$, this proves concavity for $p = 2$ without any restriction on the means. For the rest of the proof, assume $p \ge 3$.

We first establish an inequality in two variables. Fix $a, b \ge 0$ and define

$$H(t, u) = \cosh(at)\cosh(bu) + \cosh(bt)\cosh(au), \qquad L(t, u) = \log H(t, u).$$

For $t \ge u \ge 0$, we claim that

$$tL_t(t, u) - uL_u(t, u) \le \frac{a^2 + b^2}{2}(t^2 - u^2). \tag{33}$$

To see this, put $v = (t + u)/2$, $r = (t - u)/2$, $A = a + b$, and $B = |a - b|$. The addition formula for the hyperbolic cosine gives

$$\widetilde{H}(v, r) = H(v + r, v - r) = \cosh(Av)\cosh(Br) + \cosh(Bv)\cosh(Ar).$$

Let $\omega = \cosh(Av)\cosh(Br)/\widetilde{H}(v,r) \in [0,1]$. Since $\tanh(z) \le z$ for $z \ge 0$ and $v, r, A, B \ge 0$,

$$\begin{aligned} tL_t - uL_u &= r(\log \widetilde{H})_v + v(\log \widetilde{H})_r \\ &= \omega\{rA\tanh(Av) + vB\tanh(Br)\} \\ &\quad + (1-\omega)\{rB\tanh(Bv) + vA\tanh(Ar)\} \\ &\le vr(A^2+B^2) = \frac{a^2+b^2}{2}(t^2-u^2). \end{aligned}$$

This proves (33).

Now fix $k \in \{1, \dots, p-1\}$ and distinct labels $i, j$. Given $t > u > 0$, let $Q_{ij,k}(t,u)$ be the probability that exactly $k-1$ of the other $p-2$ magnitudes exceed $t$ and all the remaining $p-k-1$ magnitudes are below $u$. Thus $Q_{ij,k} \ge 0$, and it is the same whether label $i$ takes magnitude $t$ or magnitude $u$. The contribution of this unordered pair to $\Delta_k - \Delta_{k+1}$ is the integral over $t > u > 0$ of $Q_{ij,k}(t,u)$ times

$$\begin{aligned} S_{ij}(t,u) &= [h_i(t) - h_j(u)]f_i(t)f_j(u) \\ &\quad + [h_j(t) - h_i(u)]f_j(t)f_i(u). \end{aligned}$$

Writing $F_{ij} = f_i(t)f_j(u) + f_j(t)f_i(u)$, the identity for $h_j$ implies

$$S_{ij} = (-t\partial_t + u\partial_u)F_{ij}.$$

With $a = |\mu_i|$ and $b = |\mu_j|$, $F_{ij}$ is a positive constant times $\exp(-(t^2+u^2)/2)H(t,u)$. Hence

$$\begin{aligned} S_{ij}(t,u) &= F_{ij}(t,u)\{t^2 - u^2 - tL_t(t,u) + uL_u(t,u)\} \\ &\ge F_{ij}(t,u)\left(1 - \frac{\mu_i^2+\mu_j^2}{2}\right)(t^2-u^2). \end{aligned}$$

Summing over unordered pairs and integrating against the nonnegative $Q_{ij,k}$ gives $\Delta_k \ge \Delta_{k+1}$ whenever $\max_j |\mu_j| \le 1$. The same argument applies whenever the sum of the two largest squared mean magnitudes is at most 2.

Write $M = \max_j |\mu_j|$, $s = |\{j : \mu_j \ne 0\}|$, and $N = p - s$. The preceding argument covers $M \le 1$, including $s = 0$, and also $M \le \sqrt{2}$ when $s = 1$. Under (13), this proves the result for $3 \le p \le e^5$ at every sparsity, and for $3 \le p \le e^9$ when $s \ge 2$. Indeed, the function $L \mapsto \sqrt{2L} - \sqrt{2\log L}$ decreases and then increases on $L > 1$, since its derivative has the sign of $L\log L - 1$. Its value at $L = 1$ is $\sqrt{2}$, and

$$\sqrt{10} - \sqrt{2\log 5} < \sqrt{2}, \qquad \sqrt{18} - \sqrt{2\log 9} < 1 + \sqrt{2\log 2}.$$

Checking the endpoints on $[1,5]$ and $[1,9]$ therefore gives the two claims. For the remaining cases, $p > e^5$ and $M > 1$. For mean magnitude $v \ge 0$, write the folded normal density and distribution functions as

$$f(t;v) = \phi(t-v) + \phi(t+v), \qquad F(t;v) = \mathbb{P}(|Z+v| \le t), \qquad Q(t;v) = 1 - F(t;v),$$

where $Z$ is standard normal and $\phi, \Phi$ are its density and distribution function. Put $\rho_v(t) = vt\tanh(vt)$ and $\overline{\Phi} = 1 - \Phi$. With $G_k = T_{I_k}^2 - T_{I_{k+1}}^2$ and $A_k = \mathbb{E}[G_k]$, we have

$$\Delta_k - \Delta_{k+1} = A_k - C_k, \qquad C_k = \mathbb{E}\big[\rho_{|\mu_{I_k}|}(T_{I_k}) - \rho_{|\mu_{I_{k+1}}|}(T_{I_{k+1}})\big]. \tag{34}$$

The pairwise argument bounds this bias difference by

$$C_k \le \frac{1}{2}\mathbb{E}\big[G_k(\mu_{I_k}^2 + \mu_{I_{k+1}}^2)\big]. \tag{35}$$

The same bound holds after restricting both expectations to any region that depends only on the two adjacent magnitudes.

We first compare $A_k$ with $A_k^{(0)}(p)$, the expected gap between the $k$th and $(k+1)$st largest squared magnitudes when all means are zero. The density ratio of a noncentral squared normal to a central

one is $e^{-\mu^2/2}\cosh(\mu\sqrt{x})$, which increases in $x$. Its survival hazard is therefore at most the central $\chi^2_1$ hazard $h_0(x)$. Moreover,

$$h_0'(x) = h_0(x)\left\{h_0(x) - \tfrac{1}{2} - \tfrac{1}{2x}\right\} \le 0$$

by $\overline{\Phi}(t) \ge \phi(t)t/(1+t^2)$. Conditional on the $(k+1)$st largest squared value being $v$ and on the identities of the $k$ variables above it, these variables are independent truncations. Their minimum excess over $v$ has expectation at least

$$m_k(v) = \int_0^\infty \left\{\frac{Q(\sqrt{v+x};0)}{Q(\sqrt{v};0)}\right\}^k dx.$$

Decreasing central hazard makes $m_k(v)$ increasing in $v$. Coordinatewise stochastic domination gives $A_k \ge A_k^{(0)}(p)$. The same argument shows that $A_k^{(0)}(p)$ increases with $p$ for fixed $k$.

For a central folded normal, $U = F(T;0)$ is uniform and $f(t;0) \le \sqrt{2/\pi}$. Thus the inverse distribution function satisfies

$$[F^{-1}(u;0)]^2 - [F^{-1}(v;0)]^2 \ge \frac{\pi}{2}(u^2 - v^2), \qquad 0 \le v \le u < 1.$$

The smallest expected difference of squares of adjacent uniform order statistics is $4/((p+1)(p+2))$. Using monotonicity in $p$ also gives

$$A_k \ge \frac{2\pi}{(p+1)(p+2)} > p^{-2}, \qquad A_k \ge \frac{2\pi}{(k+2)(k+3)}. \tag{36}$$

For the first rank, $A_1^{(0)}(p) \ge A_1^{(0)}(2) = \mathbb{E}[|Z_1^2 - Z_2^2|] = 4/\pi$ for independent standard normals $Z_1, Z_2$, since $Z_1^2 - Z_2^2 = (Z_1 - Z_2)(Z_1 + Z_2)$ and the two factors are independent $\mathcal{N}(0,2)$ variables.

Suppose first that $s = 1$. Let $\mathbb{E}_0$ denote expectation under $p$ independent central folded normals, with ordered magnitudes $T_{(1)}^{(0)} \ge \cdots \ge T_{(p)}^{(0)}$. Put

$$r(t) = \frac{f(t;M)}{f(t;0)} = e^{-M^2/2}\cosh(Mt), \qquad J(t) = \rho_M(t)r(t) = e^{-M^2/2}Mt\sinh(Mt).$$

Multiplying the central density by $r$ for the signal label and using symmetry gives

$$C_k = \frac{1}{p}\mathbb{E}_0\big[J(T_{(k)}^{(0)}) - J(T_{(k+1)}^{(0)})\big].$$

For $P_k(t) = \binom{p}{k}Q(t;0)^k F(t;0)^{p-k}$, integration over the interval between adjacent magnitudes yields

$$C_k = \frac{1}{p}\int_0^\infty J'(t)P_k(t)\,dt, \qquad A_k^{(0)}(p) = \int_0^\infty 2tP_k(t)\,dt.$$

The function $J'(t)/(2t)$ increases, as follows from its power series. The probability densities proportional to $2tP_k(t)$ have successive likelihood ratios proportional to $Q(t;0)/F(t;0)$, which decrease. Since the covariance of an increasing and a decreasing function is nonpositive, we obtain

$$\frac{C_k}{A_k^{(0)}(p)} \le \frac{C_1}{A_1^{(0)}(p)}, \qquad 1 \le k < p. \tag{37}$$

It therefore suffices to bound the bias at rank one.

Write

$$L = \log p, \quad \ell = \log L, \quad a = \sqrt{2L}, \quad h = \sqrt{2\ell},$$
$$M_0 = a - h, \quad t_0 = \sqrt{2L - \ell}, \quad d_0 = t_0 - M_0.$$

Here $L > 5$ and $1 < M \le M_0 < t_0$. Since $P_1(t) = pQ(t;0)F(t;0)^{p-1}$,

$$C_1 = \int_0^\infty J'(t)Q(t;0)F(t;0)^{p-1}\,dt.$$

For $t \le t_0$, $J'(t)/(2t) \le M^2 r(t_0)$. This part of the integral is at most $cA_1^{(0)}(p)$, where

$$c = \frac{M^2 r(t_0)}{p} \le (1 + e^{-2t_0}) \frac{M_0^2}{2L} e^{-M_0(a-t_0)}.$$

To see the inequality, use $M > 1$ in $r(t_0) = (1+e^{-2Mt_0})e^{Mt_0 - M^2/2}/2$ and note that $M^2 e^{Mt_0 - M^2/2}$ increases for $M < t_0$. Put

$$x = \sqrt{\ell/L}, \qquad y = \sqrt{1 - x^2/2}, \qquad z = 1 - x, \qquad v = 1 - \frac{x}{2(1+y)}.$$

Then $M_0 = az$, $t_0 = ay$, and $d_0 = hv$. For $L \ge 5$,

$$x < \frac{3}{5}, \quad y > \frac{9}{10}, \quad v > \frac{5}{6}, \quad \ell \ge 3z, \quad \frac{z}{\sqrt{\ell}} < \frac{2}{5}, \quad e^{\ell(1-v^2)} < 2.$$

These bounds require only elementary calculus. Indeed, $x = \sqrt{\ell} e^{-\ell/2}$ decreases for $\ell > 1$. It exceeds $1/2$ on $[3/2, 2]$ and $1/3$ on $[2, 3]$, proving $\ell \ge 3z$ there, and the case $\ell \ge 3$ is immediate. The function $\ell^{-1/2} - e^{-\ell/2}$ is less than $2/5$ at $\ell = 1$ and tends to zero at infinity. At a stationary point its value is $(\ell - 1)/\ell^{3/2} \le 2/(3\sqrt{3}) < 2/5$. Finally, $1 - v^2 \le x/2$ gives $\ell(1 - v^2) \le 3\sqrt{3}/(2e^{3/2}) < \log 2$. Consequently,

$$c < \frac{5}{4} z^2 e^{-\ell z/2} \le \frac{5}{4} z^2 e^{-3z^2/2} \le \frac{5}{6e} < \frac{1}{3}.$$

For $t > t_0$, Mills' bound $Q(t;0) \le f(t;0)/t$ gives $J'(t)Q(t;0) \le (M^2 + M/t_0) f(t;M)$. Since $Q(t_0;M)$ increases with $M$, this part of the integral is at most

$$\begin{aligned} U &= (M_0^2 + M_0/t_0) Q(t_0; M_0) \\ &\le (M_0^2 + M_0/t_0)(1 + e^{-2t_0}) \frac{\phi(d_0)}{d_0} \\ &< \frac{5}{4} \frac{z}{\sqrt{\ell}} \frac{e^{\ell(1-v^2)}}{\sqrt{\pi} v} \left( z + \frac{1}{2Ly} \right) \\ &< \frac{5}{4} \frac{2}{5} \frac{2}{\sqrt{\pi}(5/6)} \left( 1 + \frac{1}{9} \right) = \frac{4}{3\sqrt{\pi}}. \end{aligned}$$

Here the reflected tail is at most $e^{-2M_0 t_0}$ times the other tail, and $M_0 > 1$. Therefore, combining the two integration regions,

$$\frac{C_1}{A_1^{(0)}(p)} < \frac{1}{3} + \frac{\sqrt{\pi}}{3} < 1.$$

Together with (37) and $A_k \ge A_k^{(0)}(p)$, this proves concavity when $s = 1$.

Now suppose $s \ge 2$. The elementary cases above allow us to assume $L = \log p > 9$ and $M > 1$. Put

$$a = \sqrt{2L}, \qquad h = \sqrt{2 \log L}, \qquad b = \sqrt{2 \log s}, \qquad d = h + b.$$

The coefficient bound gives $M \le a - d$. In particular,

$$a > 4, \qquad h > 2, \qquad b > \frac{7}{6}, \qquad d > 3, \qquad e^{-bh} < \frac{1}{10}.$$

Since $b < a - 3$, we have $s/p < e^{-a}$. Hence $N - 2 > 9p/10$, and Mills' bound gives $F(M;M) > 9/20$.

We first bound the probability of a signal label at either endpoint. Conditional on adjacent magnitudes $t > u > 0$,

$$\mathbb{E}\big[\mathbf{1}(\mu_{I_k} \ne 0) + \mathbf{1}(\mu_{I_{k+1}} \ne 0) \mid t, u\big] \le \frac{2s}{N-2} \max\{1, R_M(t,u)\},$$

$$R_M(t,u) = \max_{0 < v \le M} \frac{f(t;v)}{f(t;0)} \frac{F(u;0)}{F(u;v)}.$$

To see this, swap a signal endpoint with a null label outside the pair. If that null is above $t$, the ratio of the original conditional density to that after the exchange is at most one by the increasing likelihood

ratio $f(t;v)/f(t;0)$. If it is below $u$, the ratio is at most $R_M(t,u)$. Each signal endpoint has at least $N-2$ outgoing swaps, and each resulting configuration has at most $2s$ incoming swaps. Summing their conditional weights proves the bound.

Split adjacent magnitudes $t \geq u$ into three regions,

$$\begin{aligned}\mathcal{R}_1 &= \{t \leq a,\ (u \geq M \text{ or } t-u \leq 1/M)\},\\ \mathcal{R}_2 &= \{t \leq a,\ u < M,\ t-u > 1/M\},\\ \mathcal{R}_3 &= \{t > a\}.\end{aligned}$$

For $u \geq M$ and $t \leq a$, $R_M(t,u) \leq e^{Ma-M^2/2}/F(M;M)$. For $u < M$ and $t-u \leq 1/M$, use

$$\partial_t \log \frac{f(t;v)}{f(t;0)} \leq v, \qquad \frac{f(u;v)}{F(u;v)} \leq v + 1/u, \qquad \frac{F(u;0)}{f(u;0)} \leq u e^{u^2/2}.$$

The middle inequality follows by integrating $\partial_t \log f(t;v) \leq v$. These bounds give

$$R_M(t,u) \leq e(1+M^2)e^{M^2/2} \leq e^{Ma-M^2/2},$$

because $1+\log(1+M^2) \leq 2M < M(a-M)$ for $M \geq 1$. By (35), the contribution of $\mathcal{R}_1$ to $C_k$ is therefore at most

$$\frac{M^2 s}{N-2}\frac{e^{Ma-M^2/2}}{F(M;M)}A_k < \frac{40}{81}A_k < \frac{1}{2}A_k.$$

Here we used $M^2 < 2L$ and $(s/p)e^{Ma-M^2/2} \leq e^{-bh}/L$.

An interval of length $1/M$ starting below $M$ has central folded normal mass at least

$$j(M) = \frac{\sqrt{2/\pi}}{M}e^{-(M+1/(2M))^2/2} > \frac{3}{8M}e^{-M^2/2}.$$

Indeed, its smallest mass starts at $M$, and the Jensen inequality applies because $\phi$ is convex on $[1,\infty)$. On $\mathcal{R}_2$, no null lies in this interval just above $u$. Conditioning on the label of $u$ and taking a union bound gives $\mathbb{P}(\mathcal{R}_2) \leq pe^{-K}$, where

$$K = (N-1)j(M) > \frac{27}{80a}e^{3a/2} > 2a^2 = 4L.$$

The last inequality holds because $e^{3a/2}/a^3$ increases for $a \geq 4$ and $e^6 > 400$. The exact bias on this region is at most $Ma < 2L$. Using (36), its contribution is less than

$$2Lpe^{-K} < 2Le^{-L}A_k < \frac{1}{100}A_k.$$

For the last region, put

$$D = 2\phi(h)e^{-bh} < \frac{1}{10L}, \qquad \Lambda = \frac{2}{\sqrt{2\pi}a} + \frac{D}{d} < \frac{1}{4} + \frac{1}{270} < 1.$$

For $w \geq 0$, the bounds $f(a+w;v) \leq 2\phi(d+w)$ and $Q(a+w;v) \leq 2\overline{\Phi}(d+w)$ for $v \leq M$ imply

$$\sum_{j:\mu_j \neq 0} f(a+w;|\mu_j|) \leq De^{-dw}, \qquad \lambda(a+w) = \sum_{j=1}^{p} Q(a+w;|\mu_j|) \leq \Lambda e^{-dw}.$$

For the first inequality, use $s\phi(h+b+w) = \phi(h)e^{-bh-dw-w^2/2}$. For the second, apply Mills' bound to the signals and use $p\phi(a) = 1/\sqrt{2\pi}$ and $a > d$ for the nulls. The exact bias on $\mathcal{R}_3$ is at most $MT_{I_k}\mathbf{1}(\mu_{I_k} \neq 0)$. Independence and the bound using factorial moments for the number of other magnitudes above $t$ therefore bound its expectation by

$$\begin{aligned}\frac{MD\Lambda^{k-1}}{(k-1)!}\int_0^\infty (a+w)e^{-kdw}\,dw &\leq MD\left(\frac{a}{d}+\frac{1}{d^2}\right)\frac{\Lambda^{k-1}}{k!}\\ &< \frac{13}{180k!} < \frac{1}{12k!} < \frac{1}{6}A_k.\end{aligned}$$

The last inequality follows from (36) and $(k+2)(k+3) \leq 12k!$ for every $k \geq 1$. The three regional bounds sum to less than $A_k$, so (34) proves concavity for every $1 \leq k < p$.

□

### B.10 Proofs for Section 4.2

*Proof of Theorem 4.2.* For every response, the representation in terms of rank weights gives

$$\hat{f}_{k,m,p} = \sum_{j=1}^{p} w_j^{k,m,p} (\hat{f}_{j,p} - \hat{f}_{j-1,p}).$$

The weights are deterministic, so linearity of covariance proves (11).

We claim that whenever $m \le m'$, the sequence $\{w_j^{k,m',p}\}_{j=1}^{p}$ majorizes the sequence $\{w_j^{k,m,p}\}_{j=1}^{p}$. This means that

$$\sum_{j=1}^{l} w_j^{k,m',p} \ge \sum_{j=1}^{l} w_j^{k,m,p}, \qquad 1 \le l \le p, \tag{38}$$

with equality when $l = p$. Here all indices denote ranks. Let $S_{r,m}$ denote the ranks of the features selected by one randomized FS path after $r$ steps, and put $Z_{r,m}(l) = |S_{r,m} \cap [l]|$. Then $\sum_{j=1}^{l} w_j^{k,m,p} = \mathbb{E}[Z_{k,m}(l)]$.

We couple the randomized FS paths with parameters $m$ and $m'$ so that $Z_{r,m'}(l) \ge Z_{r,m}(l)$ at every step $r$ and every cutoff $l$. At step $r+1$, there are $p-r$ remaining features in each path. Set $a = \min\{m, p-r\}$ and $a' = \min\{m', p-r\}$. Draw a uniform subset $B'$ of $[p-r]$ with $a'$ elements, then a uniform subset $B \subseteq B'$ with $a$ elements. Both marginal subsets are uniform of their respective sizes. Use their elements as positions in the increasing lists of remaining ranks for the two paths. This construction also implements the cap on the candidate size when fewer than $m$ or $m'$ features remain. Write $b = \min B$ and $b' = \min B'$, so $b' \le b$. The selected rank is the $b$th remaining rank on the first path, and therefore

$$Z_{r+1,m}(l) = Z_{r,m}(l) + \mathbf{1}(b \le l - Z_{r,m}(l)),$$

with the analogous identity for $m'$ and $b'$.

We now show the desired property inductively in $r$, with the base case $r = 0$ trivial. Assume the inductive hypothesis $Z_{r,m'}(l) \ge Z_{r,m}(l)$ for every $l$. If the inequality is strict at a given $l$, its gap is a positive integer, so adding at most one to either count preserves the ordering. If the counts agree, the numbers $l - Z_{r,m'}(l)$ and $l - Z_{r,m}(l)$ of remaining ranks up to the cutoff agree, and $b' \le b$ ensures that the increment along the $m'$ path is at least that along the $m$ path. Thus $Z_{r+1,m'}(l) \ge Z_{r+1,m}(l)$ simultaneously for all $l$. Taking expectations at step $k$ proves (38). The total weights of both paths equal $k$.

Finally, put $D_j = \mathrm{df}(\hat{f}_{j,p})$, $\Delta_j = D_j - D_{j-1}$, and $A_l = \sum_{j=1}^{l} (w_j^{k,m',p} - w_j^{k,m,p})$. By the preceding comparison, $A_l \ge 0$ and $A_p = 0$. Concavity of $D_j$ means $\Delta_l \ge \Delta_{l+1}$. Summation by parts now gives

$$\begin{aligned} \mathrm{df}(\hat{f}_{k,m',p}) - \mathrm{df}(\hat{f}_{k,m,p}) &= \sum_{j=1}^{p} \Delta_j (w_j^{k,m',p} - w_j^{k,m,p}) \\ &= \sum_{l=1}^{p-1} (\Delta_l - \Delta_{l+1}) A_l \ge 0. \end{aligned}$$

No sign assumption on the increments $\Delta_j$ is needed. □

*Proof of Corollary 4.1.* The degrees of freedom comparison follows from Theorem 4.2, since $m_p^\star < p$. Combining it with (10) in (2) gives the strict prediction error improvement. □

*Proof of Proposition 4.1.* For the fixed original ordering of the features, let $\tau = \sigma/\sqrt{n}$. Throughout this proof, $k, \sigma, \gamma$ are fixed, but $n = n_p$ and $m = m_p$ may follow any sequences allowed by the proposition. Since we consider the limit as $p$ goes to infinity, we may assume $p > k$. By orthogonality and Gaussian noise, we have independent standard normal variables $Z_1, \ldots, Z_p$ such that

$$\hat{\beta}_{i,p} = \beta_{i,p} + \tau Z_i, \qquad \beta_{i,p} = \mathbf{1}(i \le k).$$

Let $I_r = \pi(r)$ be the original label of the coefficient at rank $r$, and define

$$Q_i = \tau^{-2}\hat{\beta}_{i,p}(\hat{\beta}_{i,p} - \beta_{i,p}) = Z_i^2 + \tau^{-1}\beta_{i,p}Z_i.$$

Writing $w_r = w_r^{k,m,p}$, by linearity of covariance, we can then write

$$D_k \coloneqq \mathrm{df}(\hat{f}_{k,p}) = \mathbb{E}\left[\sum_{r=1}^{k} Q_{I_r}\right], \qquad D_{k,m} \coloneqq \mathrm{df}(\hat{f}_{k,m,p}) = \mathbb{E}\left[\sum_{r=1}^{p} w_r Q_{I_r}\right].$$

Let $\mathcal{A}$ be the event that every signal coefficient is positive and its magnitude exceeds every null coefficient magnitude. For $i \le k < j$, both $\hat{\beta}_{i,p} - \hat{\beta}_{j,p}$ and $\hat{\beta}_{i,p} + \hat{\beta}_{j,p}$ have distribution $\mathcal{N}(1, 2\tau^2)$. Since

$$\{|\hat{\beta}_{j,p}| > |\hat{\beta}_{i,p}|\} \subseteq \{\hat{\beta}_{j,p} > \hat{\beta}_{i,p}\} \cup \{-\hat{\beta}_{j,p} > \hat{\beta}_{i,p}\},$$

using Gaussian tail bounds and applying a union bound, we get

$$\delta_p \coloneqq \mathbb{P}(\mathcal{A}^c) \le 2k(p-k)e^{-n/(4\sigma^2)} + ke^{-n/(2\sigma^2)}.$$

To control the contribution from $\mathcal{A}^c$, let $R_p = \sum_{i=1}^{p} |Q_i|$. By the triangle inequality in $L^2$, we have

$$\begin{aligned}(\mathbb{E}[R_p^2])^{1/2} &\le \left\{\mathbb{E}\left[\left(\sum_{i=1}^{p} Z_i^2\right)^2\right]\right\}^{1/2} + \tau^{-1}\left\{\mathbb{E}\left[\left(\sum_{i=1}^{k} |Z_i|\right)^2\right]\right\}^{1/2} \\ &\le \sqrt{p(p+2)} + k\sqrt{n}/\sigma =: B_p.\end{aligned}$$

Applying Cauchy–Schwarz, any two random variables bounded in absolute value by $R_p$ and equal on $\mathcal{A}$ have expectations differing by at most $\eta_p \coloneqq 2B_p\sqrt{\delta_p}$. Using $p \le n$, we get $\eta_p \le C_{k,\sigma} n^{3/2} e^{-n/(8\sigma^2)}$, and hence $\sup_{n \ge p} \eta_p \to 0$, uniformly over $m$. In particular,

$$|D_k - k| \le \eta_p, \tag{39}$$

because $\sum_{r=1}^{k} Q_{I_r} = \sum_{i=1}^{k} Q_i$ on $\mathcal{A}$ and $\mathbb{E}\big[\sum_{i=1}^{k} Q_i\big] = k$.

First suppose $k \ge 2$. Write $Z_{[1]} \ge \cdots \ge Z_{[k]}$ for the decreasing order of the signal noises $Z_1, \ldots, Z_k$, and write $U_{1,p} \ge \cdots \ge U_{p-k,p}$ for the decreasing order of the null squares $Z_{k+1}^2, \ldots, Z_p^2$. On $\mathcal{A}$, the first $k$ ranks are ordered by the $Z_{[r]}$, and therefore

$$\sum_{r=1}^{p} w_r Q_{I_r} = \tau^{-1}\sum_{r=1}^{k} w_r Z_{[r]} + \sum_{r=1}^{k} w_r Z_{[r]}^2 + \sum_{r=1}^{p-k} w_{k+r} U_{r,p}.$$

Notice that both sides are bounded in absolute value by $R_p$ on the whole probability space, since each is a sum of the same $Q_i$ with weights in $[0,1]$, possibly in a different order. Hence,

$$D_{k,m} = \frac{\sqrt{n}}{\sigma} a_{k,p} + b_{k,p} + T_{k,p} + O(\eta_p), \tag{40}$$

where

$$a_{k,p} = \sum_{r=1}^{k} w_r \mathbb{E}[Z_{[r]}], \qquad b_{k,p} = \sum_{r=1}^{k} w_r \mathbb{E}[Z_{[r]}^2], \qquad T_{k,p} = \sum_{r=1}^{p-k} w_{k+r}\mathbb{E}[U_{r,p}].$$

We have $0 \le b_{k,p} \le k$. Since the weights sum to $k$, Lemma B.12 also gives

$$0 \le T_{k,p} \le k\mathbb{E}\big[\max_{k < i \le p} Z_i^2\big] \le k\{2\log(2(p-k)) + 2\} = O_k(\log p).$$

Theorem 3.1 implies $a_{k,p} \to a_{k,\gamma} \coloneqq \sum_{r=1}^{k} \tilde{w}_r^{k,\gamma}\mathbb{E}[Z_{[r]}]$.

We now show that this limit is strictly positive. Let $q = 1 - \gamma \in (0,1)$ and $u_j^r = \tilde{w}_j^{r,\gamma}$. In the infinite rank process from the proof of Lemma B.8, choosing $G > j$ for the first $k-1$ steps and $G = j$ next selects rank $j$ for the first time, with probability $\gamma q^{jk-1} > 0$. Thus $u_j^k - u_j^{k-1} > 0$ for every $j$. Combining (18) and (19) gives

$$u_j^k - u_{j+1}^k = (1-q^k)(u_j^k - u_j^{k-1}) > 0.$$

Since $\sum_{r=1}^{k} \mathbb{E}[Z_{[r]}] = 0$, we have

$$a_{k,\gamma} = \frac{1}{k} \sum_{1 \le i < j \le k} (u_i^k - u_j^k)\mathbb{E}[Z_{[i]} - Z_{[j]}] > 0.$$

The last inequality follows because the ordered Gaussian noises are distinct almost surely.

Putting (39) and (40) together and using $n \ge p$, we get

$$\frac{\sigma}{\sqrt{n}}(D_{k,m} - D_k) \longrightarrow a_{k,\gamma} > 0.$$

Since $\sqrt{n}/\log p \to \infty$, the claimed logarithmically normalized limit is $+\infty$ for $k \ge 2$.

Now suppose $k = 1$. Let $S$ be the uniform candidate set of size $m$, drawn independently of the response, and let $J(S)$ denote the selected feature. By linearity of covariance, $D_{1,m} = \mathbb{E}_{S,Z}[Q_{J(S)}]$. On $\mathcal{A}$, we have $Q_{J(S)} = Q_1$ if $1 \in S$, and $Q_{J(S)} = \max_{i \in S} Z_i^2$ otherwise. By a similar calculation using $R_p$, the error in this replacement is at most $\eta_p$. Conditional on $1 \notin S$, the $m$ sampled noises are IID standard normal. Thus, writing $M_m$ for the maximum of $m$ IID $\chi_1^2$ variables,

$$D_{1,m} = \frac{m}{p} + \left(1 - \frac{m}{p}\right)\mathbb{E}[M_m] + O(\eta_p), \qquad D_{1,m} - D_1 = \left(1 - \frac{m}{p}\right)(\mathbb{E}[M_m] - 1) + O(\eta_p).$$

Here $\mathbb{E}[M_m]/\log m \to 2$ by Lemma B.12, and $\log m/\log p \to 1$ because $m/p \to \gamma \in (0,1)$. The ratio of the last display to $\log p$ therefore tends to $2(1-\gamma)$.

Finally, $\mathrm{df}(\hat{f}_{k,1,p}) = k$ and $\mathrm{df}(\hat{f}_{k,p,p}) = D_k = k + o(1)$. In either case, a subsampling ratio converging to a fixed interior value gives degrees of freedom larger than both endpoints for all sufficiently large $p$. This finishes the proof. □

*Remark* (A uniform comparison for one step). For $k = 1$, the comparison holds for every $2 \le m \le p-1$, once $p$ is sufficiently large depending only on $\sigma$, uniformly over $n \ge p$. Indeed, the comparison using candidate sets in the proof has error at most $2\eta_p$ in the difference of degrees of freedom, independently of $m$. Its bound satisfies

$$p\eta_p \le C_\sigma n^{5/2} e^{-n/(8\sigma^2)}, \qquad \sup_{n \ge p} p\eta_p \longrightarrow 0.$$

For independent standard normals $X, Y$,

$$\mathbb{E}[\max\{X^2, Y^2\}] = 1 + \tfrac{1}{2}\mathbb{E}[|(X-Y)(X+Y)|] = 1 + \frac{2}{\pi},$$

because $X - Y$ and $X + Y$ are independent $\mathcal{N}(0,2)$ variables, each with mean absolute value $2/\sqrt{\pi}$. Thus $\mathbb{E}[M_m] - 1 \ge 2/\pi$ for every $m \ge 2$. Since $1 - m/p \ge 1/p$ when $m \le p-1$, the same comparison gives

$$\mathrm{df}(\hat{f}_{1,m,p}) - \mathrm{df}(\hat{f}_{1,p}) \ge \frac{2}{\pi p} - 2\eta_p > 0$$

uniformly over the stated range for all sufficiently large $p$.

*Remark.* In this proposition, $k, \sigma$ are fixed and $m/p$ tends to a fixed $\gamma \in (0,1)$. Under orthogonality, if $1 \le k \le p$ and $m > p - k$, every base learner selects exactly the first $k$ ranks. Indeed, at most $p - k$ features lie outside those ranks, so every candidate set contains a remaining feature among the first $k$ ranks until all are selected. A capped candidate set contains all remaining features. Thus $\hat{f}_{k,m,p} = \hat{f}_{k,p}$ for every response in this range. In particular, $m = p - 1$ gives exact equality for $k \ge 2$, whereas the preceding uniform strict comparison concerns only $k = 1$. At $k = 0$ or $k = p$, all candidate sizes give the same estimator.

**Lemma B.12.** *Let $Z_1, Z_2, \ldots, Z_N$ be IID $\chi_1^2$ variables, and let $M_N = \max_{1 \le i \le N} Z_i$. We have*

$$\mathbb{E}[M_N] \le 2\log(2N) + 2, \qquad \frac{\mathbb{E}[M_N]}{\log N} \longrightarrow 2 \quad (N \to \infty).$$

*In particular, $\mathbb{E}[M_N] \ge \frac{1}{2}\log N$ for all sufficiently large $N$.*

*Proof.* Using the Gaussian tail bound, we have $\mathbb{P}(M_N > t) \le \min\{1, 2Ne^{-t/2}\}$ for $t \ge 0$. Therefore,

$$\mathbb{E}[M_N] = \int_0^\infty \mathbb{P}(M_N > t)\, dt \le 2\log(2N) + 2.$$

For the lower bound, let $\phi$ and $\Phi$ denote the standard Gaussian density and distribution function. Integration by parts gives $1 - \Phi(z) \ge z\phi(z)/(1+z^2)$ for $z > 0$. Fix $\epsilon \in (0,2)$ and set $t_N = (2-\epsilon)\log N$. Then

$$N\mathbb{P}(Z_1 > t_N) = 2N\{1 - \Phi(\sqrt{t_N})\} \ge \sqrt{\frac{2}{\pi}}\frac{\sqrt{t_N}}{1+t_N} N^{\epsilon/2} \longrightarrow \infty.$$

Independence implies $\mathbb{P}(M_N > t_N) \ge 1 - \exp\{-N\mathbb{P}(Z_1 > t_N)\} \to 1$. As such, $\mathbb{E}[M_N] \ge t_N \mathbb{P}(M_N > t_N)$, so the liminf of $\mathbb{E}[M_N]/\log N$ is at least $2-\epsilon$. Letting $\epsilon \downarrow 0$ and using the upper bound completes the proof. □

### B.11 Proof of Proposition 5.1

*Proof.* According to the decomposition (5), it suffices to prove (41) below. Fix the response and write $a_j = \hat{\beta}^2_{(j),p}$, so that $a_1 \ge \cdots \ge a_p \ge 0$. For $m \le m'$, let $u_j = w_j^{k,m,p}$, $v_j = w_j^{k,m',p}$, and $S_l = \sum_{j=1}^{l}(v_j - u_j)$, with $S_0 = 0$. By (38), we have $S_l \ge 0$, with $S_p = 0$ because both weight sequences sum to $k$. By Lemma B.1, the sequence $a_j(u_j + v_j)$ is nonnegative and nonincreasing. Factoring the difference of squares and summing by parts gives

$$\begin{aligned}\sum_{j=1}^{p} a_j(v_j^2 - u_j^2) &= \sum_{j=1}^{p} a_j(u_j + v_j)(v_j - u_j) \\ &= \sum_{l=1}^{p-1} \big\{a_l(u_l + v_l) - a_{l+1}(u_{l+1} + v_{l+1})\big\} S_l \ge 0.\end{aligned} \tag{41}$$

Under orthogonality, the two sums are the squared Euclidean norms of the corresponding coefficient vectors and the squared normalized norms of the fitted vectors. This proves the inequality, including the cases $k = 0$, $k = p$, and $m = m'$. □

### B.12 Proof of Proposition 6.1

*Proof.* For convenience, number the original features in the fixed tie order. Thus ties are resolved by the smallest index. For an active set $M$, the squared selection score of an eligible feature $j$ is $|\langle \mathbf{x}_{\cdot j,p}, \mathbf{P}_M^{\perp}\mathbf{y}\rangle|^2 / \|\mathbf{P}_M^{\perp}\mathbf{x}_{\cdot j,p}\|^2$. It equals the reduction in squared residual loss upon adding $j$. A support $T$ of size $k$ is a strict local minimum under exchanges of a single feature if $L(T) < L((T \setminus \{i\}) \cup \{j\})$ for all $i \in T$, $j \notin T$, where $L(M) = \|\mathbf{y} - \mathbf{P}_M\mathbf{y}\|^2$.

For each $(n, p, k, m)$ in the stated domain, we construct the design and noiseless response before drawing the candidate sets. We consider two cases. First suppose $1 \le m \le p - k$. Take $n \ge p$ and vectors $\mathbf{e}_1, \ldots, \mathbf{e}_p$ orthonormal in the normalized inner product. Let $S \ge 6$ be a parameter to be chosen below, and define

$$\mathbf{y}_0 = \mathbf{e}_1 + \mathbf{e}_2 + \sqrt{S-2}\,\mathbf{e}_3, \qquad a = \frac{3}{4S}, \qquad s = \frac{9}{16}.$$

Use the features

$$\mathbf{x}_{\cdot 1,p} = \mathbf{e}_1, \qquad \mathbf{x}_{\cdot 2,p} = \mathbf{e}_2, \qquad \mathbf{x}_{\cdot 3,p} = a\mathbf{y}_0 + \sqrt{1 - s/S}\,\mathbf{e}_4, \qquad \mathbf{x}_{\cdot 4,p} = a\mathbf{y}_0 - \sqrt{1 - s/S}\,\mathbf{e}_4,$$

and $\mathbf{x}_{\cdot i,p} = \mathbf{e}_i$ for $5 \le i \le p$. Write $H = \{5, \ldots, k+2\}$ for the $k-2$ padding features, with $H$ empty when $k = 2$, and set

$$\mathbf{y} = \mathbf{y}_0 + 2\sum_{i \in H} \mathbf{e}_i.$$

Features indexed by $\{k+3, \ldots, p\}$ have zero response coefficient. Every feature has norm one. Since $S > 2$ and $1 - s/S > 0$, the first four features span $\mathbf{e}_1, \ldots, \mathbf{e}_4$. The remaining features are mutually orthogonal and orthogonal to this span, so the dictionary has full column rank.

The padding features have score 2 until selected, the first two features initially have score 1, the two good features $3, 4$ have score $3/4$, and the null features have score zero. The padding features are orthogonal to the other features, so FS selects all of $H$ first. After either feature 1 or 2 is selected, the squared score of either good feature is

$$g_1(S) = \frac{s(1-1/S)^2}{1-s/S^2} < s < 1,$$

whereas the other member of $\{1, 2\}$ has squared score one. Consequently FS($k$) selects $T = H \cup \{1, 2\}$, with residual loss $S - 2$.

We now show that $T$ is a strict local minimum. A feature outside $T$ is either one of $3, 4$ or a null feature. If a feature in $\{1, 2\}$ is replaced by a good feature, the loss increases by $1 - g_1(S) > 0$. Replacement by a null feature increases it by one. If a padding feature is removed, the loss increases by 4. Adding a good feature to the remaining first two features reduces the loss by only

$$g_2(S) = \frac{s(1-2/S)^2}{1-2s/S^2} < s < 4,$$

so this exchange increases the loss by $4 - g_2(S) > 0$. Adding a null feature captures nothing and increases the loss by four. The inequalities $g_1(S) < s$ and $g_2(S) < s$ follow by comparing their numerators and denominators for $S \geq 6$. Thus every exchange of a single feature increases the loss, so $T$ is a strict local minimum. On the other hand, $H \cup \{3, 4\}$ fits the response exactly, because $\mathbf{x}_{\cdot 3,p} + \mathbf{x}_{\cdot 4,p} = 2a\mathbf{y}_0$.

We now study the base estimator of EFS, denoted $\check{f}_{k,m,p}$, which depends on a random seed $\theta$. We construct a sequence of candidate sets that selects all padding features first, then feature 3, then feature 4. For each padding step, choose any fixed candidate set containing the remaining padding feature with the smallest index. This feature is selected because its score is two and all scores outside the padding set are smaller. After these $k - 2$ steps, there are $p - k + 2$ eligible features. Exactly $p - k$ of them lie outside $\{1, 2\}$. The hypothesis $m \leq p - k$ therefore allows a fixed candidate set of size $m$ containing feature 3 and excluding features $1, 2$. Its other members are feature 4 and/or null features. It selects feature 3.

At the last step, choose a fixed candidate set containing feature 4. After feature 3 is selected, adding feature 4 eliminates the remaining core loss $S - s$. Each of features $1, 2$ instead has squared score

$$\frac{(1-s/S)^2}{1-s/S^2} < 1 < S - s,$$

and the null features have score zero. Thus feature 4 is selected, and the final loss is zero. By the score inequalities above, the same sequence of candidate sets gives this selection order for every $S \geq 6$.

At step $l \leq k$, at least $p - k + 1 > m$ features remain. Hence the probability of this particular path of candidate sets is

$$\rho = \prod_{l=1}^{k} \binom{p-l+1}{m}^{-1} > 0,$$

independently of $S$. The fit along every path is an orthogonal projection, so its residual loss is at most $\|\mathbf{y}\|^2 = S + 4(k-2)$. It follows that

$$\mathbb{E}_\theta\big[\|\mathbf{y} - \check{\mathbf{f}}_{k,m,p}\|^2\big] \leq (1-\rho)\{S + 4(k-2)\}.$$

Choose

$$S > \max\left\{6, \frac{2 + 4(k-2)(1-\rho)}{\rho}\right\}.$$

Then the last upper bound is strictly smaller than $S - 2$, the FS loss. Using the Jensen inequality, we also have

$$\|\mathbf{y} - \hat{\mathbf{f}}_{k,m,p}\|^2 \leq \mathbb{E}_\theta\big[\|\mathbf{y} - \check{\mathbf{f}}_{k,m,p}\|^2\big] < S - 2,$$

which proves the result in this case.

Now suppose $p-k+1 \le m \le p-1$. The preceding path can no longer exclude both features $1,2$ after selecting the padding features. For fixed $(n,p,k)$, we therefore use a different construction that works for every $m$ in this range. Retain the orthonormal vectors and the padding index set $H$, but put

$$\mathbf{y}_0 = 3\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3, \qquad \mathbf{x}_{\cdot 1,p} = \mathbf{e}_1, \quad \mathbf{x}_{\cdot 2,p} = \mathbf{e}_2,$$
$$\mathbf{x}_{\cdot 3,p} = \frac{2\mathbf{y}_0 + \sqrt{77}\mathbf{e}_4}{11}, \qquad \mathbf{x}_{\cdot 4,p} = \frac{2\mathbf{y}_0 - \sqrt{77}\mathbf{e}_4}{11},$$

with $\mathbf{x}_{\cdot i,p} = \mathbf{e}_i$ for $i \ge 5$, and set

$$\mathbf{y} = \mathbf{y}_0 + \frac{1}{2}\sum_{i \in H} \mathbf{e}_i.$$

The features are normalized and linearly independent. The initial core scores are $3,1,2,2$, the padding scores are $1/2$, and the null scores are zero. FS first selects feature 1. The squared scores of feature 2 and either good feature are then 1 and $16/85$, respectively, so it selects feature 2. Thereafter, either good feature has squared score $4/81 < 1/4$, so all padding features are selected before either good feature. Thus FS($k$) selects $T = H \cup \{1,2\}$ and has loss one.

For the four exchanges of a feature in $\{1,2\}$ for a good feature, we have

$$\|\mathbf{y}_0 - \mathbf{P}_{\{1,j\}}\mathbf{y}_0\|^2 = \frac{154}{85} > 1, \qquad \|\mathbf{y}_0 - \mathbf{P}_{\{2,j\}}\mathbf{y}_0\|^2 = \frac{770}{117} > 1, \qquad j \in \{3,4\}.$$

For example, after feature 1 the good feature adds $16/85$ to the squared fitted norm. After feature 2 it adds $400/117$. Replacing a padding feature by a good feature loses $1/4$ in squared fitted norm and gains only $4/81$, so the loss also increases. Replacing any selected feature by a null feature loses its positive contribution to the squared fitted norm. Therefore $T$ is again a strict local minimum under every exchange of a single feature.

Write $q = p-k-2$ for the number of null features. We first show that no null feature is selected during the first $k$ randomized steps. Notice that there are at most two eligible core features with zero score at any stage. Initially all four core scores are positive. After one core feature has been selected, the core residual is nonzero, so at least one of the three remaining core features has positive score. After two or more core selections, at most two core features remain. Consequently at most $q+2$ eligible features have zero score. Also, an exact fit is impossible before step $k$, because the unique representation in the full rank dictionary uses both good features and all $k-2$ padding features. Hence a feature with positive score remains before every step under consideration. At each step $l \le k$, the actual candidate size is $\min\{m, p-l+1\} \ge p-k+1 = q+3$. Every candidate set therefore contains a feature with positive score, including when the cap on the candidate size applies. Therefore, no null feature is selected.

Next, we show that the final support cannot contain exactly one good feature. At the last step all $p-k+1 = q+3 \le m$ eligible features are candidates. If one good feature has already been selected, its partner captures the entire remaining core residual. Projection loss is nonincreasing as features are added, so the squared norm of that residual is at least its value after both bad features and one good feature have been selected, namely $1 - 4/81 = 77/81$. The partner therefore beats every padding feature, whose squared score is $1/4$, and every other core feature, which captures strictly less of the remaining core residual. To see the latter strict inequality, note that by full rank, the unique representation of $\mathbf{y}_0$ uses both good features with nonzero coefficients. Hence a support missing one cannot fit $\mathbf{y}_0$ exactly. The partner must therefore be selected at the last step.

It remains to exclude the possibility that the first good feature is selected only at the last step. Before that step the selected set is $T \setminus \{i\}$ for some $i \in T$. Since $T$ is a strict local minimum, adding $i$ reduces the loss more than adding any feature outside $T$. The full candidate set therefore selects $i$, which excludes this possibility.

Every final support consequently either contains no good feature, in which case it is $T$ and has loss one, or contains both good features. In the latter case its core loss is zero. If it has $r$ core features in total, it omits $r-2$ padding features, so its loss is $(r-2)/4 \le 1/2$, because $2 \le r \le 4$.

To obtain strict improvement, note that the first candidate set contains feature 3 and excludes feature 1 with probability

$$\frac{m(p-m)}{p(p-1)} > 0.$$

On this event feature 3 is selected because its score two beats the other available scores, with the index tie rule choosing it if feature 4 is present. The final support then contains both good features and has loss at most $1/2$. Putting these together, we have

$$\mathbb{E}_\theta\big[\|\mathbf{y}-\check{\mathbf{f}}_{k,m,p}\|^2\big] \le 1-\frac{m(p-m)}{2p(p-1)} < 1.$$

This is strictly below the FS loss. Using the Jensen inequality again gives strictly smaller error for the infinite ensemble. The two cases cover every $1 \le m < p$, which completes the proof. □

*Remark* (An example with four features). For the first construction, take $p = 4$, $k = 2$, $m \in \{1, 2\}$, and $S = 6$. It gives $\mathbf{y} = \mathbf{e}_1 + \mathbf{e}_2 + 2\mathbf{e}_3$ and $\mathbf{x}_{\cdot 3,p} = (\mathbf{y} + \sqrt{58}\mathbf{e}_4)/8$, $\mathbf{x}_{\cdot 4,p} = (\mathbf{y} - \sqrt{58}\mathbf{e}_4)/8$. The FS support $\{1, 2\}$ has loss four, each of its four exchanges of a single feature has loss $290/63 > 4$, and $\{3, 4\}$ has loss zero. For $m = 2$, the final support is $\{1, 2\}$, $\{3, 4\}$, or a cross pair with probabilities $5/9$, $1/9$, and $1/3$, respectively. The first candidate pair contains a member of $\{1, 2\}$ with probability $5/6$, and the second candidate pair contains the partner of the selected feature with probability $2/3$. That partner has the largest score, as in the proof. Thus the mean base learner loss is $710/189 < 4$. For $m = 1$, the six supports with two features are equally likely, giving mean loss $706/189 < 4$. For $m = 3$, however, this particular design returns the FS support on every path. The first candidate set contains feature 1 or 2, and the second contains all remaining features. For $m = 3$, we instead use the second construction in the proof.

*Remark.* The exclusions are necessary for strict improvement in the mean base learner loss. At $k = 1$ or $k = p - 1$, every competing support of size $k$ can be obtained by exchanging a single feature, so a local minimum is global and no individual base learner can improve it. This does not, however, rule out an ensemble improvement, since an average of fits using $k$ features need not itself be a fit using $k$ features. At $k = 0$ every fit is zero, at $k = p$ every full rank path gives the same fit, and at $m = p$ randomization disappears.

### B.13 A finite guarantee for a fixed correlated family

We now consider a different construction, in which randomization avoids a poor greedy choice. The FS support in this construction is not a local minimum. The following bound holds for all candidate sizes $2 \le m < p$ and gives the guarantee for a fixed proportion as a special case.

**Proposition B.2** (A finite guarantee for all nontrivial candidate sizes)**.** *Let $n \ge p \ge 4$, $2 \le m \le p-1$, and $3 \le k \le p-1$ be integers. Take normalized orthonormal vectors $\mathbf{x}_{\cdot 1,p}, \dots, \mathbf{x}_{\cdot p-1,p}, \mathbf{u}$, put $\mathbf{x}_{\cdot p,p} = (\mathbf{x}_{\cdot 1,p} + \mathbf{x}_{\cdot 2,p} + \mathbf{u})/\sqrt{3}$, and let $\mathbf{y} = \beta(\mathbf{x}_{\cdot 1,p} + \mathbf{x}_{\cdot 2,p}) + \zeta \sum_{j=3}^{p-1} \mathbf{x}_{\cdot j,p}$, where $\beta/\sqrt{6} > \zeta > 0$. Resolve ties by the smallest original feature index. Define*

$$\delta = 1 - \frac{m}{p}, \qquad K_{m,p} = \max\{3, p-m+1\},$$

$$b_{k,m,p} = \begin{cases} 0, & k \ge K_{m,p}, \\ \min\{1, 2\mathbb{P}(\mathrm{Bin}(k, m/p) \le 2)\}, & k < K_{m,p}. \end{cases}$$

*Then the single randomized base learner satisfies*

$$\|\mathbf{y}-\hat{\mathbf{f}}_{k,p}\|^2 - \mathbb{E}_\theta\big[\|\mathbf{y}-\check{\mathbf{f}}_{k,m,p}\|^2\big] \ge \frac{\zeta^2\delta}{3} - 2(\beta^2-\zeta^2)b_{k,m,p}. \tag{42}$$

*In particular, for every $2 \le m \le p-1$ the gain is at least $\zeta^2(p-m)/(3p) > 0$ whenever $K_{m,p} \le k \le p-1$. The gain is also strictly positive if*

$$\frac{km}{p} \ge 2\log\left(\frac{96(\beta^2-\zeta^2)p}{\zeta^2(p-m)}\right). \tag{43}$$

*Under this condition the gain is at least $\zeta^2(p-m)/(6p) > 0$. The infinite ensemble has no larger loss than the mean base learner loss.*

*Proof.* Here, the randomness is over the choices of the candidate sets, with the response fixed. For an active set $M$ and an unselected feature $j$, define its selection score by

$$s_j(M) = \frac{|\langle \mathbf{x}_{\cdot j,p}, \mathbf{P}_M^\perp \mathbf{y}\rangle|}{\|\mathbf{P}_M^\perp \mathbf{x}_{\cdot j,p}\|}.$$

Since the construction has full column rank, all denominators are positive. Each unselected distractor $3 \le j \le p-1$ has score $\zeta$. For an unselected signal feature $j \in \{1,2\}$, write $j'$ for the other signal feature. We have

$$s_j(M) = \begin{cases} \beta, & p \notin M, \\ \beta/\sqrt{6}, & p \in M,\ j' \notin M, \\ \beta/\sqrt{2}, & p \in M,\ j' \in M. \end{cases} \tag{44}$$

For $p \notin M$, the spurious feature has score

$$s_p(M) = \frac{(2-t)\beta}{\sqrt{3-t}} = \begin{cases} 2\beta/\sqrt{3}, & t=0, \\ \beta/\sqrt{2}, & t=1, \\ 0, & t=2, \end{cases} \qquad t = |M \cap \{1,2\}|. \tag{45}$$

These formulas follow because every distractor is orthogonal to the three special features. For example, after selecting only feature $p$ among those features, a signal feature has residual correlation $\beta/3$ and squared residualized norm $2/3$. After selecting $p$ and the other signal, these are $\beta/2$ and $1/2$. If $p$ is unselected, its residual correlation and squared residualized norm are $(2-t)\beta/\sqrt{3}$ and $(3-t)/3$.

Putting (44) and (45) together, FS first selects $p, 1, 2$. This means that its loss is $L_{\mathrm{FS}} = (p-k)\zeta^2$. For $k \le p-1$, exchanging its redundant feature $p$ for any unselected distractor lowers the loss by $\zeta^2$. Thus this FS support is not a local minimum under exchanges of a single feature, including at $k = p-1$. We now study the randomized path. Every unselected signal has score at least $\beta/\sqrt{6} > \zeta$. As long as a signal remains unselected, every remaining special feature in $\{1,2,p\}$ also has score larger than the distractor score $\zeta$. Before any special feature has been selected, $p$ outranks the two signals. After one signal has been selected without $p$, the other signal outranks $p$. After both signals have been selected, $p$ has zero score and, because $m \ge 2$, is not selected before the final step.

Let $T_j$ denote the selection time of original feature $j$, extending the path through all $p$ steps, and define

$$G = \big\{\max\{T_1, T_2\} < T_p\big\}, \qquad E_3 = \big\{\max\{T_1, T_2\} > k\big\}, \qquad E_1 = G \cap E_3^c.$$

We first compute a lower bound for $\mathbb{P}(G)$. For fixed nominal candidate size $m \ge 2$ and $N \ge 2$, let $H_N$ be the probability that the remaining signal is selected before $p$, in a state containing that signal, $p$, and $N-2$ distractors, when the other signal has already been selected. For $N \ge 3$, let $G_N$ be the probability that both signals are selected before $p$ in a state containing all three special features and $N-3$ distractors, with none of the special features yet selected. Previously selected distractors do not change these scores or probabilities. If $N \le m$, the candidate set is the full remaining set, so $H_N = 1$ and $G_N = 0$ (where their respective states exist). For $N > m$, conditioning on the first selected feature gives the first recurrence below. The second holds when also $N \ge 4$.

$$\begin{aligned} H_N &= \frac{m}{N} + \frac{(N-m)(N-m-1)}{N(N-1)} H_{N-1}, \\ G_N &= a_N H_{N-1} + d_N G_{N-1}, \\ d_N &= \frac{(N-m)(N-m-1)(N-m-2)}{N(N-1)(N-2)}, \qquad a_N = 1 - \frac{m}{N} - d_N. \end{aligned}$$

Here $d_N$ is the probability of a candidate set containing only distractors, and $a_N$ is the probability that a signal is selected first, requiring exclusion of $p$. The remaining boundary case without a cap is $m = 2, N = 3$, where direct conditioning gives $G_3 = 1/3$.

At the first step whose candidate set contains either of the two remaining special features, symmetry makes each equally likely to appear alone. If both appear, the signal is selected. Thus $H_N \ge 1/2$. We next prove $G_N \ge (N-m)/(3N)$ for $N \ge \max\{m, 3\}$. For $m \ge 3$, the base case $N = m$ has both sides zero. For $m = 2$, use $G_3 = 1/3 \ge 1/9$. For the induction step, let $r = N-1 \ge m$ and $\delta_N = (N-m)/N$. Substituting $H_{N-1} \ge 1/2$ and the bound for $G_{N-1}$ under the inductive assumption, we obtain

$$\begin{aligned} G_N - \frac{\delta_N}{3} &\ge \frac{a_N}{2} + \frac{d_N(N-m-1)}{3(N-1)} - \frac{\delta_N}{3} \\ &= \frac{\delta_N m\{(3m+1)r - 2m(m+1)\}}{6r^2(r-1)} \ge 0. \end{aligned}$$

The last inequality follows from $r \geq m$, because the expression in braces is at least $m(m-1)$. Thus

$$\mathbb{P}(G) = G_p \geq \frac{p-m}{3p} = \frac{\delta}{3}.$$

To compute a bound for $\mathbb{P}(E_3)$, let us generate the candidate sets as follows. Independently at each step, first draw a uniform subset of size $m$ from the $p$ features. Retain its unselected members and fill the vacancies uniformly from the remaining eligible features. When fewer than $m$ features remain, use all of them. Conditional on the past, permutation symmetry makes the completed set uniform among eligible candidate sets. Every preliminary appearance of an unselected signal is retained. Only the other two special features can outrank it, and each can be selected only once, so three preliminary appearances force that signal to be selected. For each signal, the preliminary appearance indicators are IID $\mathrm{Bernoulli}(m/p)$ across steps. A union bound gives

$$\mathbb{P}(E_3) \leq 2\mathbb{P}\big(\mathrm{Bin}(k, m/p) \leq 2\big).$$

We can also bound the latest selection time of the signals. Suppose a signal is still unselected after $k$ steps. Throughout those steps every remaining special feature outranks every distractor. Therefore a distractor can be selected only from a candidate set containing only distractors, which requires at least $m$ remaining distractors. At most $\max\{p-m-2, 0\}$ distractors can have been selected. At most two special features have been selected, since a signal is missing. It follows that $k \leq \max\{p-m, 2\}$. Consequently $\mathbb{P}(E_3) = 0$ for $k \geq K_{m,p}$, and in all cases $\mathbb{P}(E_3) \leq b_{k,m,p}$.

We next compute the loss for each possible set of selected features to combine these probability bounds. Write $r \in \{0,1,2\}$ for the number of selected signals and $s \in \{0,1\}$ for selection of $p$. The loss is

$$L_\theta = (p-k-3+r+s)\zeta^2 + \beta^2 e_{r,s},$$

where the residual signal factors are

| | $r=0$ | $r=1$ | $r=2$ |
|---|---|---|---|
| $s=0$ | 2 | 1 | 0 |
| $s=1$ | 2/3 | 1/2 | 0 |

(each entry multiplies $\beta^2$). On $E_1$, the gain is exactly $\zeta^2$. On $E_3^c \setminus E_1$, it is zero. On $E_3$, $r \leq 1$ and the table gives $L_{\mathrm{FS}} - L_\theta \geq 3\zeta^2 - 2\beta^2$, since $\beta^2 > 6\zeta^2$. Therefore

$$\begin{aligned} L_{\mathrm{FS}} - \mathbb{E}_\theta[L_\theta] &\geq \zeta^2\mathbb{P}(E_1) - (2\beta^2 - 3\zeta^2)\mathbb{P}(E_3) \\ &\geq \zeta^2\mathbb{P}(G) - 2(\beta^2 - \zeta^2)\mathbb{P}(E_3) \\ &\geq \frac{\zeta^2\delta}{3} - 2(\beta^2 - \zeta^2)b_{k,m,p}. \end{aligned}$$

This proves (42) and the assertion for $k \geq K_{m,p}$. Using the recurrences above, we may also replace $\delta/3$ by $G_p \geq \delta/3$ throughout.

For the logarithmic sufficient condition, put $\mu = km/p$. For $\mu \geq 2$, the Chernoff bound gives

$$\mathbb{P}\big(\mathrm{Bin}(k, m/p) \leq 2\big) \leq e^{-\mu}(e\mu/2)^2 \leq 4e^{-\mu/2}.$$

The last inequality follows by maximizing $\mu^2 e^{-\mu/2}$ at $\mu = 4$. Condition (43) implies $\mu > 2$, because $\beta^2/\zeta^2 > 6$, and gives $16(\beta^2 - \zeta^2)e^{-\mu/2} \leq \zeta^2\delta/6$. Combining this with (42) proves the assertion. Finally, using the Jensen inequality, we have $\|\mathbf{y} - \hat{\mathbf{f}}_{k,m,p}\|^2 \leq \mathbb{E}_\theta[L_\theta]$. □

*Remark.* For fixed proportions, suppose $0 < a \leq m/p \leq b < 1$. Then choosing $k \geq (2/a)\log\{96(\beta^2 - \zeta^2)/(\zeta^2(1-b))\}$ suffices, uniformly in $p$, subject to $3 \leq k \leq p-1$. For $m = p-1$ and $3 \leq k \leq p-1$, the mean base learner gain is exactly $\zeta^2/p$. The first candidate set excludes $p$ with probability $1/p$, forcing both signals to be selected before $p$. Otherwise $p$ is selected first. Both signals are always selected within three steps.

For this family, $m = 1$ is, however, strictly worse in mean base learner loss for every $3 \leq k \leq p-1$. Choosing one candidate at each step gives a uniform subset of size $k$. Let $R$ be the number of omitted signal features. The squared norm of the residual signal is at least $\beta^2 R/3$ in every state, by the table above. Thus its expected loss is at least

$$\frac{p-k}{p}\left((p-3)\zeta^2 + \frac{2}{3}\beta^2\right),$$

which exceeds $(p-k)\zeta^2$ because $\beta^2 > 6\zeta^2$ and $k < p$. Even with $m \geq 2$, a threshold on $k$ independent of $m, p$ cannot suffice for this fixed family. Take $m = 2$ and $p = k^3$, for example, as $k \to \infty$. A union bound gives probability at most $6k/(p-k+1) \to 0$ that any special feature appears among the first $k$ candidate sets. Otherwise all $k$ selections are distractors, and the difference between the FS loss and the base learner loss is $3\zeta^2 - 2\beta^2 < 0$. From the table above, the loss difference on every path is bounded by a constant depending only on $\beta, \zeta$. Hence the expected difference tends to this same negative value.

*Remark* (Using increasing inclusion probabilities). Under the hypotheses of Proposition B.2, the failure term can be sharpened by using the inclusion probability at each step in place of $m/p$. Let $B_1, \ldots, B_k$ be independent with $B_l \sim \text{Bernoulli}(a_l)$, where $a_l = \min\{m/(p-l+1), 1\}$. Replace $b_{k,m,p}$ by

$$b^+_{k,m,p} = \begin{cases} 0, & k \geq K_{m,p}, \\ \min\{1, 2\mathbb{P}(B_1 + \cdots + B_k \leq 2)\}, & k < K_{m,p}. \end{cases}$$

Bound (42) remains valid with $b^+_{k,m,p}$ in place of $b_{k,m,p}$, and $b^+_{k,m,p} \leq b_{k,m,p}$. To see this, fix one signal $j$, and write $T_j$ for its selection time. At each step when $j$ is unselected, first draw its inclusion indicator as $B_l$, then sample a uniform candidate set conditional on that inclusion or exclusion. This gives the required uniform distribution over candidate sets, since each eligible feature is included with probability $a_l$. After $j$ has been selected, draw the subsequent $B_l$ independently and ignore them when sampling candidates. Three inclusions before selection force $j$ to be selected, so $\mathbb{P}(T_j > k) \leq \mathbb{P}(B_1 + \cdots + B_k \leq 2)$. Applying this construction separately to the two signals and then using a union bound gives the result.

Let

$$\mu^+_{k,m,p} = \sum_{l=1}^{k} \min\left\{\frac{m}{p-l+1}, 1\right\}.$$

For $k < K_{m,p}$, we have $k \leq p - m$, so no cap applies and $\mu^+_{k,m,p} = m(\mathsf{H}_p - \mathsf{H}_{p-k})$, where $\mathsf{H}_r = \sum_{i=1}^{r} 1/i$. The Chernoff calculation applies to independent Bernoulli variables with unequal probabilities as well. Hence the condition

$$\mu^+_{k,m,p} \geq 2\log\left(\frac{96(\beta^2-\zeta^2)p}{\zeta^2(p-m)}\right)$$

gives mean base learner gain at least $\zeta^2(p-m)/(6p)$. When $k < K_{m,p}$, apply the preceding Chernoff argument with this mean. When $k \geq K_{m,p}$, the stronger bound for this range already holds. Since $\mu^+_{k,m,p} \geq km/p$, this condition includes the earlier sufficient one.

Finally, fix $m \geq 2$ and $\beta/\sqrt{6} > \zeta > 0$ independently of $p$. For any fixed $c \in (0,1)$ satisfying

$$-m\log(1-c) > 2\log\left(\frac{96(\beta^2-\zeta^2)}{\zeta^2}\right),$$

and any integer sequence $k_p/p \to c$, we have $\mu^+_{k_p,m,p} \to -m\log(1-c)$. Therefore, the sufficient condition holds for all $p \geq p_0$, where $p_0$ may depend on $m, \beta, \zeta, c$ and the sequence $k_p$. Such a fixed $c < 1$ exists, so the result includes model sizes whose limiting fraction is strictly below one.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the theoretical contributions and scope of the paper. EFS is presented as a tractable abstraction of feature subsampled greedy fitting, with orthogonal design assumptions, correlated noiseless constructions, and empirical simulations beyond the orthogonal setting clearly described.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The introduction and setup explicitly discuss the main limitations. The analysis concerns EFS rather than full RFs, abstracts away tree structured split availability and bootstrap sampling, and relies on an orthogonal fixed design assumption for the weight and degrees of freedom results. The correlated local minimum result is existential and noiseless. The simulations are presented as empirical checks beyond orthogonality, not as formal guarantees.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theorem and proposition states its assumptions, including orthogonality, asymptotic regimes, and distributional assumptions where needed. Complete proofs are provided in the Supplementary Material, with numbered and cross referenced results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The main simulations use synthetic data. Section "Simulations" specifies the data generating distribution, $n$, $p$, covariance structure, sparsity pattern, SNR values, number of EFS base learners, cross validation procedure, and model size range, and Appendix A gives the $m$ grid and number of independent replicates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The linked repository includes synthetic results, environment requirements, and commands to reproduce the figures and rerun the experiment. The code is released under the MIT license.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section "Simulations" specifies the data distribution, covariance model, SNR definition, sparsity pattern, number of base learners, ten fold cross validation jointly selecting $(k, m)$ followed by refitting the full $k$ path at the selected $m$, and Appendix A gives the exact $m$ grid.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 2 reports pointwise 95% Monte Carlo intervals, and Appendix A defines the intervals and reports uncertainty for paired differences, including their sources of variability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The simulations used an Apple M3 CPU with 24 GB of memory and three or four workers, each with one numerical library thread. The fits for Figure 2 took about 47 worker minutes, excluding plotting. Additional runs for this comparison took about 20 worker minutes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The work is theoretical and uses synthetic simulations, so it does not involve human subjects, private data, deployed systems, or high risk models.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper is foundational theoretical work on feature subsampling in greedy model search and does not propose a deployed system or application with direct societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release data or models with a high risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use external datasets, pretrained models, or third party research implementations as part of the reported experiments, and related work is cited in the bibliography. Standard numerical library dependencies are documented in the linked repository.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The accompanying archive documents the new simulation code and synthetic outputs, including their MIT license, dependencies, limitations, and reproduction commands.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The acknowledgments describe the use of OpenAI Codex (GPT-6 Astra) during revision.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.