Precise Asymptotics of Bagging Regularized M-estimators

Takuya Koriyama*† tkoriyam@uchicago.edu
Pratik Patil*‡ pratikpatil@berkeley.edu
Jin-Hong Du§¶ jinhongd@andrew.cmu.edu
Kai Tan‖ kai.tan@rutgers.edu
Pierre C. Bellec‖ pierre.bellec@rutgers.edu

Abstract

We characterize the squared prediction risk of ensemble estimators obtained through subagging (subsample bootstrap aggregating) regularized M-estimators and construct a consistent estimator for the risk. Specifically, we consider a heterogeneous collection of $M \geq 1$ regularized M-estimators, each trained with (possibly different) subsample sizes, convex differentiable losses, and convex regularizers. We operate under the proportional asymptotics regime, where the sample size n, feature size p, and subsample sizes k_m for $m \in [M]$ all diverge with fixed limiting ratios n/p and k_m/n. Key to our analysis is a new result on the joint asymptotic behavior of correlations between the estimator and residual errors on overlapping subsamples, governed through a (provably) contractible nonlinear system of equations. Of independent interest, we also establish convergence of trace functionals related to degrees of freedom in the non-ensemble setting (with $M = 1$) along the way, extending previously known cases for square loss and ridge, lasso regularizers.

When specialized to homogeneous ensembles trained with a common loss, regularizer, and subsample size, the risk characterization sheds some light on the implicit regularization effect due to the ensemble and subsample sizes (M, k). For any ensemble size M, optimally tuning subsample size yields sample-wise monotonic risk. For the full-ensemble estimator (when $M \to \infty$), the optimal subsample size $k^\star$ tends to be in the overparameterized regime ($k^\star \leq \min\{n, p\}$), when explicit regularization is vanishing. Finally, joint optimization of subsample size, ensemble size, and regularization can significantly outperform regularizer optimization alone on the full data (without any subagging).

1 Introduction

Ensemble methods combine predictions of multiple models to improve predictive accuracy [HTF09]. Among these methods, bagging (bootstrap aggregating) trains individual models on bootstrapped samples of the dataset and averages their predictions to reduce variance and mitigate overfitting [Bre96]. A popular variant of bagging, known as subagging (subsample bootstrap aggregating), trains models on random subsamples rather than full bootstrapped samples [BY02]. Apart from the computational advantages, subagging can substantially improve predictive performance, especially in the overparameterized regimes and near model interpolation thresholds [PKWR22]. In this paper,

*Corresponding authors.
†Booth School of Business, The University of Chicago, Chicago, IL, 60637, USA.
‡Department of Statistics, University of California, Berkeley, CA 94720, USA.
§Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
¶Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
‖Department of Statistics, Rutgers University, New Brunswick, NJ 08854, USA.

Table 1: **Landscape of subagging risk analysis in high dimensions.** We summarize the various settings (estimator structure and data structure) of some of the recent works that characterize the prediction risk of subagging regularized M-estimators trained with loss function loss and regularization function reg (defined in Section 2). Convex$^\bullet$ denotes that in addition to being convex, we require the loss function to be differentiable and have Lipschitz continuous derivatives. By RMT features, we refer to features of the form $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}$ where $\boldsymbol{v}$ contains independent entries of bounded moments (of order 4^+) that are common in the random matrix theory literature. Arbitrary$^\blacktriangle$ response y refers to no additional modeling assumption on the response other than bounded moments (of order 4^+). Signal and noise refer to $\boldsymbol{\theta}$ and z in the linear data model: $y = \boldsymbol{x}^\top\boldsymbol{\theta} + z$. Arbitrary$^\blacksquare$ signal and noise distributions F_θ and F_z refer to general marginal distribution on the coordinates of signal $\boldsymbol{\theta}$ and noise z, respectively, subject to mild regularity conditions (Assumption D).

Loss (loss)	**Penalty** (reg)	**Features** ($\boldsymbol{x}$)	**Covariance** ($\boldsymbol{\Sigma}$)	**Response** ($\boldsymbol{y}$)	**Signal** ($\boldsymbol{\theta}$)	**Signal Dist.** (F_θ)	**Noise Dist.** (F_z)	**Reference**
Square		Gaussian	Isotropic	Linear	Random	Gaussian	Gaussian	[LJB20]
Square	Ridge	RMT	Anisotropic	Linear	Deterministic		Bnd. Mom.	[PDK23]
Square	Ridge	RMT	Anisotropic	Arbitrary$^\blacktriangle$				[PD23]
Huber		Gaussian	Isotropic	Linear	Deterministic		Arbitrary$^\blacksquare$	[BK24]
Logistic		Gaussian	Isotropic	Logistic	Deterministic			[BK24]
Convex	Ridge	Gaussian	Isotropic	GLM	Random	Gaussian		[CVD$^+$24]
Convex$^\bullet$	Convex	Gaussian	Isotropic	Linear	Random	Arbitrary$^\blacksquare$	Arbitrary$^\blacksquare$	Corollary 3

we analyze the squared prediction risk of subagging of regularized M-estimators trained with convex loss and regularizer.

There is growing interest in understanding the prediction risk asymptotics of ensemble methods, particularly subagging, across different types of predictors and under various data assumptions. For example, [LJB20] study subagging in the context of ordinary least squares regression without any explicit regularization in the underparameterized regime (where the number of subsamples is higher than the number of features). This is extended to ridge and ridgeless regression by [PDK23] for both the underparameterized and overparameterized regimes (where the number of subsamples is lower than the number of features). [PD23] further generalizes these results and identifies explicit equivalence paths between subsampled estimators and ridge regularized estimators. Beyond linear and ridge regression, [BK24] examines the behavior of subagging in logistic and Huber regression models without regularization. In addition to subagging, there has also been considerable work on feature sketching and other ensemble methods. For instance, [LGR$^+$22] and [PL24] study feature sketching and ensembling to optimize predictive performance in high-dimensional settings. For other recent developments in the analysis of ensemble methods and related work details, see Section 1.2.

We generalize these previous works by characterizing the prediction risk of subagging a collection of M regularized M-estimators and constructing a consistent estimator for this risk. We allow the collection to be heterogeneous, consisting of estimators trained with convex differentiable loss function loss with Lipschitz continuous derivative and convex regularization function reg, which can be non-differentiable and non-strongly convex (that includes ℓ_1-regularized Huber regression, for instance). We operate under the proportional asymptotics regime, where the sample size n, feature size p, and subsample sizes k_m for $m \in [M]$ diverge while maintaining fixed limiting ratios $n/p \to \delta \in (0, \infty)$ and $k_m/n \to c_m \in (0, 1]$.[1] Our results assume a linear response model $y = \boldsymbol{x}^\top\boldsymbol{\theta} + z$ with isotropic Gaussian features $\boldsymbol{x}$, a random signal $\boldsymbol{\theta}$ with independent coordinates drawn from an arbitrary marginal distribution F_θ, and noise variable z with an arbitrary distribution F_z, both

[1] Through the paper, we refer to n/p as inverse data aspect ratio and p/n as data aspect ratio of the design $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, viewing n as the height and p as the width of the rectangular design matrix $\boldsymbol{X}$.

subject to a mild regularity condition. We summarize our main results below and situate them within the context of recent related work in Table 1.

1.1 Summary of results and paper outline

A summary of our main results along with an outline for the paper is as follows.[2]

(a) **Risk characterization and estimation.** In Section 3, we obtain a precise characterization of the squared risk of subagging of regularized M-estimators under proportional asymptotics (Theorem 2 and Corollary 3). The asymptotic risk is governed by two nonlinear systems of equations (Systems 1a and 1b) that depend on the loss and regularization functions loss, reg, and the limiting subsample ratios c_m for $m \in [M]$ for the component estimators and the limiting inverse data aspect ratio δ. System 1a is a known system that characterizes the asymptotic risk of regularized M-estimators in non-ensemble settings [TAH18] (see Section 1.2 for more details), while System 1b is a new contribution of this paper. Each scalar unknown in the 2-dimensional System 1b is shown to be the fixed-point equation of a contraction (Theorem 1-(2)). This property plays a crucial role in proving the existence and uniqueness of the solution to System 1b. This contraction property is also crucial in establishing the asymptotic behavior of the inner products between estimator errors and their residuals for estimators trained on overlapping subsamples (Theorem 2), leading to the subagged risk asymptotics (Corollary 3). The contractility, along with a ridge smoothing technique, also allows us to maintain weak assumptions on the regularizer, specifically allowing it to be non-strongly convex. Moreover, we also construct a consistent estimator of the ensemble risk (Theorem 4 and Corollary 5), which can be employed for data-adaptive tuning of hyperparameters such as loss functions, regularizers, and subsample sizes.

(b) **Homogeneous ensembles.** In Section 4, we consider homogeneous ensembles (where components are trained on the same loss and reg functions and a common subsample size k). In Section 4.1, we analyze (oracle) optimal ensemble optimal risk with optimal ensemble size $M^\star$ and subsample size $k^\star$. We first establish the monotonicity of the risk with respect to the ensemble size M (Proposition 6), illustrating the benefits of ensembling, which leads to $M^\star = \infty$. We then prove that the risk at the optimal subsample size k_* decreases as the limiting data aspect ratio p/n decreases (Proposition 7). In particular, this implies that the optimally subsampled risk avoids the typical "double (or multiple) descents" observed in regularized M-estimators without subagging. In Section 4.2, we specialize our main result to convex regularized least squares (including ℓ_q-regularized least squares for $q \geq 1$, such as ridge and lasso estimators) and to general M-estimators (including regularized Huber regression). These recover and generalize various known results in the literature (see Section 4.2 for more details).

(c) **Subagging and overparameterization.** In Section 5, we investigate subagging of estimators with vanishing regularization ($\lambda \to 0^+$) and also contrast with estimators with optimal explicit regularization (over $\lambda \geq 0$). Our first insight is that when subagging estimators without any explicit regularization, the optimal subsample size k_* is in the overparameterized regime, regardless of whether the original data aspect ratio p/n is overparameterized. In other words, the optimal subsample size $k^\star$ satisfies $k_* \leq \min\{n, p\}$ in such cases. We verify this for the lassoless (minimum ℓ_1-norm interpolator) ensemble (Figure 5). This highlights the advantages of overparameterization in subagging in that full-ensemble subsampled lassoless can outper-

[2] The source code for experimental verifications in this paper is available at the repository `subagging-asymptotics`. The risk estimator proposed in this paper is also incorporated in the Python library `sklearn-ensemble-cv` [DP24b].

form the optimal lasso on the full data (without any subagging). Our second insight is that the joint optimization of the subsample size and explicit regularization parameter can outperform optimizing explicit regularization alone on the full data. We verify this property for ensembles of the lasso, unregularized Huber, and ℓ_1-regularizer Huber (see Figure 6, 7 and Appendix D). This highlights the benefits of subagging on top of optimal explicit regularization.

Independent results in the non-ensemble setting. In the process of characterizing the risk of the subagging ensemble, we also establish the convergence of certain trace functionals (in particular, see (13) below or the last three rows of Table 2). This implies that the observable adjustments developed in [Bel22] for inference using a single estimator converge to their deterministic counterparts defined as solutions to the System 1a, unifying the mean-field asymptotics featuring System 1a of [TAH18] and the inference results of [Bel22]. This was known only for the lasso [CMW23, Theorem 8] and unregularized M-estimators [BK24]. These convergence results are new for regularized estimators (beyond ridge and lasso) and robust loss functions (beyond squared loss) and are of independent interest even for a single estimator (that is, in the non-ensemble setting with $M = 1$).

1.2 Other related work

Resampling methods, such as bagging and subsampling, are widely used in statistics and machine learning. Given their broad applicability, there is a vast literature on these methods. In this section, we provide an overview of the literature related to the risk analysis of ensemble methods, particularly in high-dimensional regimes that have received considerable recent interest.

Classical work on bagging and subagging includes the work by [Bre96, Bre01, BY02], among others. Beyond bagging, analysis of ensemble methods of different predictors includes smooth weak predictors [BS06, FH07], nonparametric estimators [BY02, LGR$^+$22], and classifiers [HS05, Sam12]. Historically, there are also early works by [SK95, KS97] on risk asymptotics for ridge ensembles under Gaussian features. We also mention here some other early work on ensembles, including: [HS90, Per93, SK95, KS97]. For a comprehensive overview of early work on bagging and ensemble methods in general, we refer readers to [PDK23].

Substantial progress has been made in the last decade in understanding the asymptotic behavior of regularized M-estimators in high-dimensional settings, particularly under the proportional asymptotic regime where the number of features scales with the number of observations. Frameworks of Approximate Message Passing (AMP) (developed in a series of papers [DMM09, DMM11, BM11a]), Convex Gaussian Min-Max Theorem (CGMT) (developed in a series of papers [OTH13, OH16, TOH15, TAH18]), and leave-one-out (LOO) and martingale-based analysis common in random matrix theory (used in [EK13, Kar18], for example) have been instrumental in deriving the limiting test risk, often as solutions to (nonlinear) systems of self-consistent equations. More specifically, these include analyses of unregularized estimators [EKBB$^+$13, EK13, Kar18, DM16, BBEKY13], ridge estimator [DW18], lasso [BM11b, MM21], bridge estimators [WMZ18], logistic regression [SC19, MLC19, SAH19], convex regularized M-estimators [TOH15, TAH18], among others. Recently, triggered by the empirical success of neural networks that interpolate, these risk analyses have been extended to interpolating estimators with vanishing regularization (in the overparameterized regimes that allow for interpolation), such as ridgeless [HMRT22], lassoless [LW21], max-margin interpolators [MRSY19, DKT22, LS22]; see the survey papers [BMR21, Bel21] for other related references.

Beyond individual regularized M-estimators, there has now been considerable interest over the last few years in the analysis of ensembles of estimators in high-dimensional settings, especially in the

overparameterized regime just mentioned. In particular, [LJB20] consider least squares ensembles obtained by subsampling such that the final subsampled dataset has more observations than the number of features. The work of [PDK23] provides the characterization of the asymptotic risk of ensembles of ridge regression using results from Random Matrix Theory (RMT). Furthermore, recent extensions by [DPK23, PD23] expand the scope of these results by establishing risk equivalences for both optimal and suboptimal risks, considering arbitrary feature covariance and signal structures. Other follow-up works for subagging of ridge and ridgeless regression include [CZYS23, AK23]. This paper develops tools to study ensembles of regularized estimators with general loss and regularizers, beyond ridge regularization. For instance, our theory accommodates ℓ_1-regularized Huber regression.

Another line of research focuses on ensemble methods involving random features and feature sketching rather than subsampling. In random features models, the effect of ensembling on various components of the risk has been studied in [AP20, dRBK20, LGR+22]. Recently, [PL24] analyze ensembles of ridge regression with sketched features with asymptotically free sketching. There are also analyses of alternative resampling and averaging techniques. For example, in the context of distributed learning, [DS20, DS21, MRRK22] consider the divide-and-conquer approach, or splagging (split aggregating), and investigate their properties for ridge and ridgeless predictors.

Very recently, [CVD+24] analyzed the limiting equations of several resampling schemes, including bootstrap and resampling without replacement, and characterized self-consistent equations for the limiting bias and variance functionals of estimators obtained by minimization of the negative log-likelihood plus an additive ridge penalty. This is related to our risk characterization as [CVD+24] also covers sampling without replacement, but our nonlinear systems (Systems 1a and 1b) characterizing the subagging risk do not appear explicitly in their work, which instead focuses on self-consistent equations for bias and variance functionals of the specific resampling scheme. The results of [CVD+24] relies on the general AMP analysis and state evolution laid out in [LGR+22, Lemmas B.3 and B.5], generalizing [BM11b]. This analysis requires the existence of unique solution to the corresponding limiting system of equations, which is granted under strong convexity (e.g., with a ridge penalty), but was not established until the present paper for the case of ensembling of subsampled regularized estimators.

Finally, complementary to risk characterization, there has also been considerable interest in the cross-validation and model selection of ensemble methods. In particular, [DPK23] study cross-validation for bagging of ridge regression. [BDK+24] examine the consistency of generalized cross-validation (GCV) for estimating the prediction risk of arbitrary ensembles of regularized least squares estimators for strongly convex penalties. They show that GCV is inconsistent for any finite ensemble of size greater than one and identify a correction to GCV that is consistent for any finite ensemble size, termed corrected GCV (CGCV). In this paper, we generalize one of the data-dependent estimators proposed in [BDK+24] for the general setting of this paper, allowing for general convex losses and heterogeneous component estimators in the ensemble. While we do not attempt to interpret the estimator as a corrected GCV for homogeneous ensembles in the general setting, it may be possible to perform such an analysis further, which we leave for future work.

The proof strategy in this paper extends the approach of [BK24], which studies the bagging of unregularized M-estimators. While their analysis is based on a relatively simple 1-dimensional nonlinear system, the new System 1b below is 2-dimensional, introducing additional complexity to the analysis. The rise in complexity is similar to that from unregularized regression [EK13, DM16, Kar18] and its 2-dimensional system to the 4-dimensional system of regularized M-estimators [TAH18] given in System 1a. One challenge arises from the stochastic control of the trace terms

in (13). In the unregularized case, these trace terms can be approximated by a straightforward product of the norms of the error vector and residuals (see [BK24, Lemma 5.7]), and the stochastic behavior of these norms are well-understood in the existing literature (cf. [TAH18, LGC+21]). However, in the regularized case, the trace functional cannot be approximated by such a simple expression, which prevents the direct application of these existing results based on the CGMT. We overcome this by showing that, with high probability, the trace term is a stationary point of a certain (random) strongly convex function. This allows us to control the perturbation of the trace term through the behavior of the convex function (see Appendix A.4).

1.3 Notation

We denote scalars in the regular lower or upper case (e.g., a, A), vectors in bold lower case (e.g., $\boldsymbol{a}$), and matrices in bold upper case (e.g., $\boldsymbol{A}$). For a natural number n, the shorthand notation $[n]$ denotes the set $\{1, \dots, n\}$. For two real numbers x and y, we use $x \wedge y$ to denote $\min\{x, y\}$. For a vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|_q$ denotes its ℓ_q-norm for $q \geq 1$. If no subscript is present for the norm $\|\boldsymbol{u}\|$ of a vector $\boldsymbol{u}$, then it is assumed to be the ℓ_2 norm of $\boldsymbol{u}$. For a univariate function f and a vector $\boldsymbol{a} \in \mathbb{R}^n$, with a slight overload of notation, we use $f(\boldsymbol{a}) \in \mathbb{R}^n$ to denote the component-wise application of f to $\boldsymbol{a}$. We use $\mathrm{diag}[\boldsymbol{a}]$ to denote the diagonal matrix whose entries are given by the vector $\boldsymbol{a}$. Throughout, we use $\boldsymbol{0}$, $\boldsymbol{1}$, and $\boldsymbol{I}$ to respectively denote the all-zero vector, all-one vector, and identity matrix of varying dimensions, depending on the context.

For any proper, closed, convex function $f\colon \mathbb{R} \to \overline{\mathbb{R}}$, the proximal operator and Moreau envelope of f with a parameter $\tau > 0$ at a point $x \in \mathbb{R}$ are, respectively, denoted as:[3]

$$\mathsf{prox}_f(x;\tau) \coloneqq \operatorname*{argmin}_{y\in\mathbb{R}} f(y) + \frac{1}{2\tau}(x-y)^2 \quad \text{and} \quad \mathsf{env}_f(x;\tau) \coloneqq \min_{y\in\mathbb{R}} f(y) + \frac{1}{2\tau}(x-y)^2. \tag{1}$$

For a proper, closed, convex f, the argmin in (1) exists and is unique, and consequently $x \mapsto \mathsf{prox}_f(x;\tau)$ is a well-defined function. Let ∂f denote the subdifferential of f, which is the set of all subgradients of f. We jot down two key relationships between the proximal operator, subdifferential, and Moreau envelopes of f below for the reader's convenience:

$$\frac{\partial}{\partial x}\mathsf{env}_f(x;\tau) = \frac{x - \mathsf{prox}_f(x;\tau)}{\tau} \in \partial f(\mathsf{prox}_f(x;\tau)). \tag{2}$$

For simplicity, we often use $\mathsf{env}'_f(x;\tau)$ to denote the partial derivative $\frac{\partial}{\partial x}\mathsf{env}_f(x;\tau)$.[4]

Finally, we use $\mathcal{O}_{\mathbb{P}}$ and $o_{\mathbb{P}}$ to denote probabilistic big-O and little-o notation, respectively, while the convergences in probability are denoted by $\xrightarrow{\mathrm{p}}$. Most other notation we use is standard and any other non-standard notation is defined inline. For the reader's convenience, we also give a quick overview of the specific notation used in this paper in Table 4.

2 Setup and assumptions

We consider the standard supervised regression setting, in which we observe n data points $(\boldsymbol{x}_i, y_i)$ for $i \in [n]$. The feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ contains $\boldsymbol{x}_i^\top$ in its i-th row and the response vector $\boldsymbol{y} \in \mathbb{R}^n$ contains y_i in its i-th entry. We assume the following distribution on the dataset $(\boldsymbol{X}, \boldsymbol{y})$:

[3] Here $\overline{\mathbb{R}}$ is the extended real line (that does two-point ($+\infty$ and $-\infty$) compactification of the real line).

[4] In general, for a bivariate function $g(\cdot;\cdot)$, we use the notation $g'(\cdot;\cdot)$ to denote the first partial derivative with respect to the *first* argument.

Assumption A (Data distribution)**.** The distribution of $(\boldsymbol{X}, \boldsymbol{y})$ is specified by:

1. The design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ has i.i.d. entries drawn from $\mathcal{N}(0, 1/p)$.
2. The response vector $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{z}$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a random signal vector and $\boldsymbol{z} \in \mathbb{R}^n$ is a random noise vector, both independent of each other and of the feature matrix $\boldsymbol{X}$, with:
   (a) The signal vector $\boldsymbol{\theta} \in \mathbb{R}^p$ has i.i.d. entries drawn from distribution F_θ.
   (b) The noise vector $\boldsymbol{z} \in \mathbb{R}^n$ has i.i.d. entries drawn from distribution F_z.

We subsample the dataset $(\boldsymbol{X}, \boldsymbol{y})$ to create M subsampled datasets. Towards that end, define M subsample index subsets $I_m \subset [n]$ of cardinality $k_m = |I_m|$ for $m \in [M]$. The feature matrix and response vector associated with the subsampled dataset $(\boldsymbol{x}_i, y_i)$ for $i \in I_m$ are denoted as $(\boldsymbol{X}_{I_m}, \boldsymbol{y}_{I_m})$. We assume the following sampling strategy for the subsample index sets $\{I_m\}_{m \in [M]}$:

Assumption B (Subsampling strategy)**.** Given deterministic integers $\{k_m \geq 1\}_{m \in [M]}$, the M subsample index sets $\{I_m\}_{m \in [M]}$ are independent of $(\boldsymbol{X}, \boldsymbol{y})$ and are independently sampled from the uniform distribution over subsets of $[n]$ with cardinality k_m for each $m \in [M]$.

It is worth noting that if I_m and I_ℓ (for $m \neq \ell$) are any two independent subsample sets of cardinality k_m and k_ℓ per Assumption B, then the cardinality of intersection $|I_m \cap I_\ell|$ follows a hypergeometric distribution with mean $k_m k_\ell / n$. Using the properties of the hypergeometric distribution, it follows that $|I_m \cap I_\ell|/n \xrightarrow{\text{p}} c_m c_\ell$ as both $n, k_m, k_\ell, \to \infty$ with the subsample ratios $k_m/n \to c_m$ and $k_\ell/n \to c_\ell$ for some $c_m, c_\ell \in (0, 1]$ (this follows from Chebyshev's inequality and the variance formula of the hypergeometric distribution, see Section S.8.1 of [PDK23] for more details). Intuitively, each sample lands in a subsample I_m with probability c_m (the limiting ratio $|I_m|/n$) and in the overlap of two subsamples with probability $c_m c_\ell$ (the limiting ratio $|I_m \cap I_\ell|/n$), as the subsamples are drawn independently. The overlap between any two subsample sets I_m and I_ℓ is thus of order n with high probability. The randomness in subsampling in Assumption B is not important. Our results can accommodate deterministic sampling where the subsample sets $\{I_m\}_{m \in [M]}$ are selected deterministically, provided that the ratios $|I_m|/n$, I_ℓ/n, and $|I_m \cap I_\ell|/n$ converge to non-zero constants.

For each subsampled dataset $(\boldsymbol{X}_{I_m}, \boldsymbol{y}_{I_m})$ for $m \in [M]$, we define the regularized M-estimator $\widehat{\boldsymbol{\theta}}_m$ as

$$\widehat{\boldsymbol{\theta}}_m(I_m) \in \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^p} \sum_{i \in I_m} \mathsf{loss}_m(y_i - \boldsymbol{x}_i^\top \boldsymbol{b}) + \sum_{j \in [p]} \mathsf{reg}_m(\boldsymbol{b}_j). \tag{3}$$

When defining (3), we allow the argmin operator to return any one of the minimizers (as emphasized by the element notation in (3)). Here loss_m and reg_m are the loss and regularization functions that satisfy the following assumption for all $m \in [M]$:

Assumption C (Loss and regularizer structure)**.** The loss function $\mathsf{loss}\colon \mathbb{R} \to \overline{\mathbb{R}}$ is proper, closed, convex, and differentiable with derivative loss' Lipschitz continuous, and $\min_x \mathsf{loss}(x) = \mathsf{loss}(0)$. The regularizer $\mathsf{reg}\colon \mathbb{R} \to \overline{\mathbb{R}}$ is proper, closed, convex and $\min_x \mathsf{reg}(x) = \mathsf{reg}(0)$.

The final ensemble estimator, constructed using the component estimators (3), is defined as:

$$\widetilde{\boldsymbol{\theta}}_M\big(\{I_m\}_{m \in [M]}\big) \coloneqq \frac{1}{M} \sum_{m \in [M]} \widehat{\boldsymbol{\theta}}_m(I_m). \tag{4}$$

For brevity, we omit the dependency of the component and ensemble estimators on I_m and $\{I_m\}_{m\in[M]}$ and simply write $\widehat{\boldsymbol{\theta}}_m$ and $\widetilde{\boldsymbol{\theta}}_M$, respectively, when it is clear from the context. We evaluate the performance of the ensemble estimator $\widetilde{\boldsymbol{\theta}}_M$ with respect to the signal parameter $\boldsymbol{\theta}$ via:

$$R_M := \frac{1}{p}\|\widetilde{\boldsymbol{\theta}}_M - \boldsymbol{\theta}\|_2^2.$$

Note that R_M is the (excess) out-of-sample squared error in our setup because of isotropic features: For an independently sampled test feature $\boldsymbol{x}_0 \in \mathbb{R}^p$ with i.i.d. entries drawn from $\mathcal{N}(0, 1/p)$, we have $R_M = \mathbb{E}[(\boldsymbol{x}_0^\top \widetilde{\boldsymbol{\theta}}_M - \boldsymbol{x}_0^\top \boldsymbol{\theta})^2 \mid \boldsymbol{y}, \boldsymbol{X}, \{I_m\}_{m\in[M]}]$. We will refer to R_M as the risk of the ensemble. Observe that R_M is a scalar random variable that depends on both the dataset $(\boldsymbol{X}, \boldsymbol{y})$ and the random samples I_m for $m \in [M]$. The goal of the paper is to characterize the asymptotic behavior of this random variable R_M under the proportional asymptotics regime. In this regime, the sample size n, feature size p, and subsample size k_m all diverge while keeping the appropriate ratios fixed: we will assume the inverse data aspect ratio $n/p \to \delta \in (0, \infty)$ and for each $m \in [M]$, the subsample ratio $k_m/n \to c_m \in (0, 1]$ as $n, p, k_m \to \infty$.

3 Risk characterization and estimation

In this section, we will first describe a general technical result on the correlations of the error and residual vectors for regularized M-estimator in Section 3.1. We then state our general result on the risk characterization of the ensemble estimator in Section 3.3 and construct a consistent risk estimator for this risk in Section 3.4.

3.1 Asymptotics of correlations of estimator and residual errors

To state the risk characterization of the ensemble estimator, we first introduce two important nonlinear systems of equations: Systems 1a and 1b. Intuitively, these systems correspond to the corner cases where the ensemble size $M = 1$ and $M = \infty$, respectively. As we shall see in Section 3.3, these systems completely determine the risk asymptotics of the ensemble estimator (4).

System 1a (Error norms of individual regularized M-estimator)**.** Given a triple $(\mathsf{loss}, \mathsf{reg}, c\delta)$ where $c\delta \in (0, \infty)$ and $\mathsf{loss}, \mathsf{reg} : \mathbb{R} \to \mathbb{R}$ are convex functions, define the following 4-scalar system of equations in variables $(\alpha, \beta, \kappa, \nu)$:

$$\alpha^2 = \mathbb{E}\big[\big(\tfrac{1}{\nu}\mathsf{env}'_{\mathsf{reg}}(\Theta + \tfrac{\beta}{\nu}H; \tfrac{1}{\nu}) - \tfrac{\beta}{\nu}H\big)^2\big] \tag{5a}$$
$$\beta^2 = \mathbb{E}\big[\mathsf{env}'_{\mathsf{loss}}(Z + \alpha G; \kappa)^2\big] \cdot c\delta \tag{5b}$$
$$\kappa\beta = \mathbb{E}\big[\big(\tfrac{1}{\nu}\mathsf{env}'_{\mathsf{reg}}(\Theta + \tfrac{\beta}{\nu}H; \tfrac{1}{\nu}) - \tfrac{\beta}{\nu}H\big) \cdot (-H)\big] \tag{5c}$$
$$\nu\alpha = \mathbb{E}\big[\mathsf{env}'_{\mathsf{loss}}(Z + \alpha G; \kappa) \cdot G\big] \cdot c\delta \tag{5d}$$

where $H \sim \mathcal{N}(0, 1)$, $G \sim \mathcal{N}(0, 1)$, $\Theta \sim F_\theta$, $Z \sim F_z$, all mutually independent.

System 1a can be found in the literature, specifically in [TAH18, Equation 15]. To be precise, we are applying the result of [TAH18] on the subsample estimator (3) using $k = |I_m|$ observations with $k/n \to c$, so that the limiting inverse aspect ratio $k/p = k/n \cdot n/p \to c\delta$. System 1a is known to characterize the limit in probability of the risk of (3) when $|I_m|/p \to c\delta$: if $(\alpha, \beta, \kappa, \nu)$

is a solution to System 1a, then $p^{-1}\|\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2 \xrightarrow{\text{p}} \alpha^2$. The existence and uniqueness of the fixed-point parameters in this system are central to applying results from the Convex Gaussian Min-Max Theorem (CGMT) to derive precise risk characterization for regularized M-estimator (under proportional asymptotics). This is guaranteed under conditions where both loss and reg are Lipschitz and the problem parameters are such that the perfect signal recovery is not possible, leading to non-zero asymptotic risk. For a detailed discussion on these conditions, see [BK23b]. Next, we describe our second system for risk characterization of the ensemble estimators.

System 1b (Error correlations of overlapped regularized M-estimator)**.** Given $c, \widetilde{c} \in (0, 1]$ and convex pairs of functions $(\mathsf{loss}, \mathsf{reg})$, $(\widetilde{\mathsf{loss}}, \widetilde{\mathsf{reg}})$, let $(\alpha, \beta, \kappa, \nu) \in \mathbb{R}^4_{>0}$ and $(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\kappa}, \widetilde{\nu})$ be parameters that satisfy System 1a with $(\mathsf{loss}, \mathsf{reg}, c\delta)$ and $(\widetilde{\mathsf{loss}}, \widetilde{\mathsf{reg}}, \widetilde{c}\delta)$, respectively. Define the following 2-scalar system of equations in variable $(\eta_G, \eta_H) \in [-1, 1]^2$:

$$\begin{bmatrix} \eta_G \\ \eta_H \end{bmatrix} = \begin{bmatrix} F_{\mathsf{reg}}(\eta_H) \\ F_{\mathsf{loss}}(\eta_G) \end{bmatrix} \tag{6}$$

where $F_{\mathsf{loss}}, F_{\mathsf{reg}} \colon [-1, 1] \to \mathbb{R}$ are functions defined as:

$$F_{\mathsf{loss}}(\eta_G) \coloneqq \sqrt{c\widetilde{c}} \cdot \frac{\mathbb{E}\big[\mathsf{env}'_{\mathsf{loss}}(Z + \alpha G; \kappa) \cdot \mathsf{env}'_{\widetilde{\mathsf{loss}}}(Z + \widetilde{\alpha}\widetilde{G}; \widetilde{\kappa})\big]}{\mathbb{E}\big[\mathsf{env}'_{\mathsf{loss}}(Z + \alpha G; \kappa)^2\big]^{1/2} \cdot \mathbb{E}\big[\mathsf{env}'_{\widetilde{\mathsf{loss}}}(Z + \widetilde{\alpha}\widetilde{G}; \widetilde{\kappa})^2\big]^{1/2}} \tag{7a}$$

$$F_{\mathsf{reg}}(\eta_H) \coloneqq \frac{\mathbb{E}\big[\big(\frac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\big(\Theta + \frac{\beta}{\nu}H; \frac{1}{\nu}\big) - \frac{\beta}{\nu}H\big) \cdot \big(\frac{1}{\widetilde{\nu}} \cdot \mathsf{env}'_{\widetilde{\mathsf{reg}}}\big(\Theta + \frac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}; \frac{1}{\widetilde{\nu}}\big) - \frac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}\big)\big]}{\mathbb{E}\big[\big(\frac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\big(\Theta + \frac{\beta}{\nu}H; \frac{1}{\nu}\big) - \frac{\beta}{\nu}H\big)^2\big]^{1/2} \cdot \mathbb{E}\big[\big(\frac{1}{\widetilde{\nu}} \cdot \mathsf{env}'_{\widetilde{\mathsf{reg}}}\big(\Theta + \frac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}; \frac{1}{\widetilde{\nu}}\big) - \frac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}\big)^2\big]^{1/2}} \tag{7b}$$

where $\begin{pmatrix} G \\ \widetilde{G} \end{pmatrix} \sim \mathcal{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \eta_G \\ \eta_G & 1 \end{bmatrix}\Big)$, $\begin{pmatrix} H \\ \widetilde{H} \end{pmatrix} \sim \mathcal{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \eta_H \\ \eta_H & 1 \end{bmatrix}\Big)$, $\Theta \sim F_\theta$, $Z \sim F_z$, all mutually independent.

System 1b is new and one of the main contributions of this paper. Note that the parameters (η_G, η_H) in the system are correlation parameters (up to scaling factors) of the two random variables visible inside expectations in (7a) and of the two random variables in (7b), respectively. By the Cauchy–Schwarz inequality, the function F_{loss} and F_{reg} are uniformly bounded as $|F_{\mathsf{loss}}(\eta)| \le \sqrt{c\widetilde{c}}$ and $|F_{\mathsf{reg}}(\eta)| \le 1$ so that any solution (η_G, η_H) to the system (6) lies in the set $[-1, 1] \times [-\sqrt{c\widetilde{c}}, \sqrt{c\widetilde{c}}]$. As stated, it is not immediately clear if System 1b admits any solution and whether it is unique. Our first result establishes that this is indeed the case:

Theorem 1 (Existence, uniqueness, and sign pattern of solutions to System 1b)**.** The functions F_{loss} and F_{reg} defined in (7) satisfy the following properties:

1. $|F_{\mathsf{loss}}(\eta_G)| < \sqrt{c\widetilde{c}}$ for all $|\eta_G| < 1$ and $|F_{\mathsf{reg}}(\eta_H)| < 1$ for all $|\eta_H| < 1$.
2. F_{loss} and F_{reg} are non-decreasing, differentiable, and the compositions $F_{\mathsf{loss}} \circ F_{\mathsf{reg}}$ and $F_{\mathsf{reg}} \circ F_{\mathsf{loss}}$ are $\min\{c, \widetilde{c}\}$-Lipschitz.
3. If $\min\{c, \widetilde{c}\} < 1$, then System 1b admits a unique solution $(\eta_G^\star, \eta_H^\star) \in (-1, 1) \times (-\sqrt{c\widetilde{c}}, \sqrt{c\widetilde{c}})$.

4. The signs of the solution $(\eta_G^\star, \eta_H^\star)$ are characterized by the following sign pattern:

$$\operatorname{sign}\left(\begin{bmatrix}\eta_G^\star \\ \eta_H^\star\end{bmatrix}\right) = \operatorname{sign}\left(\begin{bmatrix}F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(0) \\ F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(0)\end{bmatrix}\right) \tag{8}$$

where $\operatorname{sign}(x) := \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x<0\}}$ applies component-wise.

Some remarks on Theorem 1 are in order. Among the four properties listed in Theorem 1, the most interesting is the second property: the two maps $(F_{\mathsf{loss}} \circ F_{\mathsf{reg}},\ F_{\mathsf{reg}} \circ F_{\mathsf{loss}})$ are strict contractions. Given this property, the third property (the uniqueness and existence of the solution) easily follows from the Banach fixed-point theorem. We briefly explain this next. Indeed, if $(\eta_H^\star, \eta_G^\star)$ is a solution to System 1b, then $\eta_H^\star$ automatically satisfies the following 1-dimensional fixed-point equation:

$$\eta_H^\star = F_{\mathsf{loss}}(\eta_G^\star) = F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H^\star).$$

The other direction is also true in the following sense: if $\eta_H^\star$ is a solution to the fixed-point equation $\eta_H = F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H)$, then letting $\eta_G^\star = F_{\mathsf{reg}}(\eta_H^\star)$, we observe that the pair $(\eta_H^\star, \eta_G^\star)$ satisfies System 1b. Since $F_{\mathsf{loss}} \circ F_{\mathsf{reg}}$ is a contraction mapping, the Banach fixed-point theorem implies that such $\eta_H^\star$ uniquely exists. See Appendix A.1 for the full proof details. This contraction property also certifies that the fixed-point iteration algorithm $\eta_H^{(k+1)} = F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H^{(k)})$, which we use in our experiments to solve System 1b, numerically converges to the correct solution $\eta_H^\star$ exponentially fast.

Since F_{loss} and F_{reg} are non-decreasing, combined with the fourth property in Theorem 1, we get a simple sufficient condition which determines the sign of $(\eta_H^\star, \eta_G^\star)$:

$$F_{\mathsf{loss}}(0) \text{ and } F_{\mathsf{reg}}(0) \text{ are non-negative} \Rightarrow \eta_H^\star \text{ and } \eta_G^\star \text{ are non-negative.}$$

Simplifying the denominators of F_{reg} and F_{loss} by (5a) and (5b), we can write $F_{\mathsf{loss}}(0)$ and $F_{\mathsf{reg}}(0)$ as follows:

$$F_{\mathsf{loss}}(0) = \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}} \cdot \mathbb{E}\Big[\mathbb{E}\big[\mathsf{env}'_{\mathsf{loss}}(Z + \alpha G; \kappa) \mid Z\big] \cdot \mathbb{E}\big[\mathsf{env}'_{\widetilde{\mathsf{loss}}}(Z + \widetilde{\alpha} G; \widetilde{\kappa}) \mid Z\big]\Big],$$
$$F_{\mathsf{reg}}(0) = \frac{1}{\nu\widetilde{\nu}\alpha\widetilde{\alpha}} \cdot \mathbb{E}\Big[\mathbb{E}\big[\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\beta}{\nu} H; \tfrac{1}{\nu}\big) \mid \Theta\big] \cdot \mathbb{E}\big[\mathsf{env}'_{\widetilde{\mathsf{reg}}}\big(\Theta + \tfrac{\widetilde{\beta}}{\widetilde{\nu}} H; \tfrac{1}{\widetilde{\nu}}\big) \mid \Theta\big]\Big].$$

In particular, if the same loss and regularizer are used, i.e., $\mathsf{loss} = \widetilde{\mathsf{loss}}$ and $\mathsf{reg} = \widetilde{\mathsf{reg}}$, and the subsample sizes are the same, i.e., $k = \widetilde{k}$, then the solutions satisfying System 1b are same, i.e., $(\alpha, \beta, \kappa, \nu) = (\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\kappa}, \widetilde{\nu})$, so that it is easy to see from the above formula that $F_{\mathsf{loss}}(0) \geq 0$ and $F_{\mathsf{reg}}(0) \geq 0$. This means that the solutions $(\eta_G^\star, \eta_H^\star)$ are non-negative when the same loss, regularizer, and subsample size are used.

Another case such that the solutions $(\eta_G^\star, \eta_H^\star)$ are positive is when loss and $\widetilde{\mathsf{loss}}$ are least squares and reg and $\widetilde{\mathsf{reg}}$ are ridge (but possibly different regularization parameters). This is because $\mathsf{env}'_f; (x; \tau)$ is linear in x for any squared loss (and regularizer) of the form $f(\cdot) = \lambda(\cdot)^2$ so that $F_{\mathsf{loss}}(0) = C_1 \mathbb{E}[Z^2] \geq 0$ and $F_{\mathsf{reg}}(0) = C_2 \mathbb{E}[\Theta^2] \geq 0$ for some positive constants C_1, C_2.

Remark 1 (Negative estimator error correlation)**.** Let us take $\mathsf{reg}, \widetilde{\mathsf{reg}}$ as the indicator functions

$$\mathsf{reg}(x) = \Pi_{(-\infty, -t]}(x) := \begin{cases} +\infty & x > -t \\ 0 & x \leq -t \end{cases} \quad \text{and} \quad \widetilde{\mathsf{reg}}(x) = \Pi_{[\widetilde{t}, +\infty)}(x) := \begin{cases} +\infty & x < \widetilde{t} \\ 0 & x \geq \widetilde{t} \end{cases}$$

where $t, \widetilde{t} \geq 0$ are non-negative constants. Note that the two sets, $(-\infty, -t]$ and $[\widetilde{t}, +\infty)$, are disjoint. Noting $\mathsf{prox}_{\mathsf{reg}}(x) = \min\{x, -t\}$ and $\mathsf{prox}_{\widetilde{\mathsf{reg}}}(x) = \max\{x, \widetilde{t}\}$, we have

$$F_{\mathsf{reg}}(\eta_H) = \frac{1}{\nu\widetilde{\nu}\alpha\widetilde{\alpha}}\mathbb{E}\big[\big(\min\big\{\Theta + \tfrac{\beta}{\nu}H, -t\big\} - \Theta\big) \cdot \big(\max\big\{\Theta + \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}, \widetilde{t}\big\} - \Theta\big)\big].$$

Thus, if Θ is included in the closed set $[-t, \widetilde{t}]$ with probability 1, then we have $F_{\mathsf{reg}}(\eta_H) \leq 0$ for all η_H so that $\eta_G^\star = F_{\mathsf{reg}}(\eta_H^\star)$ is non-positive. This is intuitive as we will show in Theorem 2 that η_G characterizes the limiting behavior of the correlations between two estimators trained on reg and $\widetilde{\mathsf{reg}}$.

We next show that the correlation parameters (η_G, η_H) from System 1b are the limiting correlations between the estimator and residual errors of estimators trained on overlapped samples. To do so, besides Assumptions A–C, we will need mild regularity conditions that the loss and reg in Assumption C need to satisfy in relation to the distribution F_θ and F_z of the signal and noise coordinates in Assumption A.

Assumption D (Regularity conditions)**.** Let $\Theta \sim F_\theta$ and $Z \sim F_z$ be the signal and noise random variables as in Assumption A, and $G \sim \mathcal{N}(0,1)$, $H \sim \mathcal{N}(0,1)$, all mutually independent. In addition to Assumption C, the functions $\mathsf{loss} : \mathbb{R} \to \overline{\mathbb{R}}$ and $\mathsf{reg} : \mathbb{R} \to \overline{\mathbb{R}}$ satisfy the following:

1. For all $c \in \mathbb{R}$, we have
$$\mathbb{E}[\mathsf{loss}'_+(cG + Z)^2] < +\infty \quad \text{and} \quad \mathbb{E}[\mathsf{reg}'_+(cH + \Theta)^2] < +\infty$$
where we define $f'_+(x) := \sup_{s \in \partial f(x)} |s|$ for any convex function f.
2. $\mathbb{P}(\Theta \neq 0) > 0$.
3. System 1a admits a unique positive solution $(\alpha, \beta, \kappa, \nu) \in \mathbb{R}^4_{>0}$.
4. There exists interval $\mathcal{I} \subset \mathbb{R}$ where loss' is strictly increasing. For each $z \in \mathbb{R}$, the measure $p_Z(z)$ of Z is either a Dirac delta function or it is continuous.

The conditions in Assumption D are similar to those assumed in [TAH18] when characterizing the asymptotics for the non-ensemble case. The main difference is that we do not require the second moment of Θ to be finite. These conditions ensure that the individual estimator and residual error norms converge, i.e., $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2/p \xrightarrow{\text{p}} \alpha^2$ and $\|\mathsf{loss}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})\|_2^2/p \xrightarrow{\text{p}} \beta^2$ hold, where α and β are solutions to System 1a. (See Appendices A.5 and A.6 for proofs of convergences of error vector and loss gradient norm squared under the relaxation of the conditions.)

For the upcoming statement, recall that when used on a vector, the loss and reg functions are assumed to be operated component-wise. In addition, we denote the feature matrix and response vector associated with the "overlapped" dataset $(\boldsymbol{x}_i, y_i)$ for $i \in I \cap \widetilde{I}$ using $(\boldsymbol{X}_{I\cap\widetilde{I}}, \boldsymbol{y}_{I\cap\widetilde{I}})$.

Theorem 2 (Estimator and residual error correlation characterization)**.** Let $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$ be component estimators (3) trained on subsamples $(\boldsymbol{X}_I, \boldsymbol{y}_I)$ and $(\boldsymbol{X}_{\widetilde{I}}, \boldsymbol{y}_{\widetilde{I}})$ corresponding to index sets I and $\widetilde{I}$ with parameters $(\mathsf{loss}, \mathsf{reg})$ and $(\widetilde{\mathsf{loss}}, \widetilde{\mathsf{reg}})$. Under Assumptions A–D, as $n, p, k, \widetilde{k} \to \infty$ with $n/p \to \delta \in (0, \infty)$, $k/n \to c \in (0, 1]$ and $\widetilde{k}/n \to \widetilde{c} \in (0, 1]$ with $\min\{c, \widetilde{c}\} < 1$, we have

$$\begin{aligned} p^{-1}(\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta})^\top(\widehat{\boldsymbol{\theta}}_{\widetilde{I}} - \boldsymbol{\theta}) &\xrightarrow{\text{p}} \eta_G\alpha\widetilde{\alpha} \\ p^{-1}\mathsf{loss}'(\boldsymbol{y}_{I\cap\widetilde{I}} - \boldsymbol{X}_{I\cap\widetilde{I}}\widehat{\boldsymbol{\theta}}_I)^\top \widetilde{\mathsf{loss}}'(\boldsymbol{y}_{I\cap\widetilde{I}} - \boldsymbol{X}_{I\cap\widetilde{I}}\widehat{\boldsymbol{\theta}}_{\widetilde{I}}) &\xrightarrow{\text{p}} \eta_H\beta\widetilde{\beta}, \end{aligned} \tag{10}$$

where (η_G, η_H) is the solution to System 1b. Furthermore, for any $i \in I \cap \widetilde{I}$ and $j \in [p]$, there exists a jointly normal $(G_i, \widetilde{G}_i)$ with mean 0, variance 1 and correlation η_G and $(H_j, \widetilde{H}_j)$ with mean 0, variance 1 and correlation η_H (as in System 1b) such that the residuals and estimators are jointly approximated as follows:

$$\max_{j\in[p]} \mathbb{E}\left[1 \ \wedge \ \left\| \begin{pmatrix} \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}_I \\ \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}_{\widetilde{I}} \end{pmatrix} - \begin{pmatrix} \mathsf{prox}_{\mathsf{reg}}(\theta_j + \frac{\beta}{\nu} H_j; \frac{1}{\nu}) \\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + \frac{\beta}{\widetilde{\nu}} \widetilde{H}_j; \frac{1}{\widetilde{\nu}}) \end{pmatrix} \right\|_2^2 \right] = o(1), \tag{11a}$$

$$\max_{i\in[n]} \mathbb{E}\left[\mathbb{1}_{i\in I\cap\widetilde{I}} \ \wedge \ \left\| \begin{pmatrix} y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}_I \\ y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}_{\widetilde{I}} \end{pmatrix} - \begin{pmatrix} \mathsf{prox}_{\mathsf{loss}}(z_i + \alpha G_i; \kappa) \\ \mathsf{prox}_{\widetilde{\mathsf{loss}}}(z_i + \widetilde{\alpha} \widetilde{G}_i; \widetilde{\kappa}) \end{pmatrix} \right\|_2^2 \right] = o(1). \tag{11b}$$

The proof is given in Appendix A.2.4. Put in words, η_G and η_H from System 1b encode the cosines of the angles between the estimator errors and loss gradient residual errors of estimators $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$ respectively. It is worth noting that the proximal representations in (11a) and (11b) allow one to provide limiting behavior of other functionals of the estimator and residual errors by assuming further moments on the signal and error distributions F_θ and F_z; for example, we can characterize the correlation between the raw residuals (rather than after applying the loss derivative) assuming finite second moment of F_z or consider pseudo-Lipschitz functionals other than squared error.

The main difficulty in showing Theorem 2 is the non-trivial dependence between the two estimators $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$ as they share the samples $\boldsymbol{X}_{I\cap\widetilde{I}}$. In the case of squared loss and ridge regularizer, the estimators $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$ have closed-form expressions, and prior work in [PDK23] explicitly analyze the overlapped resolvents by developing conditional calculus of resolvents. However, for general loss and regularizers, the overlapped terms are more challenging to analyze due to the lack of closed-form expressions. Our strategy in this paper is to exploit the recently developed technique in [BK24] to analyze the overlapped terms.

To prove Theorem 2, we show that the left-hand side of (10) concentrate around scalars independent of $\boldsymbol{X}$, and that these two scalars are approximate fixed-point of $F_{\mathsf{loss}} \circ F_{\mathsf{reg}}$ and $F_{\mathsf{reg}} \circ F_{\mathsf{loss}}$, respectively. This strategy of first proving the concentration of certain quantities and then obtaining approximate fixed-point equations is reminiscent of the leave-one-out analysis of [EKBB+13, Kar18] and was previously used in [BK24] to characterize the ensemble risk in unregularized M-estimators (with no explicit penalty). The setting studied here is significantly more complicated than these works due to the presence of robust loss functions, penalty functions, and shared samples between $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$.

We believe that once the contractions of $F_{\mathsf{loss}} \circ F_{\mathsf{reg}}$ and $F_{\mathsf{reg}} \circ F_{\mathsf{loss}}$ have been found and the existence and uniqueness of the solution to System 1b has been established, different techniques than those used here and discussed in the previous paragraph could also be used to derive asymptotic results similar to Theorem 2. For instance, after existence and uniqueness of the solution to System 1b is established, there is hope to carry out an AMP analysis for matrix-valued parameters (see for instance [LGR+22, Lemmas B.3 and B.5] or [JM13, GB23]), or by using the conditional CGMT technique of [CM24, Appendix F]. However, we emphasize that these alternate techniques would also first require to establish the structure of System 1b (as done in Theorem 1) in order to guarantee the existence and uniqueness of the solution to System 1b, since such existence and uniqueness result is required for both applying CGMT results and ensuring the convergence of AMP to the regularized M-estimator.

Even though we assume the existence and uniqueness of the solution to System 1a, such existence

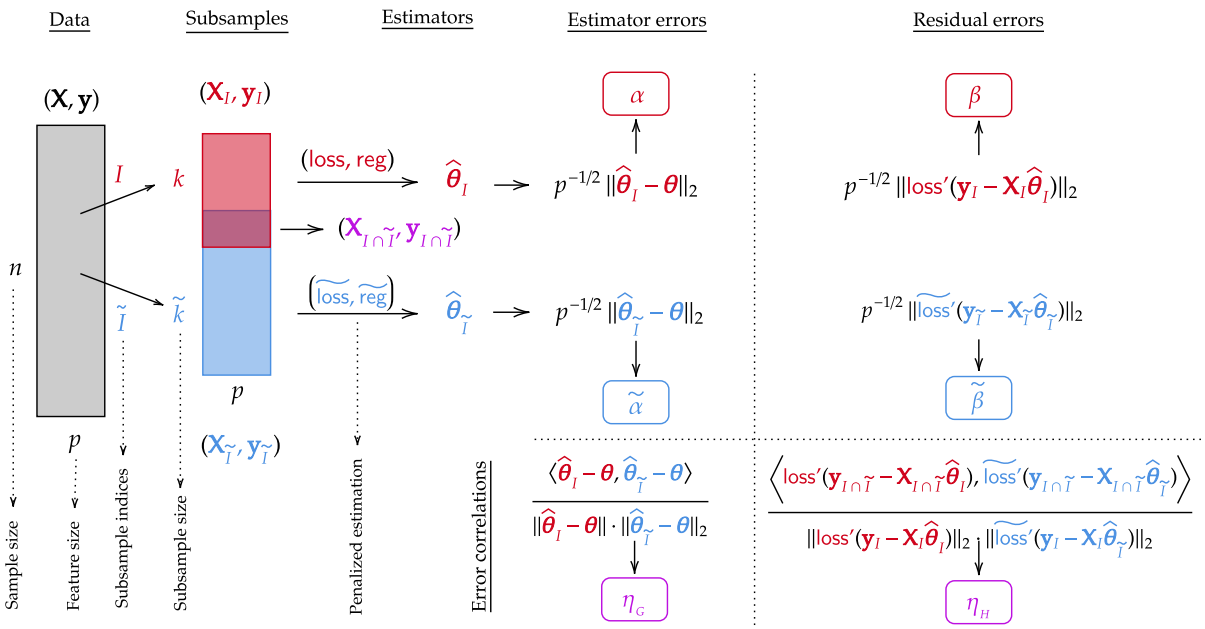

Figure 1: Illustration of the asymptotics of the estimator and residual error norms and correlations of overlapped regularized M-estimator.

and uniqueness have been established under slightly stronger assumptions in [BK23b], namely, for Lipschitz loss and regularizer. We conjecture that the assumptions of Theorem 2 (in particular, Assumption D) are sufficient for existence and uniqueness of a solution to System 1a without requiring a Lipschitz condition on the loss and penalty. We believe that extending the analysis [BK23b] to relax the Lipschitz assumption is a good starting point for this conjecture.

3.2 Interpretation of the parameters in Systems 1a and 1b

As mentioned earlier, the six parameters $(\alpha, \beta, \kappa, \nu)$ and (η_G, η_H) in Systems 1a and 1b essentially characterize the asymptotic risk of the ensemble estimator. These deterministic parameters are limits of various stochastic (observable) quantities that we now describe (see also Figure 1 for a visual illustration).

Here $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$ are the component estimators (3) trained on subsamples $(\boldsymbol{X}_I, \boldsymbol{y}_I)$ and $(\boldsymbol{X}_{\widetilde{I}}, \boldsymbol{y}_{\widetilde{I}})$ corresponding to index sets I and $\widetilde{I}$ and parameters $(\mathsf{loss}, \mathsf{reg})$ and $(\widetilde{\mathsf{loss}}, \widetilde{\mathsf{reg}})$, respectively. We further define the scalar df_I and the matrix $\boldsymbol{V}_I$ by

$$\mathsf{df}_I := \mathrm{tr}[(\partial/\partial \boldsymbol{y}_I)\boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I], \qquad \boldsymbol{V}_I := (\partial/\partial \boldsymbol{y}_I)\mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I) \in \mathbb{R}^{|I|\times|I|} \tag{12}$$

and similarly for $\widetilde{I}$. Two scalars of interest, that relates the behavior of $\widehat{\boldsymbol{\theta}}_I$ to the scalars (κ, ν) in System 1a, are df_I and $\mathrm{tr}[\boldsymbol{V}_I]$.

Assuming squared loss, $\mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I) = \boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I$ is simply the residual vector, and the matrix $\boldsymbol{V}_I$ simplifies to $\boldsymbol{V}_I = \boldsymbol{I} - (\partial/\partial \boldsymbol{y}_I)\boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I$, so that $\mathrm{tr}[\boldsymbol{V}_I] = n - \mathsf{df}_I$. That is, for the square loss case these quantities can all be related to the usual notion of effective degrees of freedom [Ste81]. The matrix $(\partial/\partial \boldsymbol{y}_I)\boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I$ is usually referred to as the "hat" or "smoothing" matrix (for linear smoothers), whose trace is the effective degrees of freedom.

Table 2: **Interpretations of various limiting quantities appearing in Systems 1a and 1b.** See Section 3.2 for definitions and notations.

Interpretation	Stochastic quantity	Limit
Error vector norm squared	$\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}\|_2^2/p$	α^2
Loss gradient norm squared	$\|\mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I)\|_2^2/p$	β^2
Inner product of error vectors	$(\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta})^\top(\widehat{\boldsymbol{\theta}}_{\widetilde{I}} - \boldsymbol{\theta})/p$	$\eta_G\alpha\widetilde{\alpha}$
Inner product of loss gradients	$\mathsf{loss}'(\boldsymbol{y}_{I\cap\widetilde{I}} - \boldsymbol{X}_{I\cap\widetilde{I}}\widehat{\boldsymbol{\theta}}_I)^\top\widetilde{\mathsf{loss}}'(\boldsymbol{y}_{I\cap\widetilde{I}} - \boldsymbol{X}_{I\cap\widetilde{I}}\widehat{\boldsymbol{\theta}}_{\widetilde{I}})/p$	$\eta_H\beta\widetilde{\beta}$
Degrees of freedom	df_I/p where $\mathsf{df}_I \coloneqq \mathrm{tr}[(\partial/\partial\boldsymbol{y}_I)\boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I]$	$\nu\kappa$
Residual degrees of freedom	$\mathrm{tr}[\boldsymbol{V}_I]/p$ where $\boldsymbol{V}_I \coloneqq (\partial/\partial\boldsymbol{y}_I)\mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}})$	ν
Generalized resolvent trace	$\mathrm{tr}[(\boldsymbol{X}_I^\top\mathrm{diag}[\mathsf{loss}''(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I)]\boldsymbol{X}_I + \mathrm{diag}[\mathsf{reg}''(\widehat{\boldsymbol{\theta}}_I)])^{-1}]$ if reg is twice differentiable and $\mathsf{df}_I/\mathrm{tr}[\boldsymbol{V}_I]$ if reg is non-smooth	κ

If loss is no the squared loss, but loss' is 1-Lipschitz (as in the Huber loss or several robust regression losses), the quantities $\mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}_I})$ is still related to a notion of residual vector, and $\mathrm{tr}[\boldsymbol{V}_I]$ is still related to a notion of degrees of freedom. By [Bel23, Lemma 9.1], the estimator $\widehat{\boldsymbol{\theta}}_I$ is the first part of a solution $(\widehat{\boldsymbol{\theta}}_I, \widehat{\boldsymbol{u}})$ to the convex optimization problem

$$\min_{\boldsymbol{b}\in\mathbb{R}^p, \boldsymbol{u}\in\mathbb{R}^{|I|}} \|\boldsymbol{y}_I - \boldsymbol{X}_I\boldsymbol{b} - \boldsymbol{u}\|_2^2 + \sum_{j=1}^p \mathsf{reg}(b_j) + \sum_{i\in I} h(u_i),$$

where $h : \mathbb{R} \to \mathbb{R}$ is a deterministic convex function related to loss. The interpretation of this new optimization problem is that in the presence of heavy-tailed errors or outliers in some components of $\boldsymbol{y}_I$, we add additional variables $(u_i)_{i\in I}$ to fit those outliers. As an example, for the Huber loss, $h(\cdot)$ is proportional to the absolute value. The solution satisfies $\mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}_I}) = \boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I - \widehat{\boldsymbol{u}} = \boldsymbol{y}_I - [\boldsymbol{X} \mid \boldsymbol{I}_I](\widehat{\boldsymbol{\theta}}_I^\top \mid \widehat{\boldsymbol{u}}^\top)^\top$. That is, $\mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}_I})$ is the residual vector of the optimization problem with enlarged design matrix $[\boldsymbol{X}_I \mid \boldsymbol{I}_I] \in \mathbb{R}^{|I|\times(|I|+p)]}$. Consequently, $\mathrm{tr}[\boldsymbol{V}_I]$ equals $|I|$ minus the effective degrees-of-freedom of the estimate $(\widehat{\boldsymbol{\theta}}_I^\top, \widehat{\boldsymbol{u}}^\top)$ fitted using this enlarged design matrix. With this in mind, we refer in Table 2 to $\mathrm{tr}[\boldsymbol{V}_I]$ as residual degrees of freedom in general, and robust residual degrees of freedom for the special case of the Huber loss.

Another interpretation of the matrix $\boldsymbol{V}_I$ in (12) is the Hessian, with respect to $\boldsymbol{y}_I$, of the objective value (3) at $\widehat{\boldsymbol{\theta}}_I$. More precisely, with

$$F(\boldsymbol{y}_I) = \sum_{i\in I} \mathsf{loss}(y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\theta}}_I) + \sum_{j\in[p]} \mathsf{reg}((\widehat{\boldsymbol{\theta}}_I)_j)$$

being the objective value at the minimizer, the envelope theorem gives $(\partial/\partial y_i)F(\boldsymbol{y}_I) = \mathsf{loss}'(y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\theta}}_I)$. Differentiating once more reveals that $\boldsymbol{V}_I$ in (12) is the Hessian of $F(\cdot)$ at $\boldsymbol{y}_I$ and $\mathrm{tr}[\boldsymbol{V}_I]$ is the Laplacian. Since partial minimization preserves convexity and the objective function in (3) is jointly convex in $(\boldsymbol{b}, \boldsymbol{y})$, the function $F(\cdot)$ is convex. This interpretation explains why $\boldsymbol{V}_I$ is a positive semi-definite matrix in cases where closed form expressions for $\boldsymbol{V}_I$ are available (see Table 5).

The convergence of the estimator and residual error norms (first two rows of Table 2) is proved in in [TAH18, CMW23, LGC+21] using the CGMT. Convergence of the corresponding inner products (third, fourth row of Table 2) is novel and established in Theorem 2. The convergence

$$\mathrm{tr}[\boldsymbol{V}_I]/p \xrightarrow{\mathrm{p}} \nu, \qquad \mathsf{df}_I/p \xrightarrow{\mathrm{p}} \nu\kappa, \qquad \mathsf{df}_I/\mathrm{tr}[\boldsymbol{V}_I] \xrightarrow{\mathrm{p}} \kappa, \tag{13}$$

was so far only known for the lasso [CMW23, Theorem 8] or the square loss [Bel23, Corollary 3.2]. To our knowledge, the present paper is the first to establish the above convergence in probability for regularized estimators and robust loss functions beyond the square loss. The proof is given in Appendix A.4.

Assuming twice differentiable loss and reg functions, the parameter κ is also the limiting trace of a resolvent-like matrix $\boldsymbol{A}_I \coloneqq (\boldsymbol{X}_I^\top \mathrm{diag}[\mathsf{loss}''(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I)]\boldsymbol{X}_I + \mathrm{diag}[\mathsf{reg}''(\widehat{\boldsymbol{\theta}}_I)])^{-1}$, which in the further special case of square loss and squared regularizer (with regularization level λ) simplifies to the standard ridge resolvent: $\boldsymbol{A}_I = (\boldsymbol{X}_I^\top \boldsymbol{X}_I + \lambda \boldsymbol{I})^{-1}$. We refer to $\boldsymbol{A}_I$ as the generalized resolvent for convenience in Table 2.

3.3 Asymptotics of ensemble risk

Using the parameters in Systems 1a and 1b, we are now ready for our main result on the squared risk asymptotics of the ensemble estimator. Observe that the squared risk of the ensemble estimator $\widetilde{\boldsymbol{\theta}}_M = \frac{1}{M}\sum_{m\in[M]} \widehat{\boldsymbol{\theta}}_m$ can be decomposed into two terms:

$$\frac{1}{p}\big\|\widetilde{\boldsymbol{\theta}}_M - \boldsymbol{\theta}\big\|_2^2 = \frac{1}{M^2}\sum_{m\in[M]} \frac{1}{p}\|\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}\|_2^2 + \frac{1}{M^2}\sum_{\substack{m,\ell\in[M]\\ m\neq\ell}} \frac{1}{p}(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta})^\top(\widehat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}).$$

Noting $\|\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}\|_2^2/p \xrightarrow{\mathrm{p}} \alpha_m^2$ and applying Theorem 2 to the cross term $(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta})^\top(\widehat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta})/p$ for each $m \neq \ell$, we arrive at the following result:

Corollary 3 (General ensemble risk characterization)**.** Suppose the assumptions of Theorem 2 hold. For $m \in [M]$, let α_m be the parameter satisfying System 1a. For $m, \ell \in [M]$, let $\eta_G(m,\ell)$ be the parameter satisfying System 1b. Then, as $n, p, k \to \infty$ with $n/p \to \delta \in (0,\infty)$ and $k_m/n \to c_m \in (0,1]$, we have

$$\frac{1}{p}\big\|\widetilde{\boldsymbol{\theta}}_M - \boldsymbol{\theta}\big\|_2^2 \xrightarrow{\mathrm{p}} \mathcal{R}_M \coloneqq \frac{1}{M^2}\sum_{m\in[M]} \alpha_m^2 + \frac{1}{M^2}\sum_{\substack{m,\ell\in[M]\\ m\neq\ell}} \eta_G(m,\ell)\cdot\alpha_m\alpha_\ell. \tag{14}$$

Since the parameters α_m and $\eta_G(m,\ell)$ implicitly depend on δ and c_m, the asymptotic risk $\mathcal{R}_M$ also implicitly depends on these parameters. For brevity, we will simply write $\mathcal{R}_M$ unless we wish to explicitly point out this dependence. The factor η_G captures the benefit of ensembling. It is worth noting that a negative η_G (which intuitively corresponds to a component that does better in a different direction) will improve the ensemble risk if the components themselves also have small risks. This aligns with the higher level intuition in ensembling that one should ensemble predictors that each does well, preferably on different parts of the input space.

3.4 Risk estimation

The risk characterization in Corollary 3 depends on the population-level characteristics (such as the signal and noise distributions F_θ and F_z) and provides useful theoretical insights into the risk behavior of the ensemble estimator in terms of these quantities. In practical applications, however, the statistician needs to estimate the risk accurately to tune ensemble hyperparameters effectively using the observed data $(\boldsymbol{X}, \boldsymbol{y})$. These hyperparameters include the choice of component estimators

(through loss and reg), their level of regularization (regularization level for reg), the subsample sizes (k), and the ensemble size (M). For this purpose, we next construct a data-dependent proxy for the squared risk, which one can then tune with respect to various hyperparameters.

Definition 1 (Risk estimator component)**.** Let $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$ be the component estimators (3) trained on subsamples $(\boldsymbol{X}_I, \boldsymbol{y}_I)$ and $(\boldsymbol{X}_{\widetilde{I}}, \boldsymbol{y}_{\widetilde{I}})$ corresponding to index sets I and $\widetilde{I}$ with parameters $(\mathsf{loss}, \mathsf{reg})$ and $(\widetilde{\mathsf{loss}}, \widetilde{\mathsf{reg}})$. Corresponding to estimators $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$, define:

1. Degrees of freedom: $\mathsf{df}_I = \mathrm{tr}[(\partial/\partial \boldsymbol{y}_I)\boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I]$ and $\mathsf{df}_{\widetilde{I}} = \mathrm{tr}[(\partial/\partial \boldsymbol{y}_{\widetilde{I}})\boldsymbol{X}_{\widetilde{I}}\widehat{\boldsymbol{\theta}}_{\widetilde{I}}]$.
2. Residual errors: $\boldsymbol{r} = \boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I$ and $\widetilde{\boldsymbol{r}} = \boldsymbol{y}_{\widetilde{I}} - \boldsymbol{X}_{\widetilde{I}}\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$.
3. Residual degrees of freedom: traces of $\boldsymbol{V}_I = (\partial/\partial \boldsymbol{y}_I)\mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\theta}}_I)$ and $\boldsymbol{V}_{\widetilde{I}} = (\partial/\partial \boldsymbol{y}_{\widetilde{I}})\widetilde{\mathsf{loss}}'(\boldsymbol{y}_{\widetilde{I}} - \boldsymbol{X}_{\widetilde{I}}\widehat{\boldsymbol{\theta}}_{\widetilde{I}})$.

Using these quantities, define an observable quantity $\mathsf{EST}_{I,\widetilde{I}}$ as follows:

$$\mathsf{EST}_{I,\widetilde{I}} := \frac{1}{n}\sum_{i\in[n]}\Big(r_i + \mathbb{1}_{\{i\in I\}}\frac{\mathsf{df}_I}{\mathrm{tr}[\boldsymbol{V}_I]}\mathsf{loss}'(r_i)\Big)\Big(\widetilde{r}_i + \mathbb{1}_{\{i\in \widetilde{I}\}}\frac{\mathsf{df}_{\widetilde{I}}}{\mathrm{tr}[\boldsymbol{V}_{\widetilde{I}}]}\widetilde{\mathsf{loss}}'(\widetilde{r}_i)\Big) \tag{15}$$

where $\mathbb{1}_\Omega$ denotes the indicator function associated with event Ω.

The quantities $\mathrm{tr}[\boldsymbol{V}_I]$ and df_I have explicit closed-form expressions for special choices of loss and reg. Some of these are summarized in Table 5. We show next that $\mathsf{EST}_{I,\widetilde{I}}$ approximates well the component of prediction risk corresponding to the inner product of estimator errors of $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_{\widetilde{I}}$. We then naturally construct a risk estimator for the prediction risk of the ensemble estimator.

Theorem 4 (Consistency of risk estimator component)**.** In addition to Assumptions A–D, assume that reg and $\widetilde{\mathsf{reg}}$ are strongly convex. Then we have

$$\frac{1}{p}(\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta})^\top(\widehat{\boldsymbol{\theta}}_{\widetilde{I}} - \boldsymbol{\theta}) + \frac{\|\boldsymbol{z}\|_2^2}{n} = \mathsf{EST}_{I,\widetilde{I}} + \mathcal{O}_{\mathbb{P}}(n^{-1/2})\Big(1 + \frac{\|\boldsymbol{z}\|_2}{\sqrt{n}}\Big).$$

The risk estimator $\mathsf{EST}_{I,\widetilde{I}}$ is a generalization of the criterion originally proposed by [BS22] for non-ensemble regularized M-estimator. Although in this paper we focus on the isotropic Gaussian design $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, the same argument in the proof of Theorem 4 works in the anisotropic design $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$. As a result, we can show that the $\mathsf{EST}_{I,\widetilde{I}}$ (without any modification) approximates $p^{-1}(\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta})^\top\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_{\widetilde{I}} - \boldsymbol{\theta}) + n^{-1}\|\boldsymbol{z}\|_2^2$ under the event that $(\mathrm{tr}[\boldsymbol{V}_I], \mathrm{tr}[\boldsymbol{V}_{\widetilde{I}}])$ are bounded from below by a positive constant as in [BS22, Theorem 5.3].

We believe that the strongly convexity assumption on reg in Theorem 4 is an artifact of our proof (see Figure 2 for an illustration where reg is not strongly convex). Note that this type of assumption of strong convexity has already been assumed in [BS22]. This assumption guarantees for free that the coefficient $\mathsf{df}/\mathrm{tr}[\boldsymbol{V}]$ does not blow up. However, we emphasize that Theorem 2 and Corollary 3 for risk characterization do not require the strong convexity assumption. The approximation argument (see Appendix A.2.1) used to prove Theorem 2 and Corollary 3 in the non-strongly convex case is again applicable in the context of Theorem 4, although it is not currently sufficient to conclude a version Theorem 4 for non-strongly convex regularizers due to the difficulty of establishing the continuity of df_I and $\mathrm{tr}[\boldsymbol{V}_I]$ with respect to the perturbation parameter μ as $\mu \to 0$ in (28).

Equipped with the component risk estimator (15), we can now construct a consistent risk estimator for the ensemble estimator (4):

Corollary 5 (General ensemble risk estimation)**.** Fix $M \geq 1$ and consider the ensemble estimator $\widehat{\boldsymbol{\theta}}_M = \frac{1}{M}\sum_{m\in[M]} \widehat{\boldsymbol{\theta}}_m$ where the component estimator $\widehat{\boldsymbol{\theta}}_m$ is trained with $(\mathsf{loss}_m, \mathsf{reg}_m)$ on subsample I_m for $m \in [M]$ as in (3). Define an estimator EST for the squared prediction risk:

$$\mathsf{EST} := \frac{1}{M^2}\sum_{m,\ell\in[M]} \mathsf{EST}_{m,\ell}$$

where $\mathsf{EST}_{m,\ell}$ is $\mathsf{EST}_{I,\widetilde{I}}$ as defined in (15) with $(\mathsf{loss}, \mathsf{reg}, I) = (\mathsf{loss}_m, \mathsf{reg}_m, I_m)$ and $(\widetilde{\mathsf{loss}}, \widetilde{\mathsf{reg}}, \widetilde{I}) = (\mathsf{loss}_\ell, \mathsf{reg}_\ell, I_\ell)$. Then we have

$$\frac{1}{p}\|\widetilde{\boldsymbol{\theta}}_M - \boldsymbol{\theta}\|_2^2 + \frac{\|\boldsymbol{z}\|_2^2}{n} = \mathsf{EST} + \mathcal{O}_{\mathbb{P}}(n^{-1/2})\left(1 + \frac{\|\boldsymbol{z}\|_2}{\sqrt{n}}\right). \tag{16}$$

If the noise distribution has enough moments, the guarantee (16) implies that EST approximates the (full) prediction risk (that includes the irreducible error) of the ensemble estimator $\widetilde{\boldsymbol{\theta}}_M$ under Assumption A:

$$\mathbb{E}\big[(y_0 - \boldsymbol{x}_0^\top \widetilde{\boldsymbol{\theta}}_M)^2 \mid \boldsymbol{y}, \boldsymbol{X}, \{I_m\}_{m\in[M]}\big] = \frac{1}{p}\|\widetilde{\boldsymbol{\theta}}_M - \boldsymbol{\theta}\|_2^2 + \mathbb{E}[Z^2],$$

where $(y_0, \boldsymbol{x}_0) \in \mathbb{R}\times\mathbb{R}^p$ is an independent test point sampled from the same distribution as the training data $(\boldsymbol{y}, \boldsymbol{X})$. More precisely, if the noise has a finite second moment, then EST is consistent for the prediction risk. Furthermore, if the noise distribution has a finite fourth moment, by the central limit theorem (on the terms involving noise averages), EST is $\sqrt{n}$-consistent for the prediction risk. This rate is not improvable because, for the single ordinary least squares (OLS) estimator (with $\mathsf{loss}(x) = x^2/2$, $\mathsf{reg} = 0$, $|I| = |\widetilde{I}| = n$) and $F_z = \mathcal{N}(0,1)$, the risk estimator gives $\mathsf{EST} = \frac{1}{n}\frac{\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\theta}}_{\mathrm{ols}}\|_2^2}{(1-p/n)^2} \overset{\mathrm{d}}{=} \frac{\chi^2_{n-p}}{n(1-p/n)^2}$ and the standard deviation of the χ^2_{n-p} incurs an unavoidable term of order $n^{-1/2}$.

If the noise distribution F_z does not have a finite second moment but has a finite $(1+\epsilon)$-moment for $\epsilon \in [0,1)$, then even if the estimator EST may not track the prediction risk (because the prediction risk does not necessarily converge), minimizing EST is approximately equivalent to minimizing the excess squared risk $p^{-1}\|\widetilde{\boldsymbol{\theta}}_M - \boldsymbol{\theta}\|_2^2$ (which does converge). This is because the moment assumption implies $\sum_{i\in[n]} z_i^2 = o_{\mathbb{P}}(n^{2/(1+\epsilon)})$ (cf. [hp24]) so that (16) yields

$$\frac{1}{p}\|\widetilde{\boldsymbol{\theta}}_M - \boldsymbol{\theta}\|_2^2 = \mathsf{EST} - \frac{\|\boldsymbol{z}\|_2^2}{n} + o_{\mathbb{P}}(n^{-\frac{\epsilon}{1+\epsilon}}),$$

where the subtraction term $\frac{\|\boldsymbol{z}\|_2^2}{n}$ is independent of hyperparameters $(\mathsf{loss}_m, \mathsf{reg}_m, I_m)_{m\in[M]}$. We illustrate this in Figure 2 with noise following Student's t_2 distribution (that does not have a finite second moment).

4 Homogeneous ensembles

While Corollary 3 applies for a generic heterogeneous ensemble (where $\mathsf{reg}_m, \mathsf{loss}_m, |I_m|$ are allowed to differ for distinct m), concrete theoretical insights can be obtained for the homogeneous case

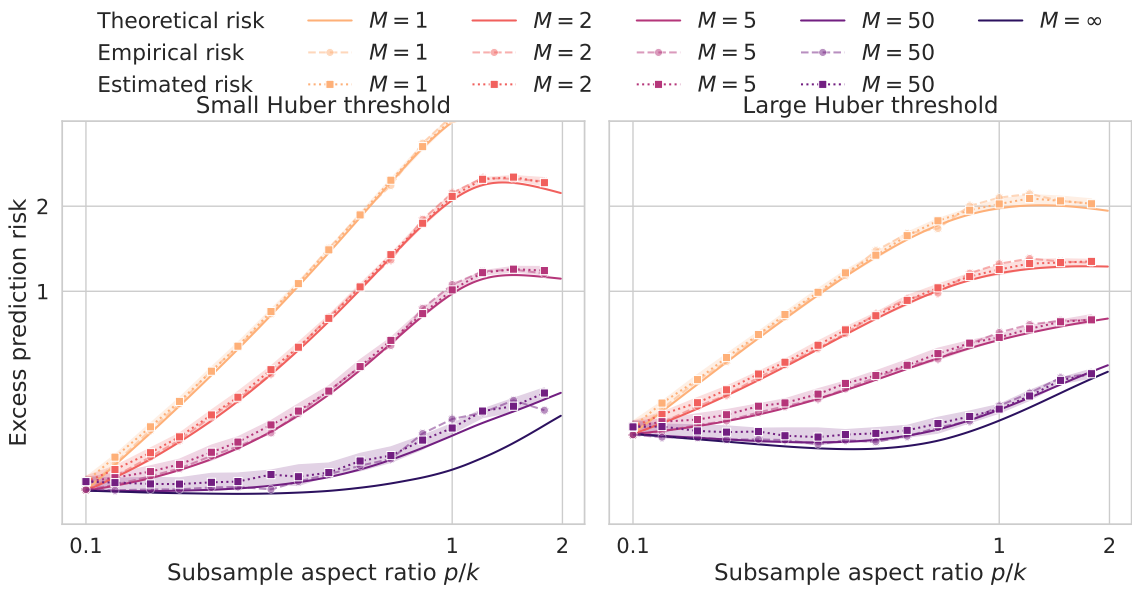

Figure 2: Risk of ℓ_1-regularized Huber ensemble at different subsample aspect ratios p/k with ℓ_1-regularization parameter $\lambda = 0.2$ and varying ensemble size M in the underparameterized regime when $p/n = 0.1$ and $n = 5000$. The solid lines represent the theoretical risks, the dashed lines represent the empirical risks averaged over 50 simulations, and the shaded regions represent the standard errors. The data model is given by Appendix D.4.2 where the noise follows Student's t distribution t_2. *Left*: Huber threshold parameter 1. *Right*: Huber threshold parameter 5.

(where $\mathsf{reg}_m, \mathsf{loss}_m, |I_m|$ are the same for every m). In the general heterogeneous case, the effect of increasing ensemble size M is not straightforward, as in general we only have:

$$\min_{m\in[M]}\{\alpha_m^2\} \overset{?}{\leq} \frac{1}{M^2}\sum_{m\in[M]}\alpha_m^2 + \frac{1}{M^2}\sum_{\substack{m,\ell\in[M]\\ m\neq\ell}}\eta_G(m,\ell)\cdot\alpha_m\alpha_\ell \leq \max_{m\in[M]}\{\alpha_m^2\}.$$

In other words, we will do no worse than the worst component but may not do better than the best component. More ensembles may or may not improve performance depending on the risks of individual estimators.[5] In contrast, for homogeneous ensembles, we show in the next subsection that increasing ensemble size does indeed help reduce the risk. This uniformity allows for more concrete analytical results and practical insights.

4.1 Risk properties

For the homogeneous ensemble of component estimators trained with the same loss, reg, and subsample size k, as $n, p, k \to \infty$ with $n/p \to \delta \in (0, \infty)$ and $k/n \to c \in (0, 1]$, the limiting risk (14) is given by:

$$\mathcal{R}_M = \frac{1}{M}\mathcal{R}_1 + \Big(1 - \frac{1}{M}\Big)\mathcal{R}_\infty \quad \text{where} \quad \begin{cases} \mathcal{R}_1 \coloneqq \alpha^2 & \text{(non-ensemble risk)},\\ \mathcal{R}_\infty \coloneqq \eta_G\alpha^2 & \text{(full-ensemble risk)}. \end{cases} \tag{17}$$

[5]Even if a predictor complements other predictors (in the sense that it has small or negative η_G with the other predictors in the ensemble), it is only "beneficial" for the ensemble if it also has a small risk (α_m) itself.

Observe that the limit $\mathcal{R}_M$ is simply a convex combination of the asymptotic risk of the single estimator $\mathcal{R}_1$ and of the full-ensemble estimator $\mathcal{R}_\infty$.[6] (Note that when $M \to \infty$, $\mathcal{R}_M$ does indeed converge to $\mathcal{R}_\infty$, justifying the notation for the limit for the case when $m \neq \ell$.) As a sanity check, note that in the special case when $c = 1$, the above setting corresponds to the non-ensemble case discussed in [TAH18, Section 4].

The following result shows the advantage of ensembling. In the classical bagging and subagging literature, it is well known that the risk of the ensemble estimator decreases as the ensemble size increases, due to the reduction in variance that comes with having more predictors in the ensemble. In the proportional asymptotic regime that we study in this paper, since subagging also introduces bias, this is not immediate. A general result along these lines that verifies the monotonicity of the risk of the ensemble itself (not the asymptotic limit) follows from Proposition 3.1 of [PDK23]; see Equation (10) of [PDK23]. Below we verify that the asymptotic risk is strictly monotonic in M by showing that $\eta_G < 1$ in general. This implies that the asymptotic risk is strictly decreasing in M.

Proposition 6 (Improvement due to ensembling)**.** Fix the subsample ratio $c = k/n \in (0, 1)$ and let $\mathcal{R}_M$ be the limiting risk as defined in (17). Then $\mathcal{R}_M$ is strictly decreasing in the number of ensembles M, i.e., $\mathcal{R}_{M+1} < \mathcal{R}_M$ for all $M \in \mathbb{N}$.

The proof follows immediately because of the form of the ensemble risk (17) and the fact that $\eta_G < 1$ when $c < 1$, which follows from Theorem 1 with $c = \widetilde{c}$. This monotonicity in the ensemble size M is illustrated by Figure 2 for the ensemble of ℓ_1-regularized Huber regression[7].

Because the risk decreases in M, the optimal ensemble size is $M = \infty$.[8] However, it may not be feasible to use an ensemble size of $M = \infty$. In practice, it suffices to use a large enough M that gives a suboptimal risk close to the full-ensemble risk. For this purpose, a natural idea is to estimate the risk of non-ensemble estimator $M = 1$ and the full estimator $M = \infty$, and obtain an estimate for the risk of M-ensemble using the relationship in (17). This is very similar to the extrapolated cross-validation estimator (ECV) of [DPRK24] which estimates the risk of $M = 1$ and $M = 2$.

We also show that for any ensemble size M, when the subsample ratio c is optimized, the resulting risk decreases in the inverse data aspect ratio $\delta = \lim n/p$. To prepare for the forthcoming statement, let us write the limiting risk $\mathcal{R}_M$ in (17) by $\mathcal{R}_M(\delta, c)$ to make the dependence on the limit $\delta = \lim n/p$ and subsample ratio $c = k/n$ clear. With this notation, we can say the following about the optimally subsampled risk.

Proposition 7 (Monotonicity of optimal subsample risk)**.** The map $\delta \mapsto \inf_{c\in(0,1]} \mathcal{R}_M(\delta, c)$ is non-increasing over $\delta \in (0, \infty)$ for all $M \in \mathbb{N}$.

A consequence of this proposition is that the risk of the optimal ensemble estimator is decreasing in the sample size n for a fixed (large enough) feature size p. Moreover, combined with Proposition 6,

[6] The reason we call this the "full" ensemble is that the ensemble estimator $\widehat{\boldsymbol{\theta}}_M$ when $M \to \infty$ is almost surely equal (coordinate-wise and conditioned on the data) to an ensemble estimator fitted on all possible (and distinct) $\binom{n}{k}$ subsamples of size k; see Lemma A.1 of [DPK23] for a precise statement and proof.

[7] The Huber loss is defined as $\mathsf{loss}(x) = \frac{x^2}{2\rho}\,\mathbb{1}(|x| \le \rho) + (|x| - \frac{\rho}{2})\,\mathbb{1}(|x| > \rho)$, where ρ is a positive constant, referred to as the Huber parameter.

[8] In practice, setting $M = \binom{n}{k}$ suffices by only averaging over estimators trained on distinct subsamples (see Appendix A.1 of [DPK23] for more details), but this still can be quite large.

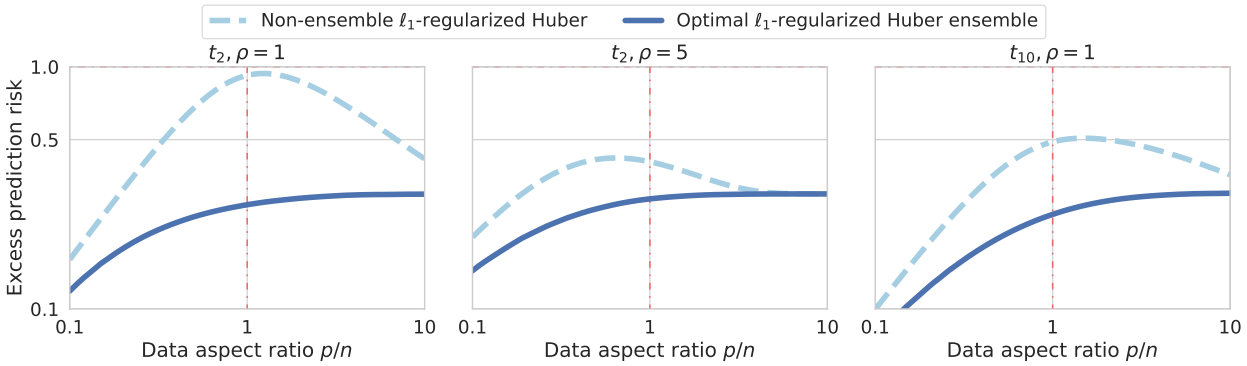

Figure 3: **Optimal subsample risk of the Huber lasso ensemble is monotonic in the data aspect ratio.** Risks of the ℓ_1-regularized Huber and optimal ℓ_1-regularized Huber ensemble, for fixed ℓ_1-regularization parameter $\lambda = 0.5$ and varying Huber parameter ρ, at different data aspect ratios p/n ranging from 0.1 to 10. The data model is given in Appendix D.4.2. *Left*: noise follows Student's t distribution t_2 and Huber parameter $\rho = 1$. *Middle*: noise follows Student's t distribution t_2 and Huber parameter $\rho = 5$. *Right*: noise follows Student's t distribution t_{10} and Huber parameter $\rho = 1$.

we also have that this monotonic risk profile lies below the function $\mathcal{R}_1(\delta, 1)$, the risk profile of the original predictor trained once on the full dataset $(\boldsymbol{X}, \boldsymbol{y})$ with no ensembling (the risk of which, as we discussed above, can be non-monotonic). Such monotonicity in the inverse data aspect ratio is important because it ensures that increasing the amount of data relative to features consistently improves the estimator's performance. In a sense, a monotonic decrease in risk with the optimal subsample ratio certifies that the model effectively utilizes all the additional data, leading to better performance as the data size grows. This result is illustrated in Figure 3 for the ℓ_1-regularized Huber regression.

4.2 Examples and connections to literature

In this section, we specialize the general risk characterization in Corollary 3 in several examples of interest. Throughout this section, we will consider ensembles of component predictors trained on the same loss, reg (with a tuning parameter λ), and subsample size k. We begin by considering convex regularized least squares, with further specialization to bridge, ridge, and lasso. We then consider general ridge regularized estimators allowing for non-squared loss in Section 4.2.2 with further specialization to the Huber loss. For the reader's convenience, the proximal operators, their derivatives, Moreau envelopes, and their derivatives for the ridge and lasso regularization and Huber loss functions are recalled in Table 6.

4.2.1 Ensembles of regularized least squares

In this section, we consider subagging regularized least squares. Given a common regularizer reg and a regularization parameter $\lambda > 0$, the component estimators are given by:

$$\widehat{\boldsymbol{\theta}}_m := \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^p} \frac{1}{2} \|\boldsymbol{y}_{I_m} - \boldsymbol{X}_{I_m} \boldsymbol{b}\|_2^2 + \lambda \sum_{j \in [p]} \mathsf{reg}(b_j).$$

Now Assumption D-(1) translates to having a bounded second moment of the noise distribution F_z, which we denote by σ^2. Using the explicit formula $\mathsf{prox}_{\mathsf{loss}}(x; \tau) = x/(1+\tau)$ for $\mathsf{loss}(x) = x^2/2$ and performing some change of variables, the risk convergence in (17) holds with $(\mathcal{R}_1, \mathcal{R}_\infty)$ given

by:

$$\mathcal{R}_1 = \tau^2 - \sigma^2 \quad \text{and} \quad \mathcal{R}_\infty = \xi^2 - \sigma^2,$$

where τ and ξ are the solutions to the following systems:

System 2 (Ensembles of regularized least square)**.** Given $\lambda \in (0, \infty)$, $\delta \in (0, \infty)$, $c \in (0, 1]$, $\sigma^2 \in [0, \infty)$, define the following 2-scalar system of equations in variables (τ, a):

$$\tau^2 = \mathbb{E}\big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)^2\big] + \sigma^2, \tag{18a}$$

$$\tfrac{\lambda}{\sqrt{c\delta}} = a\tau\big(1 - \tfrac{1}{c\delta}\mathbb{E}\big[\mathsf{prox}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big]\big), \tag{18b}$$

where $H \sim \mathcal{N}(0, 1)$ and $\Theta \sim F_\theta$ are independent. Given $(\tau, a) \in \mathbb{R}^2_{>0}$ that satisfy the above systems, define the following 1-scalar system of equations in variable ξ:

$$\xi^2 = \mathbb{E}\big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big) \cdot \big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}\widetilde{H}; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)\big] + \sigma^2 \tag{19a}$$

where $\begin{pmatrix} H \\ \widetilde{H} \end{pmatrix} \sim \mathcal{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \eta_{\scriptscriptstyle H} \\ \eta_{\scriptscriptstyle H} & 1 \end{bmatrix}\Big)$ with $\eta_{\scriptscriptstyle H} = c\dfrac{\xi^2}{\tau^2}$, and $\Theta \sim F_\theta$ independent.

Compared to Systems 1a and 1b, the parameterization in System 2 is slightly different. This is done to match the result with some of the existing results for regularized least squares (for $M = 1$). The invertible transformations are given by: $a = \frac{\lambda}{\beta}$, $\tau = \sqrt{c\delta}\frac{\beta}{\nu}$, and $\xi^2 = \eta_{\scriptscriptstyle G}\alpha^2 + \sigma^2$. These parameters also have interpretations as in Section 3.2, summarized below.

Remark 2 (Interpretation of parameters in System 2)**.** The parameter τ^2 is simply the (full) prediction risk of the non-ensemble estimator in the limit, which is $\alpha^2 + \sigma^2$. Note that $\tau^2 \geq \sigma^2$ and is sometimes referred to as effective "inflated" noise variance due to the high dimensionality [DMM09, BM11b]. Moreover, the fact that $\tau^2 = \frac{(c\delta)\beta^2}{\nu^2} = \frac{\beta^2/(c\delta)}{\nu^2/(c\delta)^2}$ is also at the core of consistency of generalized cross-validation (discussed in Section 3.4) for the non-ensemble estimator. Here, the numerator is the asymptotic training error and the denominator is the asymptotic degrees of freedom correction.[9] The parameter $a\tau$ is the effective threshold parameter at which one applies the proximal operator to the noise-inflated effective observation and appears in approximate message passing (AMP) formulations [DMM09]. The parameter a serves as a proportionality constant between the effective threshold $a\tau$ and standard deviation τ of the inflated effective noise. Finally, the parameter ξ^2, which is the main contribution of this paper, is the full-ensemble predictor risk (when $M \to \infty$), which is also $\eta_{\scriptscriptstyle G}\alpha^2 + \sigma^2$.

In the following, we isolate some special cases of regularized M-estimators to compare with existing work.

Remark 3 (Bridge ensembles)**.** Bridge estimators are also known as ℓ_q-regularized least squares and are a popular class of regularized M-estimator [FF93, Fu98]. For ℓ_q regularizer $\mathsf{reg}_q(x) = |x|^q$, the risk of the bridge for $M = 1$ is derived in [WMZ18, WWM20] for general $q \in [1, 2]$. For instance, we recover [WMZ18, Theorem 2.1] by changing of variables $(\lambda', \Theta') = (\lambda/\sqrt{c\delta}, \sqrt{c\delta}\Theta)$ such that the limiting prediction risks match $\tau' = \tau$.[10] Equation (19a) generalizes it for any $M \geq 1$. Further

[9] Observe that the factors of $c\delta$ arise because both the asymptotic training error β^2 and the asymptotic degrees of freedom correction ν are defined with normalization of p in Table 2.

[10] The parameter a requires the following change: $a' = (c\delta/\tau^2)^{(1-q)/2}a$ in the two systems.

special boundary cases of $q = 2$ (ridge) and $q = 1$ (lasso) are further isolated in the next two remarks.

Remark 4 (Ridge ensembles)**.** For ridge regression when $q = 2$, System 2 recovers Theorem 4.1 of [PDK23] under isotropic features, using a slight change of variables $(\lambda', \Theta') = (\lambda/\sqrt{c\delta}, \sqrt{c\delta}\Theta)$. The solution v to the (limiting) Stieltjes transform of the spectrum of the sample gram matrix therein satisfies that $v = (a\tau)^{-1}$. For ridge ensembles, there is a deeper connection between $\mathcal{R}_1$ and $\mathcal{R}_\infty$. It turns out that $\mathcal{R}_\infty(\lambda, c)$ is exactly equal to $\mathcal{R}_1(\mu, 1)$ for a new level of regularization $\mu = v^{-1} \geq \lambda$ that depends on c (and properties of the data distribution). The lower the subsampling proportion of c, the higher the value of this implicit regularization μ. There are entire paths of equivalences in the (λ, c) plane where not only are the asymptotic squared risks the same, but also the estimators themselves are equivalent. In a sense, one can think of the combined effect of subsampling and ensembling as an additional (implicit) ridge regularization. The prediction risk equivalences are first proved in [DPK23] and later generalized to other risks and estimator equivalences in [PD23].

Remark 5 (Lasso ensembles)**.** For the lasso predictor when $q = 1$, the first set of equations (18) for $M = 1$ in System 2 recovers [BM11b, Theorem 1.5] with a change of variables $\lambda' = \lambda/\sqrt{c\delta}$ and $\Theta' = \sqrt{c\delta}\Theta$. On the other hand, the second set of equations in System 2 for $M = \infty$ is new to the literature. We show empirically in the next section that the optimal full-ensemble subsampled lassoless is not the same as the optimal non-ensemble lasso (on full data). Thus, the subsampling and ensembling of the lasso are qualitatively different than the subsampling and ensembling of the ridge regression, as pointed out in Remark 4. In particular, the effect of subsampling for lasso is not merely additional lasso regularization.

4.2.2 Ensembles of general ridge regularized estimators

In this section, we specialize the results of Corollary 3 for ridge regularized ensembles allowing for general loss functions. Specifically, given $\lambda > 0$, consider component estimators of the form:

$$\widehat{\boldsymbol{\theta}}_m(I_m) \coloneqq \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \sum_{i\in I_m} \mathsf{loss}(y_i - \boldsymbol{x}_i^\top \boldsymbol{b}) + \lambda\|\boldsymbol{b}\|_2^2.$$

Recall that the risk convergence in (17) holds with $\mathcal{R}_1 = \alpha^2$ and $\mathcal{R}_\infty = \eta_G\alpha^2$ where α and η_G are solutions to System 1a and System 1b respectively. Here, using the explicit formula $\mathsf{prox}_{\mathsf{reg}}(x;\tau) = x/(1+\lambda\tau)$ for $\mathsf{reg}(x) = \lambda|x|^2/2$, these systems can be simplified as follows:

System 3 (Nonlinear system for general ridge regularized ensembles)**.** Given $\lambda \in (0,\infty)$, $\delta \in (0,\infty)$, and $c \in (0,1]$, define the following 2-scalar system of equations in variables (α, κ):

$$\alpha^2 = c\delta\cdot\kappa^2\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(Z+\alpha G;\kappa)^2]+\lambda^2\kappa^2\mathbb{E}[\Theta^2] \quad \text{and} \quad \alpha = c\delta\cdot\tfrac{\kappa}{1-\lambda\kappa}\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(Z+\alpha G;\kappa)\cdot G], \tag{20}$$

where $G \sim \mathcal{N}(0,1)$, $\Theta \sim F_\theta$, $Z \sim F_z$, all mutually independent. Let (β, ν) be parameters expressed in terms of (α, κ) as:

$$\beta^2 = \tfrac{1}{\kappa^2}\alpha^2 - \lambda^2\mathbb{E}[\Theta^2] \quad \text{and} \quad \nu = \tfrac{1}{\kappa} - \lambda. \tag{21}$$

Given parameters $(\alpha, \beta, \kappa, \nu)$ that satisfy (20) and (21), define the following 2-scalar system of

equations in variables (η_H, η_G):

$$\eta_H = \frac{c^2\delta}{\beta^2}\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(Z+\alpha G;\kappa)\cdot \mathsf{env}'_{\mathsf{loss}}(Z+\alpha\widetilde{G};\kappa)] \quad \text{and} \quad \eta_G = \frac{\eta_H\beta^2+\lambda^2\mathbb{E}[\Theta^2]}{\beta^2+\lambda^2\mathbb{E}[\Theta^2]}, \tag{22}$$

where $(G, \widetilde{G})$ are jointly normal with $\mathbb{E}[G\widetilde{G}] = \eta_G$.

When $M = 1$, System 3 recovers [EK13, Theorem 2.1]. When $c\delta > 1$ (underparameterized regime) and λ=0 (unregularized case), we have $\nu = \kappa^{-1}$, $\beta = \kappa^{-1}\alpha$, $\eta_H = \eta_G$. Substituting them to (22), we recover Theorem 2.3 in the recent work of [BK24].

5 Subagging and overparameterization

In Section 4, we discussed subagging of regularized estimators with an explicit regularization level $\lambda > 0$. Triggered by the success of overparameterized neural networks that can (nearly) interpolate, there has been a surge of recent work analyzing the risk behavior of estimators with vanishing regularization, such as the minimum ℓ_2-norm interpolator (ridgeless), minimum ℓ_1-norm interpolator (lassoless), and max-margin interpolators, among others. In this section, we discuss subagging of minimum ℓ_q-norm interpolators for $q \in \{1, 2\}$. We will demonstrate some interesting risk properties in Section 5.1, showcase the benefits of subagging in overparameterized regimes in Section 5.2, and contrast with optimal explicit regularization in Section 5.3.

5.1 Subagging of minimum ℓ_q-norm interpolators

We will focus in this section on subagging of bridgeless estimators, that is ℓ_q-norm regularized least squares with $\mathsf{reg}(\boldsymbol{b}) = \|\boldsymbol{b}\|_q^q$. Our main cases of interest are the "ridgeless" and "lassoless" estimators, which are the special cases when $q = 1$ [HMRT22] and $q = 2$ [LW21], respectively. The terminology "less" is motivated by the fact that these estimators can be defined in a limiting sense as $\lambda \to 0^+$ for bridge estimators with regularization level λ. We consider the ensemble of predictors $(\widehat{\boldsymbol{\theta}}_m)_{m\in[M]}$ of the form:[11]

$$\widehat{\boldsymbol{\theta}}_m \coloneqq \lim_{\lambda\to 0+} \widehat{\boldsymbol{\theta}}_m(\lambda) \quad \text{where} \quad \widehat{\boldsymbol{\theta}}_m(\lambda) \in \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{y}_{I_m} - \boldsymbol{X}_{I_m}\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{b}\|_q^q.$$

In the underparameterized regime ($p < n$), these are simply the least squares estimators: $\widehat{\boldsymbol{\theta}}_m = (\boldsymbol{X}_{I_m}^\top \boldsymbol{X}_{I_m})^{-1}\boldsymbol{X}_{I_m}^\top \boldsymbol{y}_{I_m}$. In the overparameterized regime ($p > n$), these correspond to the minimum ℓ_q-norm interpolators: $\{\|\widehat{\boldsymbol{\theta}}_m\|_q \colon \boldsymbol{y}_{I_m} = \boldsymbol{X}_{I_m}\widehat{\boldsymbol{\theta}}_m\}$, when $\boldsymbol{X}_{I_m}$ has independent rows to allow for interpolation. For $q = 2$, when reg is the ridge penalty, this also has a closed-form expression given by: $\widehat{\boldsymbol{\theta}}_m = (\boldsymbol{X}_{I_m}^\top \boldsymbol{X}_{I_m})^\dagger \boldsymbol{X}_{I_m}^\top \boldsymbol{y}_{I_m}$, where $\boldsymbol{A}^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix $\boldsymbol{A}$. In other cases, we do not have a closed-form expression for the minimum ℓ_q-norm interpolator. The next system specializes System 2 to convex regularized least squares with vanishing regularization, by taking the limit as $\lambda \to 0^+$.

System 4 (Ensembles of minimum ℓ_q-norm interpolators)**.** Given $\delta \in (0, \infty)$, $c \in (0, 1]$ such that $c\delta < 1$, and $\sigma^2 \in [0, \infty)$, define the following 2-scalar system of equations in variables

[11] We refer readers to [Tib13] for details on how the sequence of estimators is defined when the estimators for $\lambda > 0$ are not unique, as in the case of the lasso.

(τ, a):

$$\tau^2 = \mathbb{E}\big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)^2\big] + \sigma^2 \tag{23a}$$

$$0 = 1 - \tfrac{1}{c\delta}\mathbb{E}\big[\mathsf{prox}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big] \tag{23b}$$

where $H \sim \mathcal{N}(0,1)$, and $\Theta \sim F_\theta$ independent. Given (τ, a) that satisfy (23a) and (23b), define the following 1-scalar system of equations in variable ξ:

$$\xi^2 = \mathbb{E}\big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big) \cdot \big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}\widetilde{H}; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)\big] + \sigma^2 \tag{24a}$$

where $\begin{pmatrix} H \\ \widetilde{H} \end{pmatrix} \sim \mathcal{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \eta_H \\ \eta_H & 1 \end{bmatrix}\Big)$ with $\eta_H = c\dfrac{\xi^2}{\tau^2}$, and $\Theta \sim F_\theta$ independent.

To the best of our knowledge, the existence and uniqueness of the solution (τ, a) to (23) are not fully established in the literature, except for the special cases of $q = 2$ (ridgeless) and $q = 1$ (lassoless). Assuming this is the case, the existence and uniqueness of the solution ξ to (24a) follow from System 1b.

Observe that the equations (23a) and (24a) in System 4 are special cases of (18a) and (19a) in System 2 for ℓ_q penalties. Equation (23b) is the limit of (18b) as $\lambda \to 0^+$. Indeed, for $q \in \{1, 2\}$, the solution to System 2 with $\lambda > 0$ converges to the solution to System 4 as $\lambda \to 0^+$ (see Appendix C.2 for a proof). This means that

$$\lim_{\lambda \to 0^+} \operatorname*{p-lim}_{n \to +\infty} \Big\| \frac{1}{M} \sum_{m \in [M]} \widehat{\boldsymbol{\theta}}_m(\lambda) - \boldsymbol{\theta} \Big\|_2^2 = \mathcal{R}_M := M^{-1}\mathcal{R}_1 + (1 - M^{-1})\mathcal{R}_\infty$$

where $\mathcal{R}_1 = \tau^2 - \sigma^2$ and $\mathcal{R}_\infty = \xi^2 - \sigma^2$, and by the same argument, the limiting risk $\mathcal{R}_M$ satisfies Proposition 6 and 7 (see also Figure 9 and 10). However, it is challenging to show the above display with the order of $\lim_\lambda$ and p-lim_n swapped. For the ridgeless estimator ($q = 2$) and any M, this is proved in [PDK23] using a uniform convergence argument. For the lassoless estimator ($q = 1$ and $M = 1$), this is done in [LW21] where the authors directly analyze the interpolator by constructing a suitable AMP algorithm. We conjecture that this is, in general, true at least for bridgeless estimators for any $q \in [1, 2]$ and M. Since this is not the main focus of our paper and is only intended as an illustrative case, we will not work towards this goal in the current paper. We will instead investigate properties and consequences of System 4.

Further special cases of $q = 2$ (ridgeless) and $q = 1$ (lassoless) are isolated in the next two remarks. These will serve as our two main running examples in this section.

Remark 6 (Ridgeless ensembles)**.** For squared loss and ridge regularizer ($q = 2$), when $c\delta < 1$ and $\lambda \to 0^+$, we get

$$\mathcal{R}_1(\delta, c) + \sigma^2 = \mathbb{E}[\Theta^2](1 - c\delta) + \sigma^2 \frac{1}{1 - c\delta}, \quad \mathcal{R}_\infty(\delta, c) + \sigma^2 = \mathbb{E}[\Theta^2] \frac{(1 - c\delta)^2}{\delta(\delta - (c\delta)^2)} + \sigma^2 \frac{\delta}{\delta - (c\delta)^2}.$$

The result above aligns with the risk ensemble of ridgeless estimators presented in Corollary 6.1 of [PDK23]. This can be seen by substituting $\delta = 1/\phi$ and $c = \phi/\psi$, or equivalently $c\delta = 1/\psi$.

Remark 7 (Lassoless ensembles)**.** For squared loss and lasso regularizer ($q = 1$), when $c\delta < 1$ and $\lambda \to 0^+$, the total risk τ^2 is the solution to the following equations:

$$\begin{aligned} \tau^2 &= \sigma^2 + \mathbb{E}\big[\big(\,\mathsf{soft}\,\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)^2\big], \\ 1 &= \tfrac{1}{c\delta}\mathbb{P}\big(\big|\tfrac{\tau}{\sqrt{c\delta}}H + \Theta\big| > \tfrac{a\tau}{\sqrt{c\delta}}\big), \end{aligned} \tag{25}$$

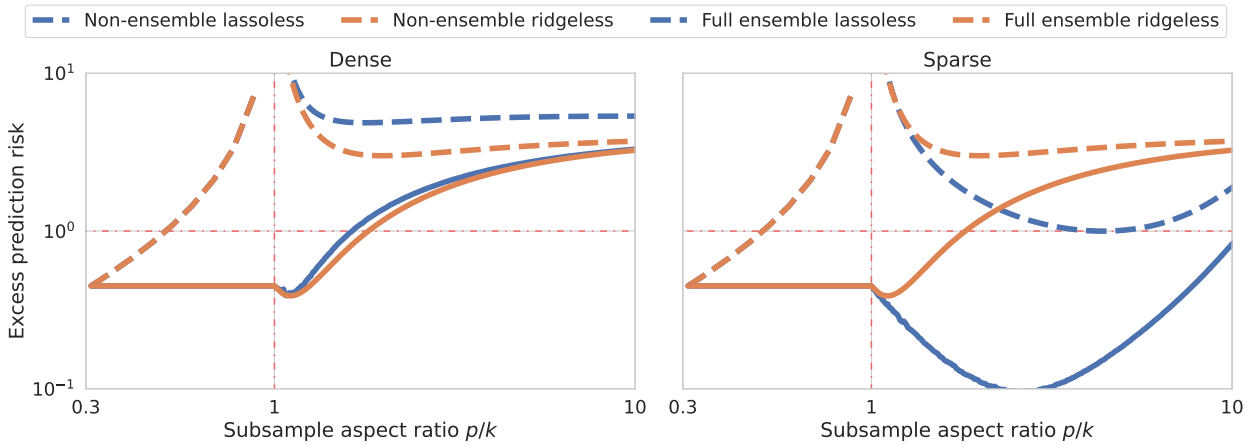

Figure 4: Prediction risks of the full-ensemble lassoless and ridgeless at different subsample aspect ratios p/k ranging from 0.3 to 10. The data model is given by (50) with signal strength $\rho = 2$, noise level $\sigma = 1$, data aspect ratio $p/n = 0.3$, and feature size $p = 500$. The support proportion s varies. *Left*: dense regime with $s = 0.9$. *Right*: sparse regime with $s = 0.01$. We observe that the full-ensemble risk is continuous at the interpolation. Also, the full-ensemble risk has a negative derivative to the right of interpolation. This implies that the optimal subsample size is in the overparameterized regime.

and ξ^2 is the solution to the following equations:

$$\xi^2 = \sigma^2 + \mathbb{E}\big[\big(\mathsf{soft}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big) \cdot \big(\mathsf{soft}\big(\tfrac{\tau}{\sqrt{c\delta}}\widetilde{H} + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)\big],$$

with $\mathbb{E}[H\widetilde{H}] = c\frac{\xi^2}{\tau^2}$. Note that soft is the soft threshold function defined by $\mathsf{soft}(x;\tau) = (|x| - \tau)_+ \operatorname{sign}(x)$. In the ensemble setting, the case of $M = 1$ corresponds to [LW21, Theorem 2] with a slight change of variables $\Theta' = \sqrt{c\delta}\Theta$. The full-ensemble case when $M = \infty$ is new.

Remark 8 (Avoiding risk divergence with ensembling)**.** Note that in the underparameterized regime when $c\delta > 1$ is fixed, the estimator of interest is simply the ensemble of least squares. Thus, a standard Stieltjes transform argument or the explicit formula for the expectation of inverse Wishart matrices give

$$\mathcal{R}_1(\delta, c) + \sigma^2 = \sigma^2 \frac{c\delta}{c\delta - 1} \quad \text{and} \quad \mathcal{R}_\infty(\delta, c) + \sigma^2 = \sigma^2 \frac{\delta}{\delta - 1} \quad \text{for all } c > \delta^{-1}.$$

It just so happens that for the full-ensemble least squares estimators, only the inverse aspect ratio δ of the original data matters! In particular, as $c \to (\delta^{-1})^+$, $\mathcal{R}_1$ diverges, while the full-ensemble risk $\mathcal{R}_\infty$ is still bounded. Now let us consider the over-parameterized regime $c\delta < 1$. By simple algebra, for any regularizer reg, the solution τ to the sub-system (23a)-(23b) in System 4 is uniformly bounded from below as:

$$\tau^2 \geq (1 - c\delta)^{-1}\sigma^2. \tag{26}$$

(See Appendix C.1 for the proof.) Recalling $\mathcal{R}_1 = \tau^2 - \sigma^2$, this means that the risk of the non-ensemble interpolators blows up as $c \to (\delta^{-1})^-$. For the minimum ℓ_2- and ℓ_1-norm interpolators, this is shown in [HMRT22, LW21]. For the full-ensemble cases, we experimentally observe from Figure 4 that the risk $\mathcal{R}_\infty$ is continuous in c for the full ridgeless and lassoless ensembles. In particular, it does not blow up around $c = \delta^{-1}$. For ridgeless, this claim is easy to verify (and holds more generally, as shown in [PDK23]). For lassoless, given the solution (a, τ) to (25), in Appendix C.3, we identify that the condition $\lim_{c\to(\delta^{-1})^-}(a\tau) = 0$ is sufficient to obtain the conclusion $\lim_{c\to(\delta^{-1})^-}(\xi^2) = \frac{\delta}{\delta-1}\sigma^2$, and we observe experimentally that $a\tau \to 0$ holds (Figure 8), however we are currently not able to provably establish that $\lim_{c\to(\delta^{-1})^-}(a\tau) = 0$.

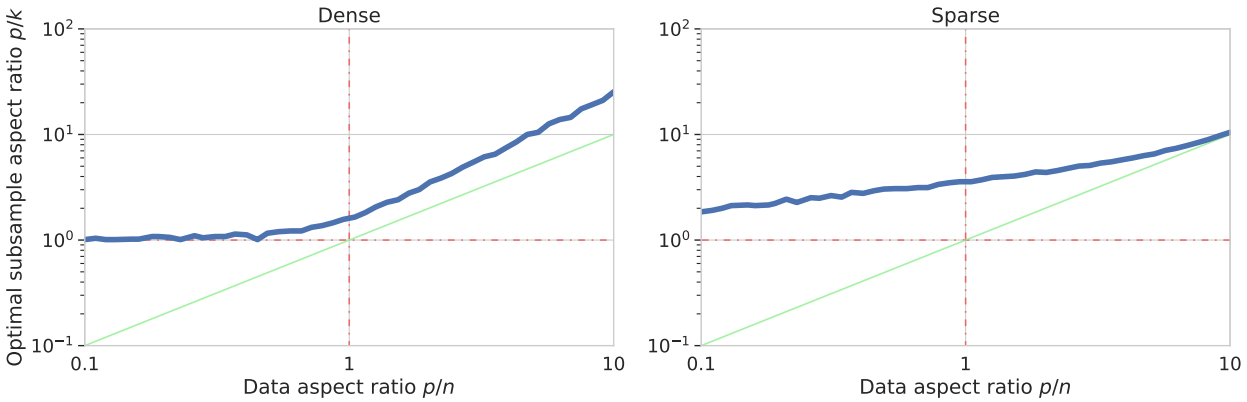

Figure 5: **Optimal subsample size for the lassoless ensemble is always in the overparameterized regime.** Optimal subsample aspect ratio p/k that achieves the optimal risk for the lassoless ensemble at different data aspect ratios p/n ranging from 0.1 to 10. The data model is as in (50) with signal strength $\rho = 2$, noise level $\sigma = 1$, data aspect ratio $p/n = 0.1$, feature size $p = 500$, and varying support proportion s. *Left*: dense regime with $s = 0.9$. *Right*: sparse regime with $s = 0.01$.

5.2 Optimal subsample size

An intriguing observation from Figure 4 concerns the optimal subsample size k_*. When the sample size n and the number of features p are fixed with $n > p$, the optimal subsample size that minimizes the full risk $\mathcal{R}_\infty$ falls below p. This suggests that even when the original sample lies in the underparameterized regime, the optimal subsample size shifts into the overparameterized regime. This phenomenon is proved for ridgeless ensembles ($\mathsf{reg}(x) = x^2$) by [PDK23]. Expanding on this, and utilizing System 4, we empirically show that this behavior extends beyond ridgeless ensembles to lassoless ensembles ($\mathsf{reg}(x) = |x|$) as well (see Figure 5).

5.3 Optimal subagging versus optimal (explicit) regularization

There are three parameters one can tune to optimize the asymptotic risk $\mathcal{R}_M(\lambda, c)$ of the ensemble estimator, as in Remarks 4 and 5: the regularization level λ, the subsample size c, and the ensemble size M. This hyperparameter optimization is simplified for ridge regression, as shown in Theorem 2.3 of [DPK23]. Minimization with respect to all three parameters is equal to the minimization over λ when $M = 1$ and $c = 1$ (non-ensemble setting), which is the same as minimization over M and c when $\lambda = 0$ (ensemble of ridgeless predictors):

$$\underbrace{\min_{\lambda\in[0,\infty],M\in\mathbb{N},c\in[0,1]} \mathcal{R}_M(\lambda, c)}_{\text{opt regularization and opt ensemble}} = \underbrace{\min_{\lambda\in[0,\infty]} \mathcal{R}_1(\lambda, 1)}_{\text{opt regularization but no ensemble}} = \underbrace{\min_{c\in[0,1]} \mathcal{R}_\infty(0, c)}_{\text{opt ensemble but no regularization}} .$$

In some situations, however, the risk minimization over all three parameters can be strictly better:

$$\underbrace{\min_{\lambda\in[0,\infty],M\in\mathbb{N},c\in[0,1]} \mathcal{R}_M(\lambda, c)}_{\text{opt regularization and opt ensemble}} < \underbrace{\min_{\lambda\in[0,\infty]} \mathcal{R}_1(\lambda, 1)}_{\text{opt regularization but no ensemble}} \wedge \underbrace{\min_{c\in[0,1]} \mathcal{R}_\infty(0, c)}_{\text{opt ensemble but no regularization}} .$$

We illustrate this through a numerical experiment with lasso. We show that the optimal full-ensemble subsampled lassoless is not the same as the optimal non-ensemble lasso (on full data). In Figure 6, we contrast the sparse and dense data settings. In each case, we show the full-ensemble risk heatmap in λ and p/k. In the left panel (the sparse setting), we see that the optimal subsample

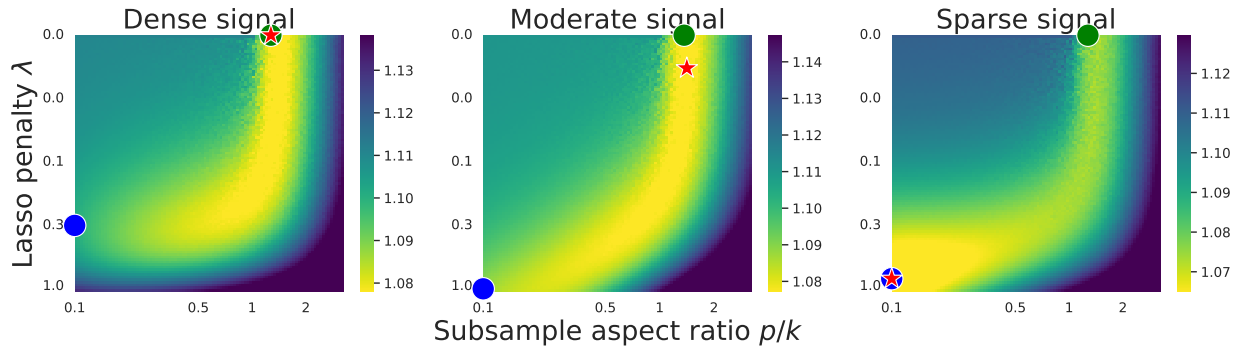

Figure 6: **Optimally subsampled lassoless regression can outperform optimal lasso regression.** Heatmaps of theoretical prediction risk in λ and p/k of full lasso ensemble in the underparameterized regime ($p/n = 0.1$). The data model is given by (50) with signal strength $\rho = 0.5$ and noise level $\sigma = 1$ at different sparsity levels s. *Left*: Dense regime with support proportion $s = 0.9$. Optimal subsample lassoless is better than optimal lasso. *Middle*: Modest sparse regime with support proportion $s = 0.5$. Optimal lasso ensemble is better than the optimal lasso and optimal subsample lassoless. *Right*: Sparse regime with support proportion $s = 0.2$. Optimal lasso is better than optimal subsample lassoless.

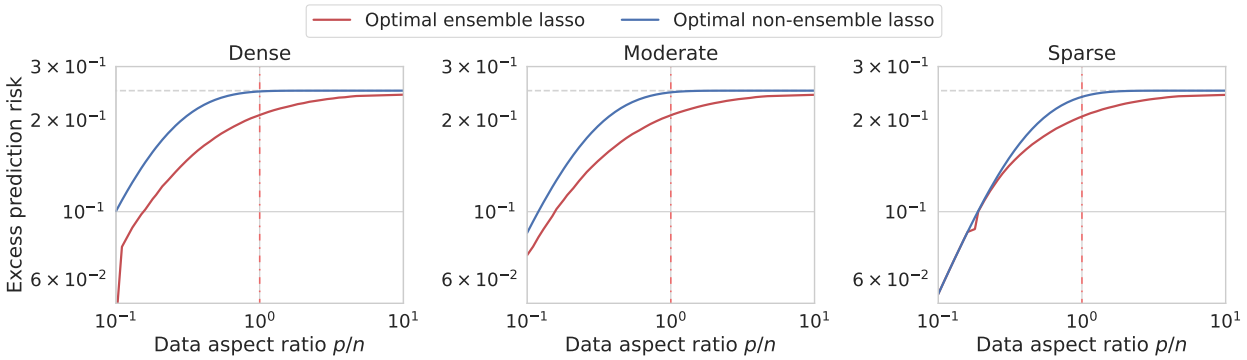

Figure 7: **Optimally subsampled ensemble lasso can uniformly beat optimally tuned non-ensemble lasso across different data aspect ratios.** The theoretical prediction risk of optimal ensemble and non-ensemble lasso at different data aspect ratios p/n is shown. The data model is given by (50) with signal strength $\rho = 0.5$ and noise level $\sigma = 1$ at different sparsity levels s. *Left*: Dense regime with support proportion $s = 0.9$. *Middle*: Modest sparse regime with support proportion $s = 0.5$. *Right*: Sparse regime with support proportion $s = 0.2$.

lassoless is worse than the optimal lasso. This is expected because the lasso is known to perform well in the sparse setting. In the right panel (the dense setting), we see that the optimal subsample lassoless is better than the optimal lasso. This is interesting because it shows that the subsample and ensemble induce an implicit regularization effect. In short, the optimal lassoless ensemble can be better or worse than the optimal lasso. In other words, the full-ensemble lassoless when c is optimized is not the same as the optimized lasso when λ is optimized on the full data.

A similar conclusion holds for overparameterized settings, as shown in Figure 11. In particular, depending on the SNR and δ, the subsample optimized risk may be smaller or larger than the optimized lasso risk. The conclusion is that, in general, it helps to optimize λ, but also to optimize the ensemble size and subsample size. The subsample and ensemble induce an implicit regularization effect. Once optimized, this implicit regularization can improve on the explicit regularization provided by the regularizer.

Finally, Figure 7 shows that the joint optimization of subsampling along with lasso penalty consistently outperforms the optimization of lasso penalty on the full data (without any subsampling) uniformly across varying data aspect ratios. Particularly, in the dense regime ($s = 0.9$) and modest sparsity ($s = 0.5$) regimes, the ensemble approach leverages implicit regularization to achieve lower prediction risk. In highly sparse scenarios ($s = 0.2$), this effect reduces as one would expect. Overall, we see complementary benefits of optimizing both subsampling and explicit regularization parameters to improve the predictive performance, with the joint optimization being the overall clear winner.

6 Open directions

In this paper, we provide a general risk characterization for the ensemble of regularized M-estimators. The characterization depends on the inverse data aspect ratio $\delta = \lim n/p$, subsample ratio c, and loss and regularizer pair ($\mathsf{loss}, \mathsf{reg}$). We also specialize the results for specific cases of interest, such as the lasso and ridge regression, and analyze various properties related to optimal subsampling and ensembling. Our goal in performing such analysis is to shed light on how subsampling and ensembling influence the risk of the ensemble estimator and the optimal choices of ensemble size and subsample size. The key takeaway is that subsampling and ensembling can be beneficial in terms of reducing the risk of the estimator, and when the subsample size c is optimized, the resulting risk is monotonic in δ for any ensemble size M.

There are several "axes" along which our results on the asymptotics of subagging can be extended. These are all apparent by inspecting the last row of Table 1 (our current results) and contrasting it against the rows above (the known results in specific cases). We briefly mention some of these open directions next to make them explicit. First, our current analysis can handle non-differentiable and non-strongly convex regularizers, but we require a differentiable loss. Extending the analysis to non-differentiable losses through techniques like Moreau smoothing [CMW23, Section B.7] is a promising future direction. In addition to relaxing conditions assumed for risk characterization, we are also interested in relaxing the assumptions for the risk estimation (Theorem 4), particularly the strong convexity assumption on reg. A promising approach in this direction is to apply the Gaussian smoothing technique recently proposed by [BK23a]. Additionally, it is of interest to extend the scope beyond separable regularizers to include non-separable regularizers, which have been studied only for special cases like generalized ridge regression [PD23]. Another potential extension is relaxing the assumption of linear response models. Partial progress in this direction includes the recent work by [BK24] (for logistic models), [CVD+24] (for generalized linear models), and [PD23] (for arbitrary response with bounded moments). Furthermore, while we assume isotropic Gaussian features, it is also of interest to extend the results to general feature distributions and anisotropic covariances. This has been studied in special cases of ridge regression ([PDK23], any M), ridgeless regression ([HX23], $M = 1$), lasso regression ([CMW23], $M = 1$), among others; see also, e.g., [PKLS23, HS23], for some progress towards establishing Gaussian universality (for $M = 1$). Finally, broadening the scope of resampling strategies beyond subsampling without replacement to include more general schemes like sampling with replacement or according to a specific distribution, as recently studied in [CVD+24, DP24a], is another future direction. We hope our current analysis serves as a step towards these open directions in various axes.

Acknowledgements

We thank Michael Celentano and Ryan Tibshirani for useful discussions. PP and JHD acknowledge the computing support provided by the grant MTH230020 for experiments run on the Bridge-2 system at the Pittsburgh Supercomputing Center. PCB acknowledges partial support from the NSF Grant DMS-1945428.

References

[AK23] Ryo Ando and Fumiyasu Komaki. On high-dimensional asymptotic properties of model averaging estimators. *arXiv preprint arXiv:2308.09476*, 2023.

[AP20] Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*, 2020.

[BBEKY13] Derek Bean, Peter J. Bickel, Noureddine El Karoui, and Bin Yu. Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013.

[BDK+24] Pierre C. Bellec, Jin-Hong Du, Takuya Koriyama, Pratik Patil, and Kai Tan. Corrected generalized cross-validation for finite ensembles of penalized estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024.

[Bel21] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[Bel22] Pierre C. Bellec. Observable adjustments in single-index models for regularized M-estimators. *arXiv preprint arXiv:2204.06990*, 2022.

[Bel23] Pierre C. Bellec. Out-of-sample error estimation for M-estimators with convex penalty. *Information and Inference: A Journal of the IMA*, 12(4):2782–2817, 2023.

[BK23a] Pierre C. Bellec and Takuya Koriyama. Error estimation and adaptive tuning for unregularized robust M-estimator. *arXiv preprint arXiv:2312.13257*, 2023.

[BK23b] Pierre C. Bellec and Takuya Koriyama. Existence of solutions to the nonlinear equations characterizing the precise error of M-estimators. *arXiv preprint arXiv:2312.13254*, 2023.

[BK24] Pierre C. Bellec and Takuya Koriyama. Asymptotics of resampling without replacement in robust and logistic regression. *arXiv preprint arXiv:2404.02070*, 2024.

[BM11a] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[BM11b] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[BMR21] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

[Bre96] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[Bre01] Leo Breiman. Using iterated bagging to debias regressions. *Machine Learning*, 45:261–277, 2001.

[BS06] Andreas Buja and Werner Stuetzle. Observations on bagging. *Statistica Sinica*, pages 323–351, 2006.

[BS22] Pierre C. Bellec and Yiwei Shen. Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning. In *Conference on Learning Theory*, 2022.

[BY02] Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

[CM22] Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *The Annals of Statistics*, 50(1):170–196, 2022.

[CM24] Michael Celentano and Andrea Montanari. Correlation adjusted debiased lasso: debiasing the lasso with inaccurate covariate model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024.

[CMW23] Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194–2220, 2023.

[CVD+24] Lucas Clarté, Adrien Vandenbroucque, Guillaume Dalle, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Analysis of bootstrap and subsampling in high-dimensional regularized regression. *arXiv preprint arXiv:2402.13622*, 2024.

[CZYS23] Xin Chen, Yicheng Zeng, Siyue Yang, and Qiang Sun. Sketched ridgeless linear regression: The role of downsampling. In *International Conference on Machine Learning*, 2023.

[DKT22] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

[DM16] David L. Donoho and Andrea Montanari. High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969, 2016.

[DMM09] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[DMM11] David L. Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

[DP24a] Jin-Hong Du and Pratik Patil. Implicit regularization paths of weighted neural representations. *arXiv preprint arxiv:2408.15784*, 2024.

[DP24b] Jin-Hong Du and Pratik Patil. Python package sklearn_ensemble_cv v0.2.3. PyPI, 2024.

[DPK23] Jin-Hong Du, Pratik Patil, and Arun Kumar Kuchibhotla. Subsample ridge ensembles: Equivalences and generalized cross-validation. In *International Conference on Machine Learning*, 2023.

[DPRK24] Jin-Hong Du, Pratik Patil, Kathryn Roeder, and Arun Kumar Kuchibhotla. Extrapolated cross-validation for randomized ensembles. *Journal of Computational and Graphical Statistics*, 2024.

[dRBK20] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, 2020.

[DS01] Kenneth R. Davidson and Stanislaw J. Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the Geometry of Banach Spaces*, 1, 2001.

[DS20] Edgar Dobriban and Yue Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52, 2020.

[DS21] Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918–943, 2021.

[DW18] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[EK13] Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

[EKBB+13] Noureddine El Karoui, Derek Bean, Peter J. Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[FF93] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[FH07] Jerome H. Friedman and Peter Hall. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683, 2007.

[Fu98] Wenjiang J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

[GB23] Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[hp24] Iosif Pinelis (https://mathoverflow.net/users/36721/iosif pinelis). Large deviations: Growth of empirical average of iid non-negative random varialbes with infinite expectations? MathOverflow, 2024. URL:https://mathoverflow.net/q/390939 (version: 2024).

[HS90] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[HS05] Peter Hall and Richard J. Samworth. Properties of bagged nearest neighbour classifiers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):363–379, 2005.

[HS23] Qiyang Han and Yandi Shen. Universality of regularized regression estimators in high dimensions. *The Annals of Statistics*, 51(4):1799–1823, 2023.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. Second edition.

[Hub64] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[HX23] Qiyang Han and Xiaocong Xu. The distribution of ridgeless least squares interpolators. *arXiv preprint arXiv:2307.02044*, 2023.

[JM13] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[Kar18] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175, 2018.

[KS97] Anders Krogh and Peter Sollich. Statistical mechanics of ensemble learning. *Physical Review E*, 55(1):811, 1997.

[LGC+21] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

[LGR+22] Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance and ensemble of learners: Exact asymptotics for convex losses in high-dimension. In *International Conference on Machine Learning*, 2022.

[LJB20] Daniel LeJeune, Hamid Javadi, and Richard Baraniuk. The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[LS22] Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers. *The Annals of Statistics*, 50(3):1669 – 1695, 2022.

[LW21] Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

[MLC19] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *International Conference on Acoustics, Speech and Signal Processing*, 2019.

[MM21] Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

[MRRK22] Nicole Mücke, Enrico Reiss, Jonas Rungenhagen, and Markus Klein. Data-splitting improves statistical performance in overparameterized regimes. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[MRSY19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[OH16] Samet Oymak and Babak Hassibi. Sharp mse bounds for proximal denoising. *Foundations of Computational Mathematics*, 16:965–1029, 2016.

[OTH13] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. In *Allerton Conference on Communication, Control, and Computing*, 2013.

[PD23] Pratik Patil and Jin-Hong Du. Generalized equivalences between subsampling and ridge regularization. In *Advances in Neural Information Processing Systems*, 2023.

[PDK23] Pratik Patil, Jin-Hong Du, and Arun Kumar Kuchibhotla. Bagging in overparameterized learning: Risk characterization and risk monotonization. *Journal of Machine Learning Research*, 24(319):1–113, 2023.

[Per93] Michael Perrone. Putting it all together: Methods for combining neural networks. In *Advances in Neural Information Processing Systems*, 1993.

[PKLS23] Luca Pesce, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Are gaussian data all you need? the extents and limits of universality in high-dimensional generalized linear estimation. In *International Conference on Machine Learning*, 2023.

[PKWR22] Pratik Patil, Arun Kumar Kuchibhotla, Yuting Wei, and Alessandro Rinaldo. Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937*, 2022.

[PL24] Pratik Patil and Daniel LeJeune. Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning. In *International Conference on Learning Representations*, 2024.

[SAH19] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, 2019.

[Sam12] Richard J. Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.

[SC19] Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[SK95] Peter Sollich and Anders Krogh. Learning with ensembles: How overfitting can be useful. In *Advances in Neural Information Processing Systems*, 1995.

[Ste81] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

[TAH18] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[Tib13] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[TOH15] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, 2015.

[VdV00] Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[WMZ18] Haolei Weng, Arian Maleki, and Le Zheng. Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *The Annals of Statistics*, 46(6A):3099 – 3129, 2018.

[WWM20] Shuaiwen Wang, Haolei Weng, and Arian Maleki. Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791 – 2823, 2020.

Supplement

This serves as a supplement to the paper "Precise Asymptotics of Bagging Regularized M-estimators." Below, we provide an outline of the supplement along with a summary of the notation used in the main paper and the supplement.

Organization

Section	Content
Appendix A	Proofs of Theorems 1, 2 and 4 from Section 3
Appendix B	Proof of Proposition 7 and other miscellaneous details from Section 4
Appendix C	Details of arguments in Remark 8 from Section 5
Appendix D	Additional numerical illustrations and experimental details

Table 3: Outline of the supplement.

Specific notation

Notation	Description
$(\boldsymbol{x}_i, y_i)$, $i \in [n]$	Observation vector with the feature vector in $\mathbb{R}^p$ and the response variable in $\mathbb{R}$
$\boldsymbol{X}, \boldsymbol{y}$	Feature matrix in $\mathbb{R}^{n\times p}$ and the response vector in $\mathbb{R}^n$
$\boldsymbol{\theta}$	Signal vector in the linear model in $\mathbb{R}^p$ with entries drawn i.i.d. from the distribution F_θ
$\boldsymbol{z}$	Noise vector in the linear model in $\mathbb{R}^n$ with entries drawn i.i.d. from the distribution F_z
k, M	Size of each subsample and the total number of subsamples (bags) used in the ensemble
I_m, $m \in [M]$	Index set (subset of $[n]$) of the m-th subsample, with $\lvert I_m\rvert = k$
$(\boldsymbol{X}_{I_m}, \boldsymbol{y}_{I_m})$, $m \in [M]$	Feature matrix in $\mathbb{R}^{k\times p}$ and response vector in $\mathbb{R}^k$ of the subsampled dataset indexed by I_m
$(\boldsymbol{X}_{I_m\cap I_\ell}, \boldsymbol{y}_{I_m\cap I_\ell})$, $m \neq \ell \in [M]$	Feature matrix in $\mathbb{R}^{\lvert I_m\cap I_\ell\rvert\times p}$ and response vector in $\mathbb{R}^{\lvert I_m\cap I_\ell\rvert}$ of the overlapped dataset corresponding to the overlap between the subsampled datasets indexed by I_m and I_ℓ
loss, reg, λ	Loss function, regularization function, and regularization level used for the regularized M-estimator
$\widehat{\boldsymbol{\theta}}_m$, $\widetilde{\boldsymbol{\theta}}_M$	Component regularized M-estimator fitted on the m-th subsample and the ensemble estimator
R_M, $\mathcal{R}_M$	Squared prediction risk of the ensemble estimator and its asymptotic limit
$\mathcal{R}_1$, $\mathcal{R}_\infty$	Asymptotic risk of the non-ensemble estimator ($M = 1$) and the full-ensemble estimator ($M = \infty$)
δ, c	Inverse data aspect ratio n/p and subsample ratio k/n
$(\alpha, \beta, \kappa, \nu)$	Parameters characterizing the ensemble risk asymptotics with $M = 1$ (System 1a)
G, H	Standard normal random variables in the ensemble risk asymptotics with $M = 1$ (System 1a)
Θ, Z	Random variables drawn according to distributions F_θ and F_z, respectively (System 1a)
(η_G, η_H)	Correlation parameters in the ensemble risk asymptotics with $M = \infty$ (System 1b)
$G, \widetilde{G}$, $H, \widetilde{H}$	Random variables appearing in the ensemble risk asymptotics with $M = \infty$ (System 1b)
(a, τ)	Parameters in alternate formulation of the ensemble risk asymptotics with $M = 1$ (System 2)
ξ	Parameter alternate formulation of the ensemble risk asymptotics with $M = \infty$ (System 2)

Table 4: Summary of some of the specific notation used in the paper.

A Proofs in Section 3

A.1 Proof of Theorem 1

Part 1. By the Cauchy–Schwarz inequality, if η_G satisfies $|F_{\mathsf{loss}}(\eta_G)| = \sqrt{c\widetilde{c}}$, then we must have

$$\operatorname{sign}(F_{\mathsf{loss}}(\eta_G)) \cdot \mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa) = \mathsf{env}'_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}(\eta_G G + \sqrt{1-\eta_G^2}\overline{G}) + Z; \widetilde{\kappa})$$

with probability 1 for any independent Gaussian $G, \overline{G}$. Multiplying both sides by $\sqrt{1-\eta_G^2}\overline{G}$ and taking expectation, using independence of $(G, \overline{G})$ and the third equation in System 1a, we are left with

$$0 = \sqrt{1-\eta_G^2} \cdot (\widetilde{\nu}\widetilde{\alpha})/(c\delta),$$

which gives $|\eta_G| = 1$ since $\widetilde{\nu}\widetilde{\alpha} > 0$. This means $|F_{\mathsf{loss}}(\eta_G)| < \sqrt{c\widetilde{c}}$ for all $|\eta_G| < 1$. By the same argument, if η_H satisfies $|F_{\mathsf{reg}}(\eta_H)| = 1$, then it holds that

$$\operatorname{sign}(F_{\mathsf{reg}}(\eta_H)) \cdot \Big[\frac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\Big(\frac{\beta}{\nu}H + \Theta; \frac{1}{\nu}\Big) - \frac{\beta}{\nu}H\Big] = \frac{1}{\widetilde{\nu}} \cdot \mathsf{env}'_{\widetilde{\mathsf{reg}}}\Big(\frac{\widetilde{\beta}}{\widetilde{\nu}}(\eta_H H + \sqrt{1-\eta_H^2}\overline{H}) + \Theta; \frac{1}{\widetilde{\nu}}\Big) - \frac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}$$

with probability 1 for any independent Gaussian $(H, \overline{H})$. Multiplying the the both side by $\sqrt{1-\eta_H^2}\overline{H}$ and taking the expectation, using the independence of $(H, \overline{H})$ and the fourth equation in System 1a, we are left with

$$0 = \sqrt{1-\eta_H^2}(-\widetilde{\kappa}\widetilde{\beta}),$$

which gives $|\eta_H| = 1$ since $\widetilde{\kappa}\widetilde{\beta} > 0$. This means $|F_{\mathsf{reg}}(\eta_H)| < 1$ for all $|\eta_H| < 1$.

Part 2. Let us prove the differentiability and contraction of compositions. By System 1a, F_{loss} and F_{reg} can be written as

$$F_{\mathsf{loss}}(\eta_G) = \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}} \cdot \mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa) \cdot \mathsf{env}'_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}\widetilde{G} + Z; \widetilde{\kappa})],$$

$$F_{\mathsf{reg}}(\eta_H) = \frac{1}{\alpha\widetilde{\alpha}} \cdot \mathbb{E}\Big[\Big(\frac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\Big(\frac{\beta}{\nu}H + \Theta; \frac{1}{\nu}\Big) - \frac{\beta}{\nu}H\Big) \cdot \Big(\frac{1}{\widetilde{\nu}} \cdot \mathsf{env}'_{\widetilde{\mathsf{reg}}}\Big(\frac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H} + \Theta; \frac{1}{\widetilde{\nu}}\Big) - \frac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}\Big)\Big].$$

We will use the following lemma to argue the differentiability of F_{loss} and F_{reg}.

Lemma 8. *Let G and Z be independent $\mathcal{N}(0,1)$ random variables. Then, for any Lipschitz functions $(f, \widetilde{f})$ with bounded second moment $\mathbb{E}[f(G)^2], \mathbb{E}[\widetilde{f}(G)^2] < +\infty$, the map*

$$\varphi : [-1, 1] \to \mathbb{R}, \quad \eta \mapsto \mathbb{E}[f(G)\widetilde{f}(\eta G + \sqrt{1-\eta^2}Z)]$$

has the derivative

$$\varphi'(\eta) = \mathbb{E}[f'(G)\widetilde{f}'(\eta G + \sqrt{1-\eta^2}Z)].$$

Proof. Since $\widetilde{f}$ is Lipschitz and $\mathcal{N}(0,1)$ has no point mass, $\widetilde{f}$ is differentiable at $G \sim \mathcal{N}(0,1)$ with probability 1. By the dominated convergence theorem, we have

$$\varphi'(\eta) = \mathbb{E}\big[f(G)\widetilde{f}'\big(\eta G + \sqrt{1-\eta^2}Z\big)\big(G - \frac{\eta}{\sqrt{1-\eta^2}}Z\big)\big].$$

Let us define $A = \eta G + \sqrt{1-\eta^2}Z$ and $B = \sqrt{1-\eta^2}G - \eta Z$ so that (A, B) are independent Gaussian $\mathcal{N}(0,1)$ and $\varphi'(\eta) = (1-\eta^2)^{-1/2}\mathbb{E}[f(\eta A + \sqrt{1-\eta^2}B)\widetilde{f}'(A)B]$. Using Stein's formula for B conditionally on A, we are left with

$$\varphi'(\eta) = \mathbb{E}[\widetilde{f}'(A)f'(\eta A + \sqrt{1-\eta^2}B)] = \mathbb{E}[\widetilde{f}'(\eta G + \sqrt{1-\eta^2}Z)f'(G)],$$

where we have used $(A, \eta A + \sqrt{1-\eta^2}B) \overset{\mathrm{d}}{=} (\eta G + \sqrt{1-\eta^2}Z, G)$. (Here and throughout $\overset{\mathrm{d}}{=}$ refers to equality in distribution.) □

Notice that $\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)$ and $\frac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\left(\frac{\beta}{\nu}H + \Theta; \frac{1}{\nu}\right) - \frac{\beta}{\nu}H$ have finite second moments, thanks to the existence of the solution to System 1a. Thus, applying Lemma 8 with $(f(x), \widetilde{f}(x)) = (\mathsf{env}'_{\mathsf{loss}}(\alpha x + Z; \kappa), \mathsf{env}'_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}x + Z; \widetilde{\kappa}))$ and $(f(x), \widetilde{f}(x)) = (\frac{\beta}{\nu}x - \frac{1}{\nu}\mathsf{env}_{\mathsf{reg}}(\frac{\beta}{\nu}x + \Theta; \frac{1}{\nu}), \frac{\widetilde{\beta}}{\widetilde{\nu}}x - \frac{1}{\widetilde{\nu}}\mathsf{env}_{\widetilde{\mathsf{reg}}}(\frac{\widetilde{\beta}}{\widetilde{\nu}}x + \Theta; \frac{1}{\widetilde{\nu}}))$, we find that F_{loss} and F_{reg} are differentiable and the derivatives are given by:

$$F'_{\mathsf{loss}}(\eta_G) = \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}} \cdot \mathbb{E}[\alpha \mathsf{env}''_{\mathsf{loss}}(\alpha G + Z; \kappa) \cdot \widetilde{\alpha}\mathsf{env}''_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}\widetilde{G} + Z; \widetilde{\kappa})]$$

$$F'_{\mathsf{reg}}(\eta_H) = \frac{1}{\alpha\widetilde{\alpha}}\mathbb{E}\Big[\frac{\beta}{\nu}\Big(1 - \frac{1}{\nu} \cdot \mathsf{env}''_{\mathsf{reg}}\Big(\frac{\beta}{\nu}H + \Theta; \frac{1}{\nu}\Big)\Big) \cdot \frac{\widetilde{\beta}}{\widetilde{\nu}}\Big(1 - \frac{1}{\widetilde{\nu}} \cdot \mathsf{env}''_{\widetilde{\mathsf{reg}}}\Big(\frac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H} + \Theta; \frac{1}{\widetilde{\nu}}\Big)\Big)\Big]$$

for all $\eta_G, \eta_H \in [-1, 1]$. Note that the non-expansiveness of the proximal operator implies that the map $x \mapsto \mathsf{env}'_f(x; \tau) = \tau^{-1}(x - \mathsf{prox}_f(x; \tau))$ is τ^{-1}-Lipschitz and non-decreasing for any convex function f. Thus, $F'_{\mathsf{loss}}(\eta_G)$ and $F'_{\mathsf{reg}}(\eta_H)$ are non-negative and uniformly bounded from above as follows:

$$\begin{aligned}
0 \le F'_{\mathsf{loss}}(\eta_G) &\le \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}}\frac{\alpha}{\kappa} \cdot \mathbb{E}[\widetilde{\alpha}\mathsf{env}''_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}\widetilde{G} + Z; \widetilde{\kappa})] && (0 \le \mathsf{env}''_{\mathsf{loss}}(\alpha G + Z; \kappa) \le \kappa^{-1}) \\
&= \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}}\frac{\alpha}{\kappa} \cdot \mathbb{E}[\widetilde{G} \cdot \mathsf{env}'_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}\widetilde{G} + Z; \widetilde{\kappa})] && (\text{by Stein's lemma}) \\
&= \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}}\frac{\alpha}{\kappa} \cdot \frac{\widetilde{\nu}\widetilde{\alpha}}{\widetilde{c}\delta} = c \cdot \frac{\alpha\widetilde{\alpha}\widetilde{\nu}}{\beta\widetilde{\beta}\kappa} && (\text{using System 1a}); \\
0 \le F'_{\mathsf{reg}}(\eta_H) &\le \frac{1}{\alpha\widetilde{\alpha}}\mathbb{E}\Big[\frac{\beta}{\nu}\Big(1 - \frac{1}{\nu} \cdot \mathsf{env}''_{\mathsf{reg}}\Big(\frac{\beta}{\nu}H + \Theta; \frac{1}{\nu}\Big)\Big)\Big] \cdot \frac{\widetilde{\beta}}{\widetilde{\nu}} && (0 \le \mathsf{env}''_{\mathsf{reg}}(\frac{\widetilde{\beta}}{\widetilde{\nu}}H + \Theta; \frac{1}{\widetilde{\nu}}) \le \widetilde{\nu}) \\
&= \frac{1}{\alpha\widetilde{\alpha}}\frac{\widetilde{\beta}}{\widetilde{\nu}}\mathbb{E}\Big[\Big(\frac{\beta}{\nu}H - \frac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\Big(\frac{\beta}{\nu}H + \Theta; \frac{1}{\nu}\Big)\Big)\Big] && (\text{by Stein's lemma}) \\
&= \frac{1}{\alpha\widetilde{\alpha}}\frac{\widetilde{\beta}}{\widetilde{\nu}} \cdot \beta\kappa = \frac{\beta\widetilde{\beta}\kappa}{\alpha\widetilde{\alpha}\widetilde{\nu}} && (\text{using System 1a}).
\end{aligned}$$

Thus, by the chain rule, noting that $(\beta\widetilde{\beta}\kappa)/(\alpha\widetilde{\alpha}\widetilde{\nu})$ is cancelled out, we have

$$0 \le (F_{\mathsf{loss}} \circ F_{\mathsf{reg}})'(\eta_H) \le c, \quad 0 \le (F_{\mathsf{reg}} \circ F_{\mathsf{loss}})'(\eta_G) \le c.$$

By switching the role of $(\alpha, \beta, \kappa, \nu, c)$ and $(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\kappa}, \widetilde{\nu}, \widetilde{c})$, it also holds that

$$0 \le (F_{\mathsf{loss}} \circ F_{\mathsf{reg}})'(\eta_H) \le \widetilde{c}, \quad 0 \le (F_{\mathsf{reg}} \circ F_{\mathsf{loss}})'(\eta_G) \le \widetilde{c}.$$

Thus, by taking the minimum of $(c, \widetilde{c})$, we find that the compositions

$$\eta_H \mapsto F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H), \quad \eta_G \mapsto F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G)$$

are $(c \wedge \widetilde{c})$-Lipschitz.

Part 3. Now let us show the uniqueness and existence of the solution to the system:

$$\eta_H = F_{\mathsf{loss}}(\eta_G), \quad \eta_G = F_{\mathsf{reg}}(\eta_H).$$

By the assumption $\min\{c, \widetilde{c}\} < 1$ and the fact that $\eta_H \mapsto F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H)$ is $\min\{c, \widetilde{c}\}$-Lipschitz, Banach's fixed-point theorem implies that the fixed-point equation

$$\eta_G = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G), \quad \eta_G \in [-1, 1]$$

admits a unique solution $\eta_G^\star \in [-1, 1]$. If we take $\eta_H^\star = F_{\mathsf{loss}}(\eta_G^\star) \in [-\sqrt{c\widetilde{c}}, \sqrt{c\widetilde{c}}]$, then we have $F_{\mathsf{reg}}(\eta_H^\star) = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G^\star) = \eta_G^\star$ so that $(\eta_G, \eta_H) = (\eta_G^\star, \eta_H^\star)$ satisfies the system: $\eta_H = F_{\mathsf{loss}}(\eta_G)$ and $\eta_G = F_{\mathsf{reg}}(\eta_H)$. This proves the existence of the solution to the system. Let us show the uniqueness. Suppose (η_G, η_H) and $(\widetilde{\eta}_G, \widetilde{\eta}_H)$ satisfy the system. Then, η_G and $\widetilde{\eta}_G$ are the solution to the fixed-point equation $\eta = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta)$:

$$\eta_G = F_{\mathsf{reg}}(\eta_H) = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G), \quad \widetilde{\eta}_G = F_{\mathsf{reg}}(\widetilde{\eta}_H) = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\widetilde{\eta}_G).$$

By the uniqueness of the solution to this fixed-point equation, we must have $\eta_G = \widetilde{\eta}_G$. By the same argument, the contraction $\|F_{\mathsf{loss}} \circ F_{\mathsf{reg}}\|_{\mathrm{Lip}} \le c \wedge \widetilde{c}$ implies that the solution to the fixed-point equation $\eta_H = F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H)$ is unique so that we must have $\eta_H = \widetilde{\eta}_H$. This proves $(\eta_G, \eta_H) = (\widetilde{\eta}_G, \widetilde{\eta}_H)$ and hence the systems admit a unique solution.

Let us show $|\eta_G^\star| < 1$. We proceed by contradiction. If $|\eta_G^\star| = 1$ then noting $\eta_G^\star = F_{\mathsf{reg}}(\eta_H^\star)$, since we have shown in part 1 that $|F_{\mathsf{reg}}(\eta_H)| < 1$ for all $|\eta_H| < 1$, we must have $|\eta_H^\star| = 1$. However, $|\eta_H^\star| = |F_{\mathsf{loss}}(\eta_G^\star)| \le \sqrt{c\widetilde{c}} < 1$ so that this is a contradiction. Thus, we must have $|\eta_G^\star| < 1$. Noting $\eta_H^\star = F_{\mathsf{loss}}(\eta_G^\star)$ and $|F_{\mathsf{loss}}(\eta_G)| < \sqrt{c\widetilde{c}}$ for all $|\eta_G| < 1$, the strict inequality $|\eta_G^\star| < 1$ in turn gives $|\eta_H^\star| < \sqrt{c\widetilde{c}}$.

Part 4. Finally, let us characterize the sign of $(\eta_G^\star, \eta_H^\star)$. Recall that $\eta_G^\star$ and $\eta_H^\star$ satisfies the fixed-point equations $\eta_G - F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G) = 0$ and $\eta_H - F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H) = 0$, respectively, and the maps $\eta_G \mapsto \eta_G - F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G)$ and $\eta_H \mapsto \eta_H - F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H)$ are strictly increasing since $F_{\mathsf{reg}} \circ F_{\mathsf{loss}}$ and $F_{\mathsf{loss}} \circ F_{\mathsf{reg}}$ are $(c \wedge \widetilde{c})$-Lipschitz with $c \wedge \widetilde{c} < 1$. Then, the characterization of sign, i.e., $\mathrm{sign}(F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(0)) = \mathrm{sign}(\eta_H^\star)$ and $\mathrm{sign}(F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(0)) = \mathrm{sign}(\eta_G^\star)$, immediately follows.

A.2 Proof of Theorem 2

In this section, for a vector $\boldsymbol{w}$, the norm $\|\boldsymbol{w}\|$ indicates the ℓ_2 norm unless specified otherwise. By the assumption $0 \in \mathrm{argmin}_x \mathsf{loss}(x) \cap \mathrm{argmin}_x \mathsf{reg}(x)$, taking $\mathsf{loss}_{\mathrm{new}}(x) = \mathsf{loss}(x) - \mathsf{loss}(0)$ and $\mathsf{reg}_{\mathrm{new}}(x) = \mathsf{reg}(x) - \mathsf{reg}(0)$ if necessary, we assume without loss of generality that loss and reg are non-negative and have 0 as a minimizer.

By the change of variable $\boldsymbol{b} \mapsto \boldsymbol{h} = (\boldsymbol{b} - \boldsymbol{\theta})/\sqrt{p}$, denoting $\boldsymbol{G} = \sqrt{p}\boldsymbol{X}$ so that $\boldsymbol{G}$ has i.i.d. $\mathcal{N}(0,1)$ entries, the regularized M-estimator $\widehat{\boldsymbol{\theta}}$ of interest and the residual vector $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}$ can be written as

$$\widehat{\boldsymbol{\theta}} = \sqrt{p}\widehat{\boldsymbol{h}} + \boldsymbol{\theta}, \quad \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{z} - \boldsymbol{G}\widehat{\boldsymbol{h}}$$

where

$$\widehat{\boldsymbol{h}} \in \operatorname*{argmin}_{\boldsymbol{h} \in \mathbb{R}^p} \mathsf{obj}(\boldsymbol{h}) \quad \text{with} \quad \mathsf{obj}(\boldsymbol{h}) := \sum_{i \in I} \mathsf{loss}(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \sum_{j \in [p]} \mathsf{reg}(\sqrt{p} h_j + \theta_j). \tag{27}$$

Throughout this section, we denote $\boldsymbol{\psi} = \sum_{i \in I} \boldsymbol{e}_i \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) \in \mathbb{R}^n$ where $\boldsymbol{e}_i$ are canonical basis of $\mathbb{R}^n$. Let $\widetilde{\boldsymbol{\psi}}, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{h}}$ be the corresponding notation for another. Then, our goal is to show

$$\widehat{\boldsymbol{h}}^\top \widetilde{\boldsymbol{h}} \xrightarrow{\mathrm{p}} \alpha\widetilde{\alpha}\eta_G \quad \text{and} \quad \boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/p \xrightarrow{\mathrm{p}} \beta\widetilde{\beta}\eta_H.$$

A.2.1 Smoothing by adding diminishing ridge penalty

For some positive and diminishing scalar $\mu = \mu_n \to 0$ to be specified later, we define the smoothed regularized M-estimator $\widehat{\boldsymbol{h}}_\mu$ as

$$\widehat{\boldsymbol{h}}_\mu \in \operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^p} \mathsf{obj}(\boldsymbol{h}) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 \tag{28}$$

where $\mathsf{obj}(\boldsymbol{h})$ is the objective function for the original regularized M-estimator $\widehat{\boldsymbol{h}}$ in (27). We denote by $\boldsymbol{\psi}_\mu = \sum_{i\in I} \boldsymbol{e}_i \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}_\mu)$ the smoothed version of $\boldsymbol{\psi}$. This strategy of adding and additive ridge penalty for the mathematical analysis as a first step, and then arguing by approximation ($\mu \to 0$) to study the original $\widehat{\boldsymbol{h}}_0 = \widehat{\boldsymbol{h}}$ is ubiquitous; see, for instance, [Kar18, CM22], [LGR+22, Appendix B.3], [BK23a]. Here, we use the following lemma to show that (28) approximates well the original estimator for any sequence $\mu = \mu_n$ indexed by n and converging to 0.

Lemma 9 (Ridge smoothing)**.** *Let $\boldsymbol{h}_\mu$ be the smoothed M-estimator and $\widehat{\boldsymbol{h}}$ be the original M-estimator.*

1. *(Monotonicity) We have $\|\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{h}}_\mu\|_2^2 \le \|\widehat{\boldsymbol{h}}\|_2^2 - \|\widehat{\boldsymbol{h}}_\mu\|_2^2$ for any $\mu > 0$.*
2. *(Convergence in $\|\cdot\|_2$) As $n, p \to +\infty$ with $n/p \to \delta$, $|I|/n \to c$, and $\mu = \mu_n$ for any $\mu_n \to 0$, we have*

$$\|\widehat{\boldsymbol{h}}_\mu\|_2^2 \xrightarrow{\mathrm{p}} \alpha^2, \quad \frac{\|\boldsymbol{\psi}_\mu\|_2^2}{p} \xrightarrow{\mathrm{p}} \beta^2, \quad \|\widehat{\boldsymbol{h}}_\mu - \widehat{\boldsymbol{h}}\|_2^2 = o_{\mathbb{P}}(1), \quad \frac{\|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\|_2^2}{p} = o_{\mathbb{P}}(1),$$

where α and β are solutions to System 1a.

Proof. By the strong convexity of the ridge term $p\mu^2/2\|\boldsymbol{h}\|^2$, the smoothed one $\widehat{\boldsymbol{h}}_\mu$ also minimizes the convex function: $\boldsymbol{h} \mapsto \mathsf{obj}(\boldsymbol{h}) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 - \frac{p\mu}{2}\|\boldsymbol{h} - \widehat{\boldsymbol{h}}_\mu\|^2$. By the optimality of $\widehat{\boldsymbol{h}}_\mu$ and $\widehat{\boldsymbol{h}}$, we have

$$\mathsf{obj}(\widehat{\boldsymbol{h}}_\mu) + \frac{p\mu}{2}\|\widehat{\boldsymbol{h}}_\mu\|^2 \le \mathsf{obj}(\widehat{\boldsymbol{h}}) + \frac{p\mu}{2}\|\widehat{\boldsymbol{h}}\|^2 - \frac{p\mu}{2}\|\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{h}}_\mu\|^2 \le \mathsf{obj}(\widehat{\boldsymbol{h}}_\mu) + \frac{p\mu}{2}\|\widehat{\boldsymbol{h}}\|^2 - \frac{p\mu}{2}\|\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{h}}_\mu\|^2$$

so that subtracting $\mathsf{obj}(\widehat{\boldsymbol{h}}_\mu)$ from the both sides and dividing by $p\mu/2 > 0$, we obtain the first claim.

Recall $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2$ and $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\mathrm{p}} \beta^2$. Then, by Lemma 9-(1) and the Lipschitz condition of loss', it suffices to show the convergence $\|\widehat{\boldsymbol{h}}_\mu\|^2 \xrightarrow{\mathrm{p}} \alpha^2$ for the smoothed estimator $\widehat{\boldsymbol{h}}_\mu$. Note that $\widehat{\boldsymbol{h}}_\mu$ minimizes the function below

$$\boldsymbol{h} \mapsto \sum_{i\in I} \mathsf{loss}(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \sum_{j\in[p]} \mathsf{reg}_j^{\mu_n}(\sqrt{p}h_j) \quad \text{where} \quad \mathsf{reg}_j^{\mu}(x) := \mathsf{reg}(x + \theta_j) + \mu\frac{x^2}{2}.$$

Now we suppose that for any standard normal $\boldsymbol{g} = (g_j)_{j=1}^p \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$, the convergence of the Moreau envelope

$$\frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j; \tau) - \mathsf{reg}_j^{\mu_n}(0) \xrightarrow{\mathrm{p}} \mathbb{E}[\mathsf{env}_{\mathsf{reg}}(cH + \Theta; \tau) - \mathsf{reg}(\Theta)]$$

holds for all $c \in \mathbb{R}$ and $\tau > 0$. Then, by [TAH18, Theorem 3.1], we have $\|\widehat{\boldsymbol{h}}_\mu\|^2 \xrightarrow{\mathrm{p}} \alpha^2$ and complete the proof. By the weak law of large numbers, the above display holds with $\mu = 0$ (see [TAH18, Lemma 4.1] for details):

$$\frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j; \tau) - \mathsf{reg}_j^{\mu=0}(0) \xrightarrow{\mathrm{p}} \mathbb{E}[\mathsf{env}_{\mathsf{reg}}(cH + \Theta; \tau) - \mathsf{reg}(\Theta)].$$

Then, noting $\mathsf{reg}_j^{\mu=0}(0) = \mathsf{reg}(\theta_j) = \mathsf{reg}_j^{\mu}(0)$ for any $\mu \geq 0$, it suffices to show

$$\frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau) - \frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) = o_{\mathbb{P}}(1).$$

For each j, the monotonicity $\mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) \leq \mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau)$ holds since the objective function is monotone in the sense of $\mathsf{reg}_j^{\mu=0}(x) \leq \mathsf{reg}_j^{\mu}(x)$ for all $x \in \mathbb{R}$ and all $\mu \geq 0$. On the other hand, by the optimality of $\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau)$ and $\mathsf{prox}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau)$, we find

$$\begin{aligned}\mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau) &\leq \frac{1}{2\tau}(cg_j - \mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau))^2 + \mathsf{reg}_j^{\mu_n}(\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau)) \\ &= \mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) + \frac{\mu_n}{2}(\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau))^2\end{aligned}$$

for each $j \in [p]$. Thus, it holds that

$$0 \leq \frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau) - \frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) \leq \frac{\mu_n}{2}\cdot\frac{1}{p}\sum_{j\in[p]}(\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau))^2,$$

where $\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) = \mathsf{prox}_{\mathsf{reg}}(cg_j+\theta_j;\tau) - \theta_j$ by the definition of reg_j^{μ}. Here, under Assumption D-(1), the expectation $\mathbb{E}[(\mathsf{prox}_{\mathsf{reg}}(cg_j+\theta_j;\tau)-\theta_j)^2]$ under $g_j \sim \mathcal{N}(0,1)$ and $\theta_j \sim \Theta$ is finite for all $c \in \mathbb{R}$ and $\tau > 9$ (cf. [TAH18, equation (123)]). Therefore, by the weak law of large numbers, we have $\frac{1}{p}\sum_{j\in[p]}(\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau))^2 \xrightarrow{\mathrm{p}} \mathbb{E}[(\mathsf{prox}_{\mathsf{reg}}(cH+\Theta;\tau)-\Theta)^2] < +\infty$. This means that for any $\mu = \mu_n \to 0$, the RHS of the previous display is $o_{\mathbb{P}}(1)$. □

A.2.2 Bounding norm of regularized M-estimators

For some positive scalar K to be specified later, we add another regularization term; we define $\widehat{\boldsymbol{h}}_{\mu,K}$ as

$$\widehat{\boldsymbol{h}}_{\mu,K} \in \operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^p} \mathsf{obj}(\boldsymbol{h}) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 + \underbrace{\frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{h}\|^2 - K}{2}\Big)}_{\text{additional term}} \quad \text{with} \qquad \widehat{\lambda} := \mathsf{obj}(\boldsymbol{0})$$

where $\mathsf{F}:\mathbb{R}\to\mathbb{R}$ is convex, non-negative, and non-decreasing with $\lim_{x\to+\infty}\mathsf{F}(x) = +\infty$, as well as differentiable with $\mathsf{F}'(u) = 0$ if $u \leq 0$ and $\mathsf{F}'(u) = 1$ if $u \geq 1$. For instance, we may take F as an integral of the smoothed step function as follows:

$$\mathsf{F}(x) := \int_{-\infty}^{x} f(u)\,\mathrm{d}u \quad \text{with} \quad f(u) := \begin{cases} 1 & u \geq 1 \\ 3u^2 - 2u & u \in (0,1) \\ 0 & u \leq 0 \end{cases}$$

Note in passing that the regularization parameter $\widehat{\lambda} = \mathsf{obj}(\boldsymbol{0}) = \sum_{i\in I}\mathsf{loss}(z_i) + \sum_{j\in[p]}\mathsf{reg}(\theta_j)$ is independent of the design matrix $\boldsymbol{G}$ and non-negative since loss and reg are non-negative.

Now we claim that for sufficiently large $K > 0$ this modified estimator $\widehat{\boldsymbol{h}}_{\mu,K}$ coincides with the smoothed estimator $\widehat{\boldsymbol{h}}_{\mu}$. Furthermore, thanks to the additional regularizer $\frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{h}\|^2-K}{2}\Big)$, the norm of $\boldsymbol{h}_{\mu,K}$ and $\boldsymbol{\psi}_{\mu,K}$ are suitably bounded as follows:

Lemma 10. *The following convergences hold.*

1. *If we set $K = 2\alpha^2$ where $\alpha > 0$ is the solution to System 1a, we have*

$$\mathbb{P}(\widehat{\boldsymbol{h}}_{\mu,K} = \widehat{\boldsymbol{h}}_\mu) \xrightarrow{\mathrm{p}} 1.$$

2. *For any $K \geq 0$, there exists a positive constant C_K that only depends on K such that*

$$\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2 \leq C_K, \quad \|\boldsymbol{\psi}_{\mu,K}\|^2 \leq C_K(1 + \|\mathsf{loss}'\|_{\mathrm{Lip}}^2)(\|\mathsf{loss}'(\boldsymbol{z})\|^2 + \|\boldsymbol{G}\|_{\mathrm{op}}^2).$$

Thus, the constant $C^* = 1.1C_K(1 + \|\mathsf{loss}'\|_{\mathrm{Lip}}^2)(\mathbb{E}[\mathsf{loss}'(Z)^2] + (1 + \delta^{-1/2}))^2$ satisfies

$$\mathbb{P}\Big(\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2 \vee (n^{-1}\|\boldsymbol{\psi}_{\mu,K}\|^2) \vee \sup_{m\geq 1} \mathbb{E}\Big[\|\widehat{\boldsymbol{h}}_{\mu,K}\|^{2m} \vee (n^{-1}\|\boldsymbol{\psi}_{\mu,K}\|^2)^m \mid \boldsymbol{z}, \boldsymbol{\theta}\Big]^{1/m} \leq C^*\Big) \to 1.$$

Proof. Let us consider the event $\Omega := \{\|\boldsymbol{h}\|^2 \leq K\}$ with $K = 2\alpha^2$, which holds with high probability, since $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2 > 0$. Combined with the monotonicity $\|\boldsymbol{h}_\mu\|^2 \leq \|\boldsymbol{h}\|^2$ by Lemma 9-(1), we have $\|\widehat{\boldsymbol{h}}_\mu\|^2 \leq K$ under the event Ω. Since $\mathsf{F}'(u) = 0$ for all $u \leq 0$, combined with the Karush-Kuhn-Tucker (KKT) condition $-p\mu\widehat{\boldsymbol{h}}_\mu \in \partial\mathsf{obj}(\widehat{\boldsymbol{h}}_\mu)$ for the smoothed estimator $\widehat{\boldsymbol{h}}_\mu$, we observe that $\widehat{\boldsymbol{h}}_\mu$ satisfies the KKT condition for $\widehat{\boldsymbol{h}}_{\mu,K}$ under the event Ω:

$$-p\mu\widehat{\boldsymbol{h}}_\mu - \widehat{\lambda}\mathsf{F}'\Big(\frac{\|\widehat{\boldsymbol{h}}_\mu\|^2 - K}{2}\Big)\widehat{\boldsymbol{h}}_\mu = -p\mu\widehat{\boldsymbol{h}}_\mu - 0 \cdot \widehat{\boldsymbol{h}}_\mu \in \partial\mathsf{obj}(\widehat{\boldsymbol{h}}_\mu).$$

This implies $\mathbb{P}(\widehat{\boldsymbol{h}}_\mu = \widehat{\boldsymbol{h}}_{\mu,K}) \geq \mathbb{P}(\Omega) \to 1$.

By the non-negativity of $(\mathsf{obj}(\cdot), p\mu\|\cdot\|^2, \frac{\widehat{\lambda}}{2}\mathsf{F}(\cdot))$ and the optimality of $\widehat{\boldsymbol{h}}_{\mu,K}$, it holds that

$$\begin{aligned}
0 \leq \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2 - K}{2}\Big) &\leq \mathsf{obj}(\widehat{\boldsymbol{h}}_{\mu,K}) + \frac{p\mu}{2}\|\boldsymbol{h}_{\mu,K}\|^2 + \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{h}_{\mu,K}\|^2 - K}{2}\Big) \\
&\leq \mathsf{obj}(\boldsymbol{0}) + \frac{p\mu}{2}\|\boldsymbol{0}\|^2 + \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{0}\|^2 - K}{2}\Big) \\
&= \widehat{\lambda} + 0 + 0.
\end{aligned}$$

When $\widehat{\lambda} = 0$ then all inequalities above holds with equality, which means that $\boldsymbol{0}$ minimizes the objective function $\mathsf{obj}(\boldsymbol{h}) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 + \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{h}\|^2 - K}{2}\Big)$ for $\widehat{\boldsymbol{h}}_{\mu,K}$. Since this objective function is strongly convex due to the ridge term, the minimizer is unique. This means $\widehat{\boldsymbol{h}}_{\mu,K} = \boldsymbol{0}$ when $\widehat{\lambda} = 0$. On the other hand, if $\widehat{\lambda} > 0$ then dividing the above display by $\widehat{\lambda} > 0$ we are left with $\mathsf{F}\Big(\frac{\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2 - K}{2}\Big) \leq 2$. Since F is non-decreasing and coercive on the positive side, i.e., $\lim_{x\to+\infty}\mathsf{F}(x) \to +\infty$, this gives $\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2 \leq C(K)$ for a constant $C(K) > 0$ depending on K only. Combined with the Lipschitz condition of loss', the norm of $\boldsymbol{\psi}_{\mu,K} = \sum_{i\in I}\boldsymbol{e}_i\mathsf{loss}'(z_i - \boldsymbol{g}_i^\top\boldsymbol{h}_{\mu,K})$ is bounded as

$$\|\boldsymbol{\psi}_{\mu,K}\| \leq \|\mathsf{loss}'(\boldsymbol{z})\| + \|\mathsf{loss}'(\boldsymbol{z} - \boldsymbol{G}\widehat{\boldsymbol{h}}_{\mu,K}) - \mathsf{loss}'(\boldsymbol{z})\| \leq \|\mathsf{loss}'(\boldsymbol{z})\| + \|\mathsf{loss}'\|_{\mathrm{Lip}}\|\boldsymbol{G}\|_{\mathrm{op}}\|\widehat{\boldsymbol{h}}_{\mu,K}\|.$$

Since $\mathsf{loss}'(z_i)^2$ has a finite second moment by Assumption D-(1), the weak law of large numbers gives $n^{-1}\|\mathsf{loss}'(\boldsymbol{z})\|^2 \xrightarrow{\mathrm{p}} \mathbb{E}[\mathsf{loss}'(Z)^2] < +\infty$. Since $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ is a Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries, we also have $\|\boldsymbol{G}\|_{op}^2/n \xrightarrow{\mathrm{p}} (1 + \delta^{-1/2})^2$ by standard results of the maximal singular value of a Gaussian matrix, e.g., [DS01, Theorem II.13]. This completes the proof. □

A.2.3 Derivative formulae for strongly convex regularizer

Lemma 11 ([BS22]). *Let $\widehat{\boldsymbol{h}} \in \mathbb{R}^p$ be the regularized M-estimator of the form*

$$\widehat{\boldsymbol{h}} \in \operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^p} \sum_{i\in I} \mathsf{loss}(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \mathsf{R}(\boldsymbol{h}),$$

where $I \subset [n]$ is a subset independent of $(\boldsymbol{g}_i)_{i\in[n]}$, $\mathsf{loss} : \mathbb{R} \to \mathbb{R}$ is a convex and differentiable function with Lipschitz derivative, $\mathsf{R} : \mathbb{R}^p \to \mathbb{R}$ is a τ-strongly convex regularizer for some $\tau > 0$, and R is independent from $(\boldsymbol{g}_i)$. Then, there exists a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ satisfying $\|\boldsymbol{A}\|_{\mathrm{op}} \le \tau^{-1}$ and $\mathrm{tr}[\boldsymbol{A}] \ge 0$ such that $\widehat{\boldsymbol{h}} \in \mathbb{R}^p$ and $\boldsymbol{\psi} = \sum_{i\in I} \boldsymbol{e}_i \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \widehat{\boldsymbol{h}}) \in \mathbb{R}^n$ are both differentiable as functions of the design $(\boldsymbol{G}) = (g_{ij})$ with the derivative given by

$$\forall i \in [n], \forall j \in [p], \quad \frac{\partial \widehat{\boldsymbol{h}}}{\partial g_{ij}} = \boldsymbol{A}\Big(\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{\psi} - \boldsymbol{G}^\top \boldsymbol{D} \boldsymbol{e}_i \boldsymbol{e}_j^\top \widehat{\boldsymbol{h}}\Big), \quad \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} = -\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{\psi} - \boldsymbol{V}\boldsymbol{e}_i \boldsymbol{e}_j^\top \widehat{\boldsymbol{h}}. \tag{29}$$

where $\boldsymbol{D}$ and $\boldsymbol{V}$ are $n \times n$ matrices defined by $\boldsymbol{D} = \sum_{i\in I} \boldsymbol{e}_i \boldsymbol{e}_i^\top \mathsf{loss}''(z_i - \boldsymbol{g}_i^\top \widehat{\boldsymbol{h}})$ and $\boldsymbol{V} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}$. Furthermore, the matrix $\boldsymbol{V}$ defined above is positive semidefinite with its operator norm bounded as $\|\boldsymbol{V}\|_{\mathrm{op}} \le \|\mathsf{loss}'\|_{\mathrm{Lip}}$.

Recall that the modified estimator $\widehat{\boldsymbol{h}}_{\mu,K}$ in Appendix A.2.2 minimizes the objective function $\sum_{i\in I} \mathsf{loss}(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \mathsf{R}(\boldsymbol{h})$ where R is the regularizer of the form

$$\mathsf{R}(\boldsymbol{h}) := \sum_{j\in[p]} \mathsf{reg}(\sqrt{p} h_j + \theta_j) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 + \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{h}\|^2 - K}{2}\Big),$$

which is $(p\mu)$-strongly convex. Thus, we can apply Lemma 11; there exists a matrix $\boldsymbol{A}_{\mu,K} \in \mathbb{R}^{p\times p}$ such that

$$\|\boldsymbol{A}_{\mu,K}\|_{\mathrm{op}} \le (p\mu)^{-1}, \quad \mathrm{tr}[\boldsymbol{A}_{\mu,K}] \ge 0, \tag{30}$$

and $\widehat{\boldsymbol{h}}_{\mu,K}$ and $\boldsymbol{\psi}_{\mu,K} = \sum_{i\in I} \boldsymbol{e}_i \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \widehat{\boldsymbol{h}}_{\mu,K})$ are differentiable with respect to the design matrix $\boldsymbol{G} = (g_{ij})$ as in (29) with

$$\boldsymbol{D}_{\mu,K} = \sum_{i\in I} \boldsymbol{e}_i \boldsymbol{e}_i^\top \mathsf{loss}''(z_i - \boldsymbol{g}_i^\top \widehat{\boldsymbol{h}}_{\mu,K}) \qquad \boldsymbol{V}_{\mu,K} = \boldsymbol{D}_{\mu,K} - \boldsymbol{D}_{\mu,K}\boldsymbol{G}\boldsymbol{A}_{\mu,K}\boldsymbol{G}^\top \boldsymbol{D}_{\mu,K}.$$

Now we claim that the trace of $\boldsymbol{V}_{\mu,K}$ and $\boldsymbol{A}_{\mu,K}$ are empirical quantities that converge to the remaining solution ν and κ:

Lemma 12. *For any $\mu \to 0$ such that $\mu^{-1} = O(n^{1/8})$, we have*

$$\mathrm{tr}[\boldsymbol{V}_{\mu,K}]/p \xrightarrow{\mathrm{p}} \nu \quad \text{and} \quad \mathrm{tr}[\boldsymbol{A}_{\mu,K}] \xrightarrow{\mathrm{p}} \kappa.$$

Proof. See Appendix A.4. □

A.2.4 Proof of Theorem 2

Below we will take the diminishing constant $\mu_n \to 0$ as in Lemma 12. Dropping the dependence on (μ, K) for simplicity, we denote $\boldsymbol{h} = \widehat{\boldsymbol{h}}_{\mu,K}$, $\boldsymbol{\psi} = \boldsymbol{\psi}_{\mu,K}$, and $\boldsymbol{V} = \boldsymbol{V}_{\mu,K}$, $\boldsymbol{A} = \boldsymbol{A}_{\mu,K}$. In the same way, we use the notation $(\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{\psi}}, \widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{A}})$ for the other estimator. Then, by the convergence in $\|\cdot\|_2$ from Lemma 9 and Lemma 10, it suffices to show the convergence of correlation for the modified ones:

$$\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha}) \xrightarrow{\mathrm{p}} \eta_G \quad \text{and} \quad \boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \xrightarrow{\mathrm{p}} \eta_H.$$

First, we will argue by the Gaussian Poincaré inequality that $\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha})$ and $\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta})$ concentrate on random quantities that are independent of the scaled design matrix $\boldsymbol{G} = \sqrt{p}\boldsymbol{X}$. Throughout this section, we denote by $\bar{\mathbb{E}}[\cdot] = \mathbb{E}[\cdot|\boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I}]$ the conditional expectation with respect to the design matrix $\boldsymbol{G}$ given signal $\boldsymbol{\theta}$, noise $\boldsymbol{z}$ and subsample index $I, \widetilde{I}$. Since $\boldsymbol{G}$ is independent of $(\boldsymbol{\theta}, \boldsymbol{z}, I, \widetilde{I})$, the conditional distribution of $\boldsymbol{G}$ given $(\boldsymbol{\theta}, \boldsymbol{z}, I, \widetilde{I})$ is still that of a matrix with i.i.d. $\mathcal{N}(0,1)$ entries.

Lemma 13. *For any $\mu_n > 0$ such that $\mu_n^{-1} = o(n^{1/2})$, we have*

$$\bar{\mathbb{E}}\big[(\boldsymbol{h}^\top \widetilde{\boldsymbol{h}} - \bar{\mathbb{E}}[\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}])^2\big] = \mathcal{O}_{\mathbb{P}}(n^{-1}\mu^{-2}) = o_{\mathbb{P}}(1).$$
$$\bar{\mathbb{E}}\big[(\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}} - \bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}])^2\big] = \mathcal{O}_{\mathbb{P}}(n\mu^{-2}) = o_{\mathbb{P}}(n^2).$$

Proof. By the Gaussian Poincaré inequality, noting $\partial_{ij}(\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}) = \widetilde{\boldsymbol{h}}^\top \partial_{ij}\boldsymbol{h} + \boldsymbol{h}^\top \partial_{ij}\widetilde{\boldsymbol{h}}$ with $\partial_{ij} = \frac{\partial}{\partial g_{ij}}$, the conditional variance of $\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}$ is bounded from above as

$$\bar{\mathbb{E}}\Big[(\boldsymbol{h}^\top \widetilde{\boldsymbol{h}} - \bar{\mathbb{E}}[\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}])^2\Big] \le \sum_{ij} \bar{\mathbb{E}}\Big[(\partial_{ij}(\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}))^2\Big] \le 2\sum_{ij} \bar{\mathbb{E}}\Big[(\widetilde{\boldsymbol{h}}^\top \partial_{ij}\boldsymbol{h})^2\Big] + 2\sum_{ij} \bar{\mathbb{E}}\Big[(\boldsymbol{h}^\top \partial_{ij}\widetilde{\boldsymbol{h}})^2\Big],$$

where we denote by $\sum_{ij} = \sum_{i=1}^n \sum_{j=1}^p$ for brevity. By the derivative formula (29), the first term on the RHS is bounded as

$$\begin{aligned}\sum_{ij}\Big(\widetilde{\boldsymbol{h}}^\top \partial_{ij}\boldsymbol{h}\Big)^2 &= \sum_{i,j}\Big(\widetilde{\boldsymbol{h}}^\top \boldsymbol{A}(\boldsymbol{e}_j\psi_i - \boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i h_j)\Big)^2 \\ &\le 2\|\boldsymbol{A}\|_{\mathrm{op}}^2\|\boldsymbol{h}\|^2\|\boldsymbol{\psi}\|^2 + 2\|\boldsymbol{A}\|_{\mathrm{op}}^2\|\boldsymbol{G}\|_{\mathrm{op}}^2\|\boldsymbol{D}\|_{\mathrm{op}}^2\|\widetilde{\boldsymbol{h}}\|^2\|\boldsymbol{h}\|^2.\end{aligned}$$

Using the upper bound $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu_n)^{-1}$ from (30) and the moment bound in Lemma 10-(2), the conditional expectation (with respect to $\bar{\mathbb{E}}$) of the RHS is $\mathcal{O}_{\mathbb{P}}(n^{-1}\mu_n^{-2})$. By symmetry, we also get $\bar{\mathbb{E}}[\sum_{ij} \bar{\mathbb{E}}[(\boldsymbol{h}^\top \partial_{ij}\widetilde{\boldsymbol{h}})^2]] = \mathcal{O}_{\mathbb{P}}(n^{-1}\mu_n^{-2})$.

We use a similar argument for the variance of $\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}$. The Gaussian Poincaré gives $\bar{\mathbb{E}}[(\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}} - \bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}])^2] \le 2\sum_{ij} \bar{\mathbb{E}}[(\widetilde{\boldsymbol{\psi}}^\top \partial_{ij}\boldsymbol{\psi})^2 + (\boldsymbol{\psi}^\top \partial_{ij}\widetilde{\boldsymbol{\psi}})^2]$, where

$$\begin{aligned}\sum_{ij}\Big(\widetilde{\boldsymbol{\psi}}^\top \partial_{ij}\boldsymbol{\psi}\Big)^2 &= \sum_{ij}\Big(\widetilde{\boldsymbol{\psi}}^\top(-\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j\psi_i - \boldsymbol{V}\boldsymbol{e}_i\boldsymbol{w}_j)\Big)^2 \\ &\lesssim \|\boldsymbol{A}^\top \boldsymbol{G}^\top \boldsymbol{D}^\top\|_{\mathrm{op}}^2\|\widetilde{\boldsymbol{\psi}}\|^2\|\boldsymbol{\psi}\|^2 + \|\boldsymbol{V}\|_{\mathrm{op}}^2\|\widetilde{\boldsymbol{\psi}}\|^2\|\boldsymbol{h}\|^2.\end{aligned}$$

by the derivative formula. Using $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu_n)^{-1}$ and the moment bound in Lemma 10-(2), the conditional expectation $\bar{\mathbb{E}}[\cdot]$ of the RHS is $\mathcal{O}_{\mathbb{P}}(n\mu_n^{-2} + n) = \mathcal{O}_{\mathbb{P}}(n\mu_n^{-2})$. □

Let us define $\widehat{\eta}_G$ and $\widehat{\eta}_H$ as the "truncated" values of the conditional expectation of $\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}$ and $\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}$, respectively:

$$\widehat{\eta}_G := \Pi_{[-1,1]}\big(\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]/(\alpha\widetilde{\alpha})\big), \quad \widehat{\eta}_H := \Pi_{[-1,1]}\big(\overline{\mathbb{E}}[\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}]/(p\beta\widetilde{\beta})\big)$$

where $\Pi_{[-1,1]}(x) := \max(\min(x,1),-1)$ is the projection map onto $[-1,1]$. Here, we emphasize that $(\widehat{\eta}_H, \widehat{\eta}_G)$ is independent from the design matrix $\boldsymbol{G}$ since $\boldsymbol{G}$ is integrated out. Furthermore, the absolute values of $\widehat{\eta_G}$ and $\widehat{\eta_H}$ are less than 1 due to the truncation, and hence the two matrices

$$\begin{pmatrix} 1 & \widehat{\eta}_H \\ \widehat{\eta}_H & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & \widehat{\eta}_G \\ \widehat{\eta}_G & 1 \end{pmatrix}$$

are both positive semidefinite. By the concentration from Lemma 13 and the convergence $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2 > 0, \|\boldsymbol{\psi}\|^2/p \xrightarrow{\mathrm{p}} \beta^2 > 0$, noting the truncation $x \mapsto \Pi_{[-1,1]}(x)$ is continuous, we find that the random variables $(\widehat{\eta}_H, \widehat{\eta}_G)$ defined above still capture the correlations, that is,

$$\begin{aligned}
\widehat{\eta}_H &= \Pi_{[-1,1]}\Big(\frac{\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}}{\|\boldsymbol{\psi}\|\|\widetilde{\boldsymbol{\psi}}\|}\Big) + o_{\mathbb{P}}(1) = \frac{\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}}{\|\boldsymbol{\psi}\|\|\widetilde{\boldsymbol{\psi}}\|} + o_{\mathbb{P}}(1).\\
\widehat{\eta}_G &= \Pi_{[-1,1]}\Big(\frac{\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\Big) + o_{\mathbb{P}}(1) = \frac{\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|} + o_{\mathbb{P}}(1).
\end{aligned}$$

where the second equation follows from the fact that the correlations are less than 1 in absolute values by the Cauchy–Schwarz inequality.

Now, we use the multivariate normal approximation below to invite System 1b.

Lemma 14 (Proposition 5.1 in [BK24]). *Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{I}_q)$ and let $\boldsymbol{F} : \mathbb{R}^q \to \mathbb{R}^{q\times M}$ be a locally Lipschitz function with $M \le q$. Then there exists a standard normal vector $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{I}_M)$ such that*

$$\mathbb{E}\Big[\big\|\boldsymbol{F}(\boldsymbol{z})^\top\boldsymbol{z} - \sum_{l\in[q]}\frac{\partial\boldsymbol{F}(\boldsymbol{z})^\top\boldsymbol{e}_l}{\partial z_l} - \big\{\boldsymbol{F}(\boldsymbol{z})^\top\boldsymbol{F}(\boldsymbol{z})\big\}^{1/2}\boldsymbol{w}\big\|^2\Big] \le C_3\sum_{l\in[q]}\mathbb{E}\Big[\big\|\frac{\partial\boldsymbol{F}(z)}{\partial z_l}\big\|_F^2\Big],$$

where $\{\cdot\}^{1/2}$ is the square root of the positive semidefinite matrix.

For each $j \in [p]$, applying Lemma 14 with $\boldsymbol{F} = \Big[\frac{\boldsymbol{\psi}}{\sqrt{p}\beta} \quad \frac{\widetilde{\boldsymbol{\psi}}}{\sqrt{p}\widetilde{\beta}}\Big] \in \mathbb{R}^{n\times 2}$ and $\boldsymbol{z} = \boldsymbol{G}\boldsymbol{e}_j \in \mathbb{R}^n$, using the derivative formula (29), we find that there exists a random vector $\boldsymbol{w}_j \in \mathbb{R}^2$ such that

$$\boldsymbol{w}_j \mid (\boldsymbol{\theta}, \boldsymbol{\epsilon}, \boldsymbol{G}^{-j}, I, \widetilde{I}) \overset{\mathrm{d}}{=} \mathcal{N}(\boldsymbol{0}_2, \boldsymbol{I}_2)$$

and

$$\begin{aligned}
&\overline{\mathbb{E}}\Big[\Big\|\frac{1}{\sqrt{p}}\begin{pmatrix}(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\boldsymbol{V}]h_j + \boldsymbol{\psi}^\top\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j)/\beta\\ (\widetilde{\boldsymbol{\psi}}^\top\boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\widetilde{\boldsymbol{V}}]\widetilde{h}_j + \widetilde{\boldsymbol{\psi}}^\top\widetilde{\boldsymbol{D}}\boldsymbol{G}\widetilde{\boldsymbol{A}}\boldsymbol{e}_j)/\widetilde{\beta}\end{pmatrix} - \begin{pmatrix}\|\boldsymbol{\psi}\|^2/(p\beta^2) & \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta})\\ \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) & \|\widetilde{\boldsymbol{\psi}}\|^2/(p\widetilde{\beta}^2)\end{pmatrix}^{1/2}\boldsymbol{w}_j\Big\|^2\Big]\\
&\le C_4\sum_{i\in[n]}\overline{\mathbb{E}}\Big[\frac{1}{p\beta^2}\|\frac{\partial\boldsymbol{\psi}}{\partial g_{ij}}\|^2 + \frac{1}{p\widetilde{\beta}^2}\|\frac{\partial\widetilde{\boldsymbol{\psi}}}{\partial g_{ij}}\|^2\Big].
\end{aligned}$$

Using $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu_n)^{-1}$ from (30) and Lemma 10, we have

$$\overline{\mathbb{E}}\Big[\|\boldsymbol{\psi}^\top\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\|^2\Big] = \mathcal{O}_{\mathbb{P}}(\mu^{-2}),$$

$$\sum_{ij} \bar{\mathbb{E}}\Big[\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\|^2\Big] \le \bar{\mathbb{E}}\Big[2\|\boldsymbol{DGA}\|_{\mathrm{F}}^2\|\boldsymbol{\psi}\|^2 + 2\|\boldsymbol{V}\|_{\mathrm{F}}^2\|\boldsymbol{h}\|^2\Big] = \mathcal{O}_{\mathbb{P}}(n\mu^{-2}).$$

Thus, summing over $j \in [p]$ in the previous display, noting $\mu^{-2} = o(n)$, we are left with

$$\sum_{j\in[p]} \Big\| \frac{1}{\sqrt{p}} \begin{pmatrix} (\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\boldsymbol{V}]h_j)/\beta \\ (\widetilde{\boldsymbol{\psi}}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\widetilde{\boldsymbol{V}}]h_j)/\widetilde{\beta} \end{pmatrix} - \begin{pmatrix} \|\boldsymbol{\psi}\|^2/(p\beta^2) & \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \\ \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) & \|\widetilde{\boldsymbol{\psi}}\|^2/(p\widetilde{\beta}^2) \end{pmatrix}^{1/2} \boldsymbol{w}_j \Big\|^2 = o_{\mathbb{P}}(n).$$

By the (1/2)-Hölder continuity for the matrix square root: $\|\boldsymbol{M}^{1/2} - \boldsymbol{N}^{1/2}\|_{\mathrm{op}} \le \|\boldsymbol{M} - \boldsymbol{N}\|_{\mathrm{op}}^{1/2}$ for positive semidefinite matrices $\boldsymbol{M}, \boldsymbol{N}$, combined with the convergence $\|\boldsymbol{\psi}\|^2/p \to \beta^2$, $\|\widetilde{\boldsymbol{\psi}}\|^2/p \to \widetilde{\beta}^2$ and the concentration $\widehat{\eta}_{_H} = \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) + o_{\mathbb{P}}(1)$, we have

$$\Big\| \begin{pmatrix} \|\boldsymbol{\psi}\|^2/(p\beta^2) & \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \\ \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) & \|\widetilde{\boldsymbol{\psi}}\|^2/(p\widetilde{\beta}^2) \end{pmatrix}^{1/2} - \begin{pmatrix} 1 & \widehat{\eta}_{_H} \\ \widehat{\eta}_{_H} & 1 \end{pmatrix}^{1/2} \Big\|_{\mathrm{op}} = o_{\mathbb{P}}(1).$$

Combined with the previous display, we get

$$\sum_{j\in[p]} \Big\| \frac{1}{\sqrt{p}} \begin{pmatrix} (\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\boldsymbol{V}]h_j)/\beta \\ (\widetilde{\boldsymbol{\psi}}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\widetilde{\boldsymbol{V}}]\widetilde{h}_j)/\widetilde{\beta} \end{pmatrix} - \begin{pmatrix} 1 & \widehat{\eta}_{_H} \\ \widehat{\eta}_{_H} & 1 \end{pmatrix}^{1/2} \boldsymbol{w}_j \Big\|^2 = o_{\mathbb{P}}(n) + o_{\mathbb{P}}(1)\sum_{j\in[p]} \|\boldsymbol{w}_j\|^2 = o_{\mathbb{P}}(n),$$

where the last equation follows from $\sum_{j\in[p]} \bar{\mathbb{E}}[\|\boldsymbol{w}_j\|^2] = 2p$, which follows from the fact that the marginal distribution of $\boldsymbol{w}_j$ given $(\boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I})$ is $\mathcal{N}(0,1)$ (note, however, that we do not establish or take for granted that $(\boldsymbol{w}_j)_{j\in[p]}$ are i.i.d.). Using $\mathrm{tr}[\boldsymbol{V}]/p \xrightarrow{\mathrm{p}} \nu$ (Lemma 12) and $\|\boldsymbol{h}\|^2 = \mathcal{O}_{\mathbb{P}}(1)$ from Lemma 10, we can also replace $\mathrm{tr}[\boldsymbol{V}]h_j$ by $p\kappa h_j$. As a consequence, we get

$$\sum_{j\in[p]} \Big\| \begin{pmatrix} \frac{1}{\sqrt{p}\beta}(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + p\nu h_j) \\ \frac{1}{\sqrt{p}\widetilde{\beta}}(\widetilde{\boldsymbol{\psi}}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + p\widetilde{\nu}\widetilde{h}_j) \end{pmatrix} - \begin{pmatrix} \widehat{w}_j \\ \widetilde{w}_j \end{pmatrix} \Big\|^2 = o_{\mathbb{P}}(n) \text{ where } \begin{pmatrix} \widehat{w}_j \\ \widetilde{w}_j \end{pmatrix} \coloneqq \begin{pmatrix} 1 & \widehat{\eta}_{_H} \\ \widehat{\eta}_{_H} & 1 \end{pmatrix}^{1/2} \boldsymbol{w}_j. \tag{31}$$

Here, we emphasize that the conditional distribution of $(\widehat{w}_j, \widetilde{w}_j)$ is given by

$$\begin{pmatrix} \widehat{w}_j \\ \widetilde{w}_j \end{pmatrix} \mid (\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{G}^{-j}, I, \widetilde{I}) \overset{\mathrm{d}}{=} \mathcal{N}\Big(\boldsymbol{0}_2, \begin{pmatrix} 1 & \widehat{\eta}_{_H} \\ \widehat{\eta}_{_H} & 1 \end{pmatrix}\Big)$$

since the conditional distribution of $\boldsymbol{w}_j$ given $(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{G}^{-j}, I, \widetilde{I})$ is standard normal $\mathcal{N}(0_2, I_2)$ and $\widehat{\eta}_{_H}$ is $\sigma(\boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I})$-measurable. For each $j \in [p]$, let us define Ξ_j and $\widetilde{\Xi}_j$ as

$$\Xi_j = \frac{\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + p\nu h_j}{\sqrt{p}\beta} - \widehat{w}_j, \quad \widetilde{\Xi}_j = \frac{\widetilde{\boldsymbol{\psi}}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + p\widetilde{\nu}\widetilde{h}_j}{\sqrt{p}\widetilde{\beta}} - \widetilde{w}_j$$

so that the bound (31) reads $\sum_{j\in[p]}[\Xi_j^2 + \widetilde{\Xi}_j^2] = o_{\mathbb{P}}(n)$. By the KKT conditions $\boldsymbol{G}^\top\boldsymbol{\psi} \in \sqrt{p}\partial\mathsf{reg}(\sqrt{p}\boldsymbol{h} + \boldsymbol{\theta})$ and $\boldsymbol{G}^\top\widetilde{\boldsymbol{\psi}} \in \sqrt{p}\partial\widetilde{\mathsf{reg}}(\sqrt{p}\widetilde{\boldsymbol{h}} + \boldsymbol{\theta})$ for $\boldsymbol{h}$ and $\widetilde{\boldsymbol{h}}$, respectively, $\sqrt{p}h_j$ and $\sqrt{p}\widetilde{h}_j$ can be written as

$$\begin{pmatrix} \sqrt{p}h_j \\ \sqrt{p}\widetilde{h}_j \end{pmatrix} = \begin{pmatrix} \mathsf{prox}_{\mathsf{reg}}(\theta_j + \frac{\beta}{\nu}(\widehat{w}_j + \Xi_j); \nu^{-1}) - \theta_j \\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + \frac{\widetilde{\beta}}{\widetilde{\nu}}(\widetilde{w}_j + \widetilde{\Xi}_j); \widetilde{\nu}^{-1}) - \theta_j \end{pmatrix}.$$

Now we claim that the bound $\sum_{j\in[p]}\Xi_j^2 = o_{\mathbb{P}}(n)$ also holds in the conditional expectation $\overline{\mathbb{E}}$, i.e., $\overline{\mathbb{E}}[\sum_{j\in[p]}\Xi_j^2] = o_{\mathbb{P}}(n)$. To this end, it suffices to show $\overline{\mathbb{E}}[(p^{-1}\sum_{j\in[p]}\Xi_j^2)^2] \le C$ with high probability for a constant C. Using the upper estimate

$$0 \le \frac{1}{p}\sum_{j\in[p]}\Xi_j^2 \le 3\Big(\frac{\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2}{\beta^2p^2} + \frac{\nu^2}{\beta^2}\|\boldsymbol{h}\|^2 + \frac{1}{p}\sum_{j\in[p]}(\widehat{w}_j)^2\Big)$$

and the moment bound form Lemma 10, we get

$$\overline{\mathbb{E}}\Big[\Big(\frac{1}{p}\sum_{j\in[p]}\Xi_j^2\Big)^2\Big] \le C_5\Big(1 + \overline{\mathbb{E}}\Big[\Big(\frac{1}{p}\sum_{j\in[p]}(\widehat{w}_j)^2\Big)^2\Big]\Big)$$

with high probability. For the second term, expanding the square of the summation,

$$\overline{\mathbb{E}}\Big[\Big(p^{-1}\sum_{j\in[p]}(\widehat{w}_j)^2\Big)^2\Big] = \frac{1}{p^2}\sum_{i,j}\overline{\mathbb{E}}[(\widehat{w}_i)^2(\widehat{w}_j)^2] \le \frac{1}{p^2}\sum_{i,j}\sqrt{\overline{\mathbb{E}}[(\widehat{w}_i)^4]}\sqrt{\overline{\mathbb{E}}[(\widehat{w}_j)^4]} = 6.$$

where we have used $\overline{\mathbb{E}}[(\widehat{w}_j)^4] = 6$ for all j, which follows from the fact that the marginal law of $\widehat{w}_j$ is $\mathcal{N}(0,1)$. This gives $\overline{\mathbb{E}}[(\frac{1}{p}\sum_{j\in[p]}\Xi_j^2)^2] \le C$ with high probability for a constant C, and hence we get the estimate $\overline{\mathbb{E}}[\frac{1}{p}\sum_{j\in[p]}\Xi_j^2] = o_{\mathbb{P}}(1)$. The same argument yields $\overline{\mathbb{E}}[\frac{1}{p}\sum_{j\in[p]}\widetilde{\Xi}_j^2] = o_{\mathbb{P}}(1)$.

Combined with the proximal representation of $(\sqrt{p}h_j, \sqrt{p}\widetilde{h}_j)$ using $(\Xi_j, \widetilde{\Xi}_j)$, since $\mathsf{prox}_f(\cdot)$ is 1-Lipschitz for any convex function f, the upper bounds of $\overline{\mathbb{E}}[\frac{1}{p}\sum_{j\in[p]}\Xi_j^2]$ and $\overline{\mathbb{E}}[\frac{1}{p}\sum_{j\in[p]}\widetilde{\Xi}_j^2]$ yield the following simple proximal approximation of $(\sqrt{p}h_j, \sqrt{p}\widetilde{h}_j)$:

$$\frac{1}{p}\sum_{j\in[p]}\overline{\mathbb{E}}\Big[\Big\|\begin{pmatrix}\sqrt{p}h_j\\ \sqrt{p}\widetilde{h}_j\end{pmatrix} - \begin{pmatrix}\mathsf{prox}_{\mathsf{reg}}(\theta_j + (\beta/\nu)\widehat{w}_j;\nu^{-1}) - \theta_j\\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + (\widetilde{\beta}/\widetilde{\nu})\widetilde{w}_j;\widetilde{\nu}^{-1}) - \theta_j\end{pmatrix}\Big\|^2\Big] = o_{\mathbb{P}}(1).$$

Noting $\overline{\mathbb{E}}[\|\boldsymbol{h}\|^2] = \mathcal{O}_{\mathbb{P}}(1)$, this lets us approximate $\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]/\alpha\widetilde{\alpha}$ by the inner product of proximal operators:

$$\begin{aligned}\frac{\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]}{\alpha\widetilde{\alpha}} + o_{\mathbb{P}}(1) &= \frac{1}{p}\sum_{j\in[p]}\frac{1}{\alpha\widetilde{\alpha}}\overline{\mathbb{E}}\Big[\Big(\mathsf{prox}_{\mathsf{reg}}(\theta_j + (\beta/\nu)\widehat{w}_j;\nu^{-1}) - \theta_j\Big)\Big(\mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + (\widetilde{\beta}/\widetilde{\nu})\widetilde{w}_j;\widetilde{\nu}^{-1}) - \theta_j\Big)\Big]\\ &= \frac{1}{p}\sum_{j\in[p]}F_{\mathsf{reg}}(\widehat{\eta}_H;\theta_j),\end{aligned}$$

where for the second inequality, we used the fact that the marginal law of $(\widehat{w}_j, \widetilde{w}_j)$ given $(\boldsymbol{\theta}, \boldsymbol{z}, I, \widetilde{I})$ is jointly Gaussian, with zero mean, unit variance, and correlation $\widehat{\eta}_H$, and where $F_{\mathsf{reg}}(\cdot;\theta_j) : [-1,1] \to \mathbb{R}$ is the function defined by

$$F_{\mathsf{reg}}(\eta;\theta_j) = \iint \frac{\big(\mathsf{prox}_{\mathsf{reg}}(\theta_j + \frac{\beta}{\nu}x;\frac{1}{\nu}) - \theta_j\big)\big(\mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + \frac{\widetilde{\beta}}{\widetilde{\nu}}(\eta x + \sqrt{1-\eta^2}y);\frac{1}{\widetilde{\nu}}) - \theta_j\big)}{\alpha\widetilde{\alpha}}\varphi(x)\varphi(y)\,\mathrm{d}x\,\mathrm{d}y$$

with φ being the standard normal probability function and $\iint = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}$. Notice that for each $\eta \in [-1,1]$, the sequence $(F_{\mathsf{reg}}(\eta;\theta_j))_{j=1}^p$ are i.i.d. random variables with mean $\mathbb{E}[F_{\mathsf{reg}}(\eta;\theta_j)] = F_{\mathsf{reg}}(\eta)$. Furthermore, by the Jensen's inequality and the Cauchy–Schwarz inequality, the expectation of the absolute value is finite:

$$\mathbb{E}\Big[|F_{\mathsf{reg}}(\eta;\theta_j)|\Big] \le \frac{1}{\alpha\widetilde{\alpha}}\mathbb{E}\Big[\big|\big(\mathsf{prox}_{\mathsf{reg}}(\Theta + \tfrac{\beta}{\nu}H;\tfrac{1}{\nu}) - \Theta\big)\cdot\big(\mathsf{prox}_{\widetilde{\mathsf{reg}}}(\Theta + \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H};\tfrac{1}{\widetilde{\nu}}) - \Theta\big)\big|\Big]$$

$$\leq \frac{1}{\alpha\widetilde{\alpha}}\mathbb{E}\Big[\big(\mathsf{prox}_{\mathsf{reg}}(\Theta+\tfrac{\beta}{\nu}H;\tfrac{1}{\nu})-\Theta\big)^2\Big]^{1/2}\cdot\mathbb{E}\Big[\big(\mathsf{prox}_{\widetilde{\mathsf{reg}}}(\Theta+\tfrac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H};\tfrac{1}{\widetilde{\nu}})-\Theta\big)^2\Big]^{1/2}$$
$$= 1 \qquad \text{(by (5a) in System 1a)}.$$

Thus, the weak law of large numbers gives $p^{-1}\sum_{j\in[p]} F_{\mathsf{reg}}(\eta;\theta_j) \xrightarrow{\mathrm{p}} F_{\mathsf{reg}}(\eta)$ for each $\eta\in[-1,1]$. Next, we claim that this convergence holds uniformly over $\eta\in[-1,1]$.

Lemma 15. *Let I be a bounded and closed interval of $\mathbb{R}$ and let $(f_n)_{n\geq 1}: I\to\mathbb{R}$ be a sequence of non-decreasing functions such that $f_n(x)\xrightarrow{\mathrm{p}} f(x)$ pointwise for some function $f: I\to\mathbb{R}$. If f is non-decreasing and continuous, then the uniform convergence $\sup_{x\in I}|f_n(x)-f(x)|\xrightarrow{\mathrm{p}} 0$ holds.*

Note that Lemma 15 is a probabilistic analogue of a similar statement for deterministic functions: if a sequence of real-valued monotone functions (on $\mathbb{R}$) converges pointwise to a continuous function on a compact set $I\subset\mathbb{R}$, then the convergence is uniform on the set I. The probabilistic version is known but in a more general setting, so to keep the treatment self-contained we give a basic proof below which is similar to proof of Glivenko-Cantelli theorem (cf. [VdV00, Theorem 19.1]).

Proof. Let us write $I=[a,b]$. Since f is continuous and I is compact, f is uniformly continuous on I. For any $\epsilon>0$, there exists some $\delta_\epsilon>0$ such that $|f(x)-f(y)|<\epsilon/2$ for all $x,y\in I$ such that $|x-y|\leq\delta_\epsilon$. Now for sufficiently large integer $k=k_\epsilon\in\mathbb{N}$ such that $(b-a)/k<\delta_\epsilon$, let us take equally spaced grids $(x_i)_{i=0}^k$ over $[a,b]$ such that $a=x_0<x_1<\cdots<x_k=b$ and $(x_i-x_{i-1})=(b-a)/k$ for all $i\in\{0,\dots k\}$. Let $\Omega_\epsilon=\cap_{i=0}^k\{|f_n(x_i)-f(x_i)|\leq\epsilon/2\}$ be the event under which f_n and f are sufficiently close at the finite grids. Note that this event holds with probability converging to 1, thanks to the pointwise convergence $f_n(x_i)\xrightarrow{\mathrm{p}} f(x_i)$ at finitely many x_i. Then, since f_n is non-decreasing while f does not move more than $\epsilon/2$ in $[x_{i-1},x_i]$, for all $i\in\{0,1,\dots,k\}$ and for all $x\in[x_{i-1},x_i]$ we have

$$f_n(x_{i-1})\leq f_n(x)\leq f_n(x_i),$$
$$-\epsilon/2-f(x_{i-1})\leq -f(x)\leq -f(x_i)+\epsilon/2.$$

In the event Ω_ϵ, summing the two lines it holds that $-\epsilon\leq f_n(x)-f(x)\leq\epsilon$ for all $i\in\{0,1,\dots,k\}$ and for all $x\in[x_{i-1},x_i]$. □

Recall that we have shown in the proof of Theorem 1 that F_{reg} is differentiable with a non-negative derivative. By the same argument, the map $\eta\mapsto F_{\mathsf{reg}}(\eta;\theta_j)$ is differentiable with a non-negative derivative. Thus, applying Lemma 15 with $f_n(\cdot)=p^{-1}\sum_{j\in[p]}F_{\mathsf{reg}}(\cdot;\theta_j)$, $f(\cdot)=F_{\mathsf{reg}}(\cdot)$ and $I=[-1,1]$, we get the uniform convergence:

$$\sup_{\eta\in[-1,1]}\Big|\frac{1}{p}\sum_{j\in[p]}F_{\mathsf{reg}}(\eta;\theta_j)-F_{\mathsf{reg}}(\eta)\Big|\xrightarrow{\mathrm{p}} 0.$$

Combined with $\frac{\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]}{\alpha\widetilde{\alpha}}=\frac{1}{p}\sum_{j\in[p]}\widehat{F}_{\mathsf{reg}}(\widehat{\eta}_H;\theta_j)+o_{\mathbb{P}}(1)$ and $\widehat{\eta}_H\in[-1,1]$, we are left with

$$\frac{\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]}{\alpha\widetilde{\alpha}}=\frac{1}{p}\sum_{j\in[p]}F_{\mathsf{reg}}(\widehat{\eta}_H;\theta_j)+o_{\mathbb{P}}(1)=F_{\mathsf{reg}}(\widehat{\eta}_H)+o_{\mathbb{P}}(1).$$

Recall the definition $\widehat{\eta}_G = \Pi_{[-1,1]}(\frac{\bar{\mathbb{E}}[\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}]}{\alpha\widetilde{\alpha}})$ where $\Pi_{[-1,1]}$ is the projection onto $[-1,1]$. By the continuity of the projection map and the bound $\sup_{\eta\in[-1,1]} |F_{\mathsf{reg}}(\eta)| \le 1$ (see Theorem 1), the above display yields

$$\widehat{\eta}_G = \Pi_{[-1,1]}(F_{\mathsf{reg}}(\widehat{\eta}_H)) + o_{\mathbb{P}}(1) = F_{\mathsf{reg}}(\widehat{\eta}_H) + o_{\mathbb{P}}(1).$$

Next, let us show $\widehat{\eta}_H = F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1)$ using the same argument. Using Lemma 14 with $\boldsymbol{F} = [\boldsymbol{h}/\alpha \mid \widetilde{\boldsymbol{h}}/\widetilde{\alpha}] \in \mathbb{R}^{p\times 2}$ and $\boldsymbol{z} = \boldsymbol{g}_i$, there exists a random vector $\boldsymbol{u}_i \in \mathbb{R}^2$ with conditional distribution

$$\boldsymbol{u}_i \mid \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{G}_{-i}, I, \widetilde{I} \overset{\mathrm{d}}{=} \mathcal{N}(\boldsymbol{0}_2, \boldsymbol{I}_2)$$

and

$$\bar{\mathbb{E}}\Big[\Big\|\begin{pmatrix}(\boldsymbol{g}_i^\top \boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\psi_i + \boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i)/\alpha \\ (\boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \mathrm{tr}[\widetilde{\boldsymbol{A}}]\widetilde{\psi}_i + \widetilde{\boldsymbol{h}}^\top \widetilde{\boldsymbol{A}}\boldsymbol{G}^\top \widetilde{\boldsymbol{D}}\boldsymbol{e}_i)/\widetilde{\alpha}\end{pmatrix} - \begin{pmatrix}\|\boldsymbol{h}\|^2/\alpha^2 & \boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/\alpha\widetilde{\alpha} \\ \widetilde{\boldsymbol{h}}^\top\boldsymbol{h}/\alpha\widetilde{\alpha} & \|\widetilde{\boldsymbol{h}}\|^2/\widetilde{\alpha}^2\end{pmatrix}^{1/2} \boldsymbol{u}_i\Big\|^2\Big]$$
$$\le C_6 \sum_{j=1}^{p} \bar{\mathbb{E}}\frac{1}{\alpha^2}\Big\|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\Big\|^2 + \frac{1}{\widetilde{\alpha}^2}\Big\|\frac{\partial \widetilde{\boldsymbol{h}}}{\partial g_{ij}}\Big\|^2. \quad (32)$$

Using $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ in (30) Lemma 10, noting $\mu^{-2} = o(n)$, it follows that

$$\bar{\mathbb{E}}\big[\|\boldsymbol{h}\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}\|^2\big] = \mathcal{O}_{\mathbb{P}}(n^{-1}\mu^{-2}) = o_{\mathbb{P}}(1)$$
$$\sum_{i\in[n]}\sum_{j\in[p]} \bar{\mathbb{E}}\Big[\|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\|^2\Big] \le 2\bar{\mathbb{E}}\Big[\|\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{\psi}\|^2 + \|\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}\|_F^2\|\boldsymbol{h}\|^2\Big] = \mathcal{O}_{\mathbb{P}}(\mu^{-2}) = o_{\mathbb{P}}(n)$$

so that summing over $i \in [n]$ the inequality (32), we get

$$\sum_{i\in[n]}\Big\|\begin{pmatrix}(\boldsymbol{g}_i^\top \boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\psi_i)/\alpha \\ (\boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \mathrm{tr}[\widetilde{\boldsymbol{A}}]\widetilde{\psi}_i)/\widetilde{\alpha}\end{pmatrix} - \begin{pmatrix}\|\boldsymbol{h}\|^2/\alpha^2 & \boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/\alpha\widetilde{\alpha} \\ \widetilde{\boldsymbol{h}}^\top\boldsymbol{h}/\alpha\widetilde{\alpha} & \|\widetilde{\boldsymbol{h}}\|^2/\widetilde{\alpha}^2\end{pmatrix}^{1/2} \boldsymbol{u}_i\Big\|^2 = o_{\mathbb{P}}(n).$$

Appealing to the (1/2)-Hölder continuity for the matrix square root again, now using the two convergences $\|\boldsymbol{h}\|^2 \to \alpha^2, \|\widetilde{\boldsymbol{h}}\|^2 \xrightarrow{\mathrm{p}} \widetilde{\alpha}^2$ and the concentration $\widehat{\eta}_G = \boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha}) + o_{\mathbb{P}}(1)$, we get

$$\Big\|\begin{pmatrix}\|\boldsymbol{h}\|^2/\alpha^2 & \boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/\alpha\widetilde{\alpha} \\ \widetilde{\boldsymbol{h}}^\top\boldsymbol{h}/\alpha\widetilde{\alpha} & \|\widetilde{\boldsymbol{h}}\|^2/\widetilde{\alpha}^2\end{pmatrix}^{1/2} - \begin{pmatrix}1 & \widehat{\eta}_G \\ \widehat{\eta}_G & 1\end{pmatrix}^{1/2}\Big\|_{\mathrm{op}} = o_{\mathbb{P}}(1).$$

Combined with the convergence $\mathrm{tr}[\boldsymbol{A}] \xrightarrow{\mathrm{p}} \kappa, \mathrm{tr}[\widetilde{\boldsymbol{A}}] \xrightarrow{\mathrm{p}} \widetilde{\kappa}$, noting that $\sum_{i\in[n]}\|\boldsymbol{u}_i\|^2$ and $\|\boldsymbol{\psi}\|^2 + \|\widetilde{\boldsymbol{\psi}}\|^2$ are both $\mathcal{O}_{\mathbb{P}}(n)$, we are left with

$$\frac{1}{n}\sum_{i\in[n]}\Big\|\begin{pmatrix}\boldsymbol{g}_i^\top\boldsymbol{h} - \kappa\psi_i \\ \boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \widetilde{\kappa}\widetilde{\psi}_i\end{pmatrix} - \begin{pmatrix}\alpha\widehat{u}_i \\ \widetilde{\alpha}\widetilde{u}_i\end{pmatrix}\Big\|^2 = o_{\mathbb{P}}(1) \quad \text{where} \quad \begin{pmatrix}\widehat{u}_i \\ \widetilde{u}_i\end{pmatrix} = \begin{pmatrix}1 & \widehat{\eta}_G \\ \widehat{\eta}_G & 1\end{pmatrix}^{1/2}\boldsymbol{u}_i.$$

By the same argument that we used to bound Ξ_j and $\widetilde{\Xi}_j$, using Lemma 10 and the fact that the marginal law of $\widehat{u}_i$ (and $\widetilde{u}_i$) is $\mathcal{N}(0,1)$, we can show that the conditional expectation $\bar{\mathbb{E}}$ of the square of LHS is bounded from above by a constant C with high probability. Thus, the above approximation also holds in the conditional expectation $\bar{\mathbb{E}}$:

$$\frac{1}{n}\sum_{i\in[n]}\bar{\mathbb{E}}\Big[\Big\|\begin{pmatrix}\boldsymbol{g}_i^\top\boldsymbol{h} - \kappa\psi_i \\ \boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \widetilde{\kappa}\widetilde{\psi}_i\end{pmatrix} - \begin{pmatrix}\alpha\widehat{u}_i \\ \widetilde{\alpha}\widetilde{u}_i\end{pmatrix}\Big\|^2\Big] = o_{\mathbb{P}}(1).$$

Let us define $\Xi^i := \boldsymbol{g}_i^\top \boldsymbol{h} - \kappa\psi_i - \alpha\widehat{u}_i$ and $\widetilde{\Xi}^i := \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \widetilde{\kappa}\widetilde{\psi}_i - \widetilde{\alpha}\widetilde{u}_i$ so that the above display reads $\sum_{i\in[n]} \bar{\mathbb{E}}[(\Xi^i)^2 + (\widetilde{\Xi}^i)^2] = o(n)$. Since $\psi_i = \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h})$ and $\widetilde{\psi}_i = \widetilde{\mathsf{loss}}'(\widetilde{z}_i - \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}})$ for all $i \in I \cap \widetilde{I}$, the residuals can be written as

$$z_i - \boldsymbol{g}_i^\top \boldsymbol{h} = \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i - \Xi^i; \kappa), \quad z_i - \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} = \mathsf{prox}_{\widetilde{\mathsf{loss}}}(z_i - \widetilde{\alpha}\widetilde{u}_i - \widetilde{\Xi}^i; \widetilde{\kappa})$$

for all $i \in I \cap \widetilde{I}$. Since the map $x \mapsto \mathsf{env}'_f(x;\tau) = f' \circ \mathsf{prox}_f(x;\tau)$ is a composition of Lipschitz functions if f is convex and differentiable with Lipschitz derivative, the moment bound $\sum_{i\in[n]} \bar{\mathbb{E}}[(\Xi^i)^2 + (\widetilde{\Xi}^i)^2] = o(n)$ lets us approximate ψ_i and $\widetilde{\psi}_i$ by $\mathsf{env}'_{\mathsf{loss}}$ and $\mathsf{env}'_{\widetilde{\mathsf{loss}}}$ as follows:

$$\frac{1}{n}\sum_{i\in I\cap\widetilde{I}} \bar{\mathbb{E}}\left\| \begin{pmatrix} \psi_i - \mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \kappa) \\ \widetilde{\psi}_i - \mathsf{env}'_{\widetilde{\mathsf{loss}}}(z_i - \widetilde{\alpha}\widetilde{u}_i; \widetilde{\kappa}) \end{pmatrix} \right\|^2 = o_{\mathbb{P}}(1) \quad \text{with} \quad \begin{pmatrix} \widehat{u}_i \\ \widetilde{u}_i \end{pmatrix} | \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{G}_{-i}, I, \widetilde{I} \overset{\mathrm{d}}{=} \mathcal{N}(0_2, \begin{pmatrix} 1 & \widehat{\eta}_G \\ \widehat{\eta}_G & 1 \end{pmatrix})$$

Noting $\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2] = \mathcal{O}_{\mathbb{P}}(n)$ by Lemma 10, this lets us approximate $\bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}]/(p\beta\widetilde{\beta})$ by the inner product of $(\mathsf{env}'_{\mathsf{loss}}, \mathsf{env}'_{\widetilde{\mathsf{loss}}})$:

$$\begin{aligned} \frac{\bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}]}{p\beta\widetilde{\beta}} &= \frac{|I\cap\widetilde{I}|}{p\beta\widetilde{\beta}} \frac{1}{|I\cap\widetilde{I}|} \bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}] \\ &= \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}} \frac{1}{|I\cap\widetilde{I}|} \sum_{i\in I\cap\widetilde{I}} \bar{\mathbb{E}}[\mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \kappa) \cdot \mathsf{env}'_{\widetilde{\mathsf{loss}}}(z_i - \widetilde{\alpha}\widetilde{u}_i; \widetilde{\kappa})] + o_{\mathbb{P}}(1). \end{aligned}$$

Since the marginal distribution of $(u_i, \widetilde{u}_i)$ given $(\boldsymbol{\theta}, \boldsymbol{z}, I, \widetilde{I})$ is centered normal with unit variance and correlation $\widehat{\eta}_G$, the above display reads

$$\frac{\bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}]}{p\beta\widetilde{\beta}} = \frac{1}{|I\cap\widetilde{I}|} \sum_{i\in I\cap\widetilde{I}} F_{\mathsf{loss}}(\widehat{\eta}_G; z_i) + o_{\mathbb{P}}(1)$$

where $F_{\mathsf{loss}}(\eta; z_i) : [-1,1] \to \mathbb{R}$ is the function defined as:

$$F_{\mathsf{loss}}(\eta; z_i) = \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x)\varphi(y) \mathsf{env}'_{\mathsf{loss}}(z_i + \alpha x; \kappa) \mathsf{env}'_{\widetilde{\mathsf{loss}}}(z_i + \widetilde{\alpha}(\eta x + \sqrt{1-\eta^2}y); \widetilde{\kappa})\, \mathrm{d}x\, \mathrm{d}y.$$

By the same argument for $F_{\mathsf{reg}}(\eta; \theta_j)$, the sequence $(F_{\mathsf{loss}}(\eta; z_i))_{i\in[n]}$ are i.i.d. random variables with mean $\mathbb{E}[F_{\mathsf{loss}}(\eta; z_i)] = F_{\mathsf{loss}}(\eta)$ and the expectation of the absolute value $|F_{\mathsf{loss}}(\eta; z_i)|$ is finite. Thus, by the weak law of large numbers, we have $\frac{1}{m}\sum_{i\in[m]} F_{\mathsf{loss}}(\eta; z_i) \xrightarrow{\mathrm{p}} \mathbb{E}[F_{\mathsf{loss}}(\eta; z_1)] = F_{\mathsf{loss}}(\eta)$ for any deterministic integer $m = m_n \to +\infty$. Then, we have that for any $\epsilon > 0$, denoting by $\sum_K$ the sum over all possible value K taken by $I \cap \widetilde{I}$,

$$\begin{aligned} &\mathbb{P}\Big(|\frac{1}{|I\cap\widetilde{I}|}\sum_{i\in I\cap\widetilde{I}} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big) \\ &= \sum_K \mathbb{P}\Big(|\frac{1}{|K|}\sum_{i\in K} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon,\ I\cap\widetilde{I} = K\Big) \qquad \text{(by additivity of disjoint events)} \\ &= \sum_K \mathbb{P}\Big(|\frac{1}{|K|}\sum_{i\in K} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big)\mathbb{P}(I\cap\widetilde{I} = K) \qquad \text{(by independence)} \\ &= \sum_{m\geq 0} \mathbb{P}\Big(|\frac{1}{m}\sum_{i\in[M]} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big)\mathbb{P}(|I\cap\widetilde{I}| = m) \quad \text{(since } (z_i)_{i\in K} \overset{\mathrm{d}}{=} (z_i)_{i\in[m]} \text{ for } m = |K|). \end{aligned}$$

We now split the sum over $m \geq 0$ into two as follows:

$$\begin{aligned}
&\mathbb{P}\Big(|\frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big) \\
&= \Big(\sum_{m \leq nc\widetilde{c}/2} + \sum_{m > nc\widetilde{c}/2}\Big)\mathbb{P}\Big(|\frac{1}{m} \sum_{i \in [M]} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big)\mathbb{P}\Big(|I \cap \widetilde{I}| = m\Big) \\
&\leq \mathbb{P}(|I \cap \widetilde{I}| \leq nc\widetilde{c}/2) + \sup_{m > nc\widetilde{c}/2} \mathbb{P}\Big(|\frac{1}{m} \sum_{i \in [M]} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big)\mathbb{P}(|I \cap \widetilde{I}| > nc\widetilde{c}/2)
\end{aligned}$$

where $\mathbb{P}(|I \cap \widetilde{I}| > nc\widetilde{c}/2) \leq 1$ in the rightmost term. The second term converges to 0 by the weak law of large numbers, and the first by Chebyshev's inequality applied to the hypergeometric distribution.

This gives $\frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\eta; z_i) \xrightarrow{\mathrm{p}} F_{\mathsf{loss}}(\eta)$ pointwise for any $\eta \in [-1, 1]$. By the same argument we used for F_{reg} and $F_{\mathsf{reg}}(\cdot; \theta_j)$, F_{loss} and $F_{\mathsf{loss}}(\cdot; z_i)$ are non-decreasing and continuous. Thus, we can apply Lemma 15 with $f_n(\cdot) = \frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\cdot; z_i)$, $f(\cdot) = F_{\mathsf{loss}}(\cdot)$ and obtain the uniform convergence:

$$\sup_{\eta \in [-1,1]} \Big|\frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)\Big| = o_{\mathbb{P}}(1).$$

Combined with $\frac{\bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}]}{p\beta\widetilde{\beta}} = \frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\widehat{\eta}_G; z_i)$ and $\widehat{\eta}_G \in [-1, 1]$, we are left with

$$\bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}]/(p\beta\widetilde{\beta}) = F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1).$$

Recalling $\widehat{\eta}_H = \Pi_{[-1,1]}(\bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}]/(p\beta\widetilde{\beta}))$ where $\Pi_{[-1,1]}$ is the projection onto $[-1, 1]$, by the continuity of the projection map and $|F_{\mathsf{loss}}(\eta)| \leq \sqrt{c\widetilde{c}} \leq 1$ for all $\eta \in [-1, 1]$ (see Theorem 1), we finally obtain

$$\widehat{\eta}_H = \Pi_{[-1,1]}(F_{\mathsf{loss}}(\widehat{\eta}_G)) + o_{\mathbb{P}}(1) = F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1).$$

In summary, we have shown that $\widehat{\eta}_G = F_{\mathsf{reg}}(\widehat{\eta}_H) + o_{\mathbb{P}}(1)$ and $\widehat{\eta}_H = F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1)$. By the continuity of F_{reg} and F_{loss}, it holds that

$$\widehat{\eta}_G = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1) \quad \text{and} \quad \widehat{\eta}_H = F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\widehat{\eta}_H) + o_{\mathbb{P}}(1).$$

Since $F_{\mathsf{reg}} \circ F_{\mathsf{loss}}$ and $F_{\mathsf{loss}} \circ F_{\mathsf{reg}}$ are $(c \wedge \widetilde{c})$-Lipschitz with $c \wedge \widetilde{c} < 1$, we have

$$|\widehat{\eta}_G - \eta_G| = |F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1) - F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G)| \leq (c \wedge \widetilde{c})|\widehat{\eta}_G - \eta_G| + o_{\mathbb{P}}(1),$$

so $\widehat{\eta}_G = \eta_G + o_{\mathbb{P}}(1)$, and similarly $\widehat{\eta}_H = \eta_H + o_{\mathbb{P}}(1)$ for the fixed points η_G and η_H satisfying $\eta_G = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G)$ and $\eta_H = F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H)$. Recalling $\widehat{\eta}_G = \boldsymbol{h}^\top \widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha}) + o_{\mathbb{P}}(1)$ and $\widehat{\eta}_H = \boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) + o_{\mathbb{P}}(1)$, this gives $\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha}) \xrightarrow{\mathrm{p}} \eta_G$ and $\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \xrightarrow{\mathrm{p}} \eta_H$, thereby completing the proof.

Finally, let us show the proximal approximation of estimators $(\sqrt{p}h_j, \sqrt{p}\widetilde{h}_j)$ and residuals $(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}, z_i - \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}})$. Using the convergence $\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \xrightarrow{\mathrm{p}} \eta_H$ that we have shown, applying the above

argument with $\widehat{\eta}_H$ replaced by η_H, we have

$$\frac{1}{p}\sum_{j\in[p]}\overline{\mathbb{E}}\Big[\underbrace{\Big\|\begin{pmatrix}\sqrt{p}h_j\\ \sqrt{p}\widetilde{h}_j\end{pmatrix}-\begin{pmatrix}\mathsf{prox}_{\mathsf{reg}}(\theta_j+(\beta/\nu)\widehat{w}_j;\nu^{-1})-\theta_j\\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j+(\widetilde{\beta}/\widetilde{\nu})\widetilde{w}_j;\widetilde{\nu}^{-1})-\theta_j\end{pmatrix}\Big\|^2}_{\mathrm{LHS}_j}\Big]=o_{\mathbb{P}}(1).$$

Since $|X_n|=o_{\mathbb{P}}(1)$ is equivalent to $\mathbb{E}[1\wedge|X_n|]=o(1)$ for any random variable X_n, the above display reads

$$\mathbb{E}\Big[1\wedge\frac{1}{p}\sum_{j\in[p]}\overline{\mathbb{E}}[\mathrm{LHS}_j]\Big]=o(1)$$

where the expectation $\mathbb{E}$ is with respect to $(\boldsymbol{z},\boldsymbol{\theta},I,\widetilde{I})$ (recall the definition $\overline{\mathbb{E}}[\cdot]=\mathbb{E}[\cdot|\boldsymbol{z},\boldsymbol{\theta},I,\widetilde{I}]$). Note that the integrand is bounded from below as

$$1\wedge\frac{1}{p}\sum_{j\in[p]}\overline{\mathbb{E}}[\mathrm{LHS}_j]\geq\frac{1}{p}\sum_{j\in[p]}\Big(1\wedge\overline{\mathbb{E}}[\mathrm{LHS}_j]\Big)\geq\frac{1}{p}\sum_{j\in[p]}\overline{\mathbb{E}}[1\wedge\mathrm{LHS}_j],$$

where the second inequality follows from the Jensen's inequality $\overline{\mathbb{E}}[f(X)]\leq f(\overline{\mathbb{E}}[X])$ applied with the concave function $f(x)=1\wedge x$ and $X=\mathrm{LHS}_j$. Taking the expectation of the above display and using the tower property, we are left with

$$\frac{1}{p}\sum_{j\in[p]}\mathbb{E}[1\wedge\mathrm{LHS}_j]=o(1).$$

Since the marginal distribution of the integrand $(1\wedge\mathrm{LHS}_j)$ is the same for all $j\in[p]$ by symmetry, the LHS equals to $\mathbb{E}[1\wedge\mathrm{LHS}_{j'}]$ for any $j'\in[p]$. This gives $\max_{j\in[p]}\mathbb{E}[1\wedge\mathrm{LHS}_j]=o(1)$ and completes the proof of the joint approximation of estimators $(\sqrt{p}h_j,\sqrt{p}\widetilde{h}_j)$. The joint approximation of residual follows from the same argument so we omit the proof.

A.3 Proof of Theorem 4

We assume that reg and $\widetilde{\mathsf{reg}}$ are μ-strongly convex for a fixed $\mu>0$. Then [BS22, Theorem 5.1] implies that

$$\mathrm{tr}[\boldsymbol{A}]\,\mathrm{tr}[\boldsymbol{V}]-\mathsf{df}=\mathcal{O}_{\mathbb{P}}(\sqrt{n}).$$

On the other hand, by the same argument in the proof of Lemma 12 with the diminishing ridge penalty μ_n replaced by the strongly convexity parameter $\mu>0$, we have $\mathrm{tr}[\boldsymbol{V}]/p\xrightarrow{\mathrm{p}}\nu>0$ and $\mathrm{tr}[\boldsymbol{A}]\xrightarrow{\mathrm{p}}\kappa$. Therefore, we get

$$\mathrm{tr}[\boldsymbol{A}]-\mathsf{df}/\,\mathrm{tr}[\boldsymbol{V}]=\mathcal{O}_{\mathbb{P}}(n^{-1/2})\quad\text{and}\quad\mathsf{df}/\,\mathrm{tr}[\boldsymbol{V}]=\mathcal{O}_{\mathbb{P}}(1).$$

Note in passing that the same things hold for $\widetilde{\mathsf{df}}$ and $\mathrm{tr}[\widetilde{\boldsymbol{V}}]$. By the Cauchy–Schwarz inequality, the error term due to the replacement of $(\mathsf{df}/\,\mathrm{tr}[\boldsymbol{V}],\widetilde{\mathsf{df}}/\,\mathrm{tr}[\widetilde{\boldsymbol{V}}])$ by $(\mathrm{tr}[\boldsymbol{A}],\mathrm{tr}[\widetilde{\boldsymbol{A}}])$ is

$$\Big|\sum_{i\in[n]}\Big(r_i+\frac{\mathsf{df}}{\mathrm{tr}[\boldsymbol{V}]}\,\mathbb{1}_{\{i\in I\}}\,\psi_i\Big)^\top\Big(\widetilde{r}_i+\frac{\widetilde{\mathsf{df}}}{\mathrm{tr}[\widetilde{\boldsymbol{V}}]}\,\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i\Big)-\sum_{i\in[n]}\Big(r_i+\mathrm{tr}[\boldsymbol{A}]\,\mathbb{1}_{\{i\in I\}}\,\psi_i\Big)^\top\Big(\widetilde{r}_i+\mathrm{tr}[\widetilde{\boldsymbol{A}}]\,\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i\Big)\Big|$$

$$\leq\sqrt{\sum_{i\in[n]}\big(r_i+\frac{\mathsf{df}}{\mathrm{tr}[\boldsymbol{V}]}\,\mathbb{1}_{\{i\in I\}}\,\psi_i\big)^2}\cdot\Big|\frac{\widetilde{\mathsf{df}}}{\mathrm{tr}[\widetilde{\boldsymbol{V}}]}-\mathrm{tr}[\widetilde{\boldsymbol{A}}]\Big|\|\widetilde{\boldsymbol{\psi}}\|+\sqrt{\sum_{i\in[n]}(\widetilde{r}_i+\mathrm{tr}[\widetilde{\boldsymbol{A}}]\,\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i)^2}\cdot\Big|\frac{\mathsf{df}}{\mathrm{tr}[\boldsymbol{V}]}-\mathrm{tr}[\boldsymbol{A}]\Big|\|\boldsymbol{\psi}\|$$

$= (\|\boldsymbol{z}\| + \mathcal{O}_{\mathbb{P}}(\sqrt{n})) \cdot \mathcal{O}_{\mathbb{P}}(1)$

where the last equality follows from the following fact: $\|\boldsymbol{r}\|^2 \le 2(\|\boldsymbol{z}\|^2 + \|\boldsymbol{G}\boldsymbol{h}\|^2)$, $\|\boldsymbol{G}\boldsymbol{h}\|^2 = \mathcal{O}_{\mathbb{P}}(n)$, $\|\boldsymbol{\psi}\|^2 = \mathcal{O}_{\mathbb{P}}(n)$, and $\mathsf{df}/\operatorname{tr}[\boldsymbol{V}] = \mathcal{O}_{\mathbb{P}}(1)$. Therefore, it suffcies to show

$$n\boldsymbol{h}^\top \widetilde{\boldsymbol{h}} + \|\boldsymbol{z}\|^2 - \sum_{i\in[n]} (z_i - \boldsymbol{g}_i^\top \boldsymbol{h} + \operatorname{tr}[\boldsymbol{A}]\, \mathbb{1}_{\{i\in I\}}\, \psi_i)(z_i - \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} + \operatorname{tr}[\widetilde{\boldsymbol{A}}]\, \mathbb{1}_{\{i\in \widetilde{I}\}}\, \widetilde{\psi}_i) = \mathcal{O}_{\mathbb{P}}(1)\|\boldsymbol{z}\| + \mathcal{O}_{\mathbb{P}}(\sqrt{n}).$$

By simple algebra, the LHS can be decomposed into three terms $\xi_1 + \widetilde{\xi}_1 + \xi_2$ with:

$$\xi_1 = \sum_{i\in[n]} z_i(\boldsymbol{g}_i^\top \boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\, \mathbb{1}_{\{i\in I\}}\, \psi_i), \qquad \widetilde{\xi}_1 = \sum_{i\in[n]} z_i(\boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \operatorname{tr}[\widetilde{\boldsymbol{A}}]\, \mathbb{1}_{\{i\in \widetilde{I}\}}\, \widetilde{\psi}_i)$$
$$\xi_2 = n\boldsymbol{h}^\top \widetilde{\boldsymbol{h}} - \sum_{i\in[n]} \big(\boldsymbol{g}_i^\top \boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\, \mathbb{1}_{\{i\in I\}}\, \psi_i\big)\big(\boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \operatorname{tr}[\widetilde{\boldsymbol{A}}]\, \mathbb{1}_{\{i\in \widetilde{I}\}}\, \widetilde{\psi}_i\big)$$

For $(\xi_1, \widetilde{\xi}_1)$, the derivative formula (35) and the argument in the proof of [BS22, Proposition 18] yield

$$\overline{\mathbb{E}}\Big[\frac{|\xi_1|}{\{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2\}^{1/2} \cdot \|\boldsymbol{z}\|}\Big] + \overline{\mathbb{E}}\Big[\frac{|\widetilde{\xi}_1|}{\{\|\widetilde{\boldsymbol{h}}\|^2 + p^{-1}\|\widetilde{\boldsymbol{\psi}}\|^2\}^{1/2} \cdot \|\boldsymbol{z}\|}\Big] \le C(\mu, \delta).$$

Since the denominators are $\mathcal{O}_{\mathbb{P}}(1)\|\boldsymbol{z}\|$, we have $\xi_1 + \widetilde{\xi}_1 = \mathcal{O}_{\mathbb{P}}(1)\|\boldsymbol{z}\|$.

Below we bound ξ_2 using the following moment inequality, which is a variant of [Bel23, Theorem 7.2].

Lemma 16. *Let $\boldsymbol{\rho}, \widetilde{\boldsymbol{\rho}} : \mathbb{R}^{K\times Q} \to \mathbb{R}^Q$ be locally Lipschitz functions and $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} : \mathbb{R}^{K\times Q} \to \mathbb{R}^L$ be locally Lipschitz functions. If $(\boldsymbol{z}_k)_{k\in[K]}$ are i.i.d. $\mathcal{N}(\boldsymbol{0}_Q, \boldsymbol{I}_Q)$, we have*

$$\mathbb{E}\Big|\frac{K\boldsymbol{\rho}^\top \widetilde{\boldsymbol{\rho}} - \sum_{k\in[K]}(\boldsymbol{z}_k^\top \boldsymbol{\rho} - \sum_{q\in Q} \frac{\partial \rho_q}{z_{kq}})(\boldsymbol{z}_k^\top \widetilde{\boldsymbol{\rho}} - \sum_{q\in Q} \frac{\partial \widetilde{\rho}_q}{z_{kq}})}{\{\|\boldsymbol{\rho}\|^2 + \|\boldsymbol{\zeta}\|^2\}^{1/2}\{\|\widetilde{\boldsymbol{\rho}}\|^2 + \|\widetilde{\boldsymbol{\zeta}}\|^2\}^{1/2}}\Big| \le C_7(\sqrt{K}(1 + \sqrt{\mathbb{E}[\Xi + \widetilde{\Xi}]}) + \mathbb{E}[\Xi + \widetilde{\Xi}])$$

where $\Xi := \frac{1}{\|\boldsymbol{\rho}\|^2 + \|\boldsymbol{\zeta}\|^2} \sum_{k\in[K]} \sum_{q\in[Q]} \Big(\big\|\frac{\partial \boldsymbol{\rho}}{\partial g_{kq}}\big\|^2 + \big\|\frac{\partial \boldsymbol{\zeta}}{\partial g_{kq}}\big\|^2\Big)$.

(The proof of Lemma 16 is given in Appendix A.3.1.) By Lemma 16 with $(\boldsymbol{\rho}, \widetilde{\boldsymbol{\rho}}) = (\boldsymbol{h}, \widetilde{\boldsymbol{h}})$, $(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}}) = (\boldsymbol{\psi}/\sqrt{p}, \widetilde{\boldsymbol{\psi}}/\sqrt{p})$, and $K = [n]$, we get

$$\overline{\mathbb{E}}\Big|\frac{n\boldsymbol{h}^\top \widetilde{\boldsymbol{h}} - \sum_{i\in n}(\boldsymbol{g}_i^\top \boldsymbol{h} - \sum_{j\in[p]} \frac{\partial h_j}{\partial g_{ij}})(\boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \sum_{j\in[p]} \frac{\partial \widetilde{h}_j}{\partial g_{ij}})}{\{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2\}^{1/2}\{\|\widetilde{\boldsymbol{h}}\|^2 + p^{-1}\|\widetilde{\boldsymbol{\psi}}\|^2\}^{1/2}}\Big| \le C_8(\sqrt{n}(1 + \sqrt{\overline{\mathbb{E}}[\Xi + \widetilde{\Xi}]}) + \overline{\mathbb{E}}[\Xi + \widetilde{\Xi}]).$$

where

$$\Xi = \frac{1}{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2} \sum_{i\in[n]} \sum_{j\in[p]} \Big(\big\|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\big\|^2 + \frac{1}{p}\big\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\big\|^2\Big) \tag{33}$$

Let us bound Ξ. By the derivative formula (29), we have

$$\sum_{ij} \|\partial_{ij}\boldsymbol{h}\|^2 \le 2\|\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{\psi}\|^2 + 2\|\boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}\|_F^2\|\boldsymbol{h}\|^2, \quad \sum_{ij} \|\partial_{ij}\boldsymbol{\psi}\|^2 \le 2\|\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{\psi}\|^2 + 2\|\boldsymbol{V}\|_{\mathrm{F}}^2\|\boldsymbol{h}\|^2$$

where $\boldsymbol{V} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}$ and $\boldsymbol{D} = \mathrm{diag}\{\mathsf{loss}''(z_i - \boldsymbol{g}_i^\top\boldsymbol{h})\}$. By $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ and $\|\boldsymbol{D}\|_{\mathrm{op}} \le \|\mathsf{loss}'\|_{\mathrm{Lip}}$, Ξ is bounded from above by

$$\Xi \le 2p\|\boldsymbol{A}\|_F^2 + 2\|\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}\|_F^2 + 2\|\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\|_F^2 + 2p^{-1}\|\boldsymbol{V}\|_{\mathrm{F}}^2 \le C(\delta, \|\mathsf{loss}'\|_{\mathrm{Lip}}) \cdot \|\boldsymbol{A}\|_{\mathrm{op}}^2\|\boldsymbol{G}\|_{\mathrm{op}}^4 \tag{34}$$

By $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ in (30) and Lemma 10, we obtain $\bar{\mathbb{E}}[\Xi] \le C_9$. This gives

$$\bar{\mathbb{E}}\Big|\frac{n\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \sum_{i\in n}(\boldsymbol{g}_i^\top\boldsymbol{h} - \sum_{j\in[p]}\frac{\partial h_j}{\partial g_{ij}})(\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \sum_{j\in[p]}\frac{\partial \widetilde{h}_j}{\partial g_{ij}})}{\{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2\}^{1/2}\{\|\widetilde{\boldsymbol{h}}\|^2 + p^{-1}\|\widetilde{\boldsymbol{\psi}}\|^2\}^{1/2}}\Big| \le C_{10}\sqrt{n}$$

Since the denominator is $\mathcal{O}_{\mathbb{P}}(1)$, we have

$$n\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \sum_{i\in[n]}\Big(\boldsymbol{g}_i^\top\boldsymbol{h} - \sum_{j\in[p]}\frac{\partial h_j}{\partial g_{ij}}\Big)\Big(\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \sum_{j\in[p]}\frac{\partial \widetilde{h}_j}{\partial g_{ij}}\Big) = \mathcal{O}_{\mathbb{P}}(\sqrt{n}).$$

By the derivative formula (29), noting that $\frac{\partial h_j}{\partial g_{ij}} = 0$ for all $i \notin I$ and $j \in [p]$, it holds that

$$\sum_{j\in[p]}\frac{\partial h_j}{\partial g_{ij}} = \begin{cases} 0 & i \notin I \\ \mathrm{tr}[\boldsymbol{A}]\psi_i - \boldsymbol{h}^\top\boldsymbol{A}\boldsymbol{G}\boldsymbol{D}\boldsymbol{e}_i & i \in I \end{cases} \tag{35}$$

Combined with $\|\boldsymbol{h}^\top\boldsymbol{A}\boldsymbol{G}\boldsymbol{D}\|^2 = \mathcal{O}_{\mathbb{P}}(n^{-1})$, the Cauchy–Schwarz inequality leads to

$$\sum_{i\in[n]}\Big(\boldsymbol{g}_i^\top\boldsymbol{h} - \sum_{j\in[p]}\frac{\partial h_j}{\partial g_{ij}}\Big)\Big(\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \sum_{j\in[p]}\frac{\partial \widetilde{h}_j}{\partial g_{ij}}\Big) = \sum_{i\in[n]}\Big(\boldsymbol{g}_i^\top\boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\,\mathbb{1}_{\{i\in I\}}\,\psi_i\Big)\Big(\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \mathrm{tr}[\widetilde{\boldsymbol{A}}]\,\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i\Big) + \mathcal{O}_{\mathbb{P}}(1).$$

This gives $\xi_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{n}) + \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(\sqrt{n})$ and completes the proof.

A.3.1 Proof of Lemma 16

Let us denote $\boldsymbol{f} = (\boldsymbol{\rho}^\top, \boldsymbol{\zeta}^\top)^\top \in \mathbb{R}^{Q+L}$ and $\widetilde{\boldsymbol{f}} = (\widetilde{\boldsymbol{\rho}}^\top, \widetilde{\boldsymbol{\zeta}}^\top)^\top\mathbb{R}^{Q\times L}$. Let us write the product as

$$\frac{\boldsymbol{\rho}^\top\widetilde{\boldsymbol{\rho}}}{\|\boldsymbol{f}\|\|\widetilde{\boldsymbol{f}}\|} = \Big\|\frac{1}{2}\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} + \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big\|^2 - \Big\|\frac{1}{2}\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} - \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big\|^2$$

Note that $\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|}$ and $\frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}$ are bounded by 1 in the standard Euclid norm $\|\cdot\|$. Applying the χ-square type moment inequality [Bel23, Theorem 7.2] to these terms respectively, we observe that the following two terms are bounded from above by $\sqrt{K}\{1 + \mathbb{E}[\Xi + \widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\Xi + \widetilde{\Xi}]$ up to some universal constant:

$$\mathbb{E}\Big|K\Big\|\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} + \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big\|^2 - \sum_{k\in[K]}\Big(\boldsymbol{z}_k^\top\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} + \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big) - \sum_{q\in[Q]}\frac{\partial}{\partial z_{kq}}\Big(\frac{\rho_q}{\|\boldsymbol{f}\|} + \frac{\widetilde{\rho}_q}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big)^2\Big|$$

$$\mathbb{E}\Big|K\Big\|\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} - \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big\|^2 - \sum_{k\in[K]}\Big(\boldsymbol{z}_k^\top\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} - \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big) - \sum_{q\in[Q]}\frac{\partial}{\partial z_{kq}}\Big(\frac{\rho_q}{\|\boldsymbol{f}\|} - \frac{\widetilde{\rho}_q}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big)^2\Big|$$

Thus, by the triangle inequality, we are left with the bound of cross terms:

$$\mathbb{E}\Big|K\frac{\boldsymbol{\rho}^\top\widetilde{\boldsymbol{\rho}}}{\|\boldsymbol{f}\|\|\widetilde{\boldsymbol{f}}\|} - \sum_{k\in[K]}b_k\widetilde{b}_k\Big| \lesssim \sqrt{K}\{1 + \mathbb{E}[\Xi + \widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\Xi + \widetilde{\Xi}] \tag{36}$$

where $b_k = \boldsymbol{z}_k^\top \boldsymbol{\rho}/\|\boldsymbol{f}\| - \sum_{q\in[Q]}(\partial/\partial z_{kq})(\rho_q/\|\boldsymbol{f}\|)$. Expanding the derivative of the second term, b_k can be written as

$$b_k = \underbrace{\frac{\boldsymbol{z}_k^\top \boldsymbol{\rho} - \sum_{q\in[Q]} \frac{\partial \rho_q}{\partial z_{kq}}}{\|\boldsymbol{f}\|}}_{=:a_k} - \sum_{q\in[Q]} \rho_q \frac{\partial}{\partial z_{kq}} \frac{1}{\|\boldsymbol{f}\|} = a_k + \sum_{q\in[Q]} \rho_q \frac{\boldsymbol{f}^\top}{\|\boldsymbol{f}\|^3} \frac{\partial \boldsymbol{f}}{\partial z_{kq}}$$

Thus, by multiple applications of Cauchy–Schwarz inequality, using $\|\boldsymbol{\rho}\|^2 \le \|\boldsymbol{f}\|^2$, we find that the error $\|\boldsymbol{b}-\boldsymbol{a}\|^2 = \sum_{k\in[K]}(b_k - a_k)^2$ is bounded from above by 2Ξ:

$$\|\boldsymbol{a}-\boldsymbol{b}\|^2 = \sum_{k\in K}\Big(\sum_{q\in[Q]} \rho_q \frac{\boldsymbol{f}^\top}{\|\boldsymbol{f}\|^3}\frac{\partial \boldsymbol{f}}{\partial z_{kq}}\Big)^2 \le \sum_k \|\boldsymbol{\rho}\|^2 \sum_q \Big(\frac{\boldsymbol{f}^\top}{\|\boldsymbol{f}\|^3}\frac{\partial \boldsymbol{f}}{\partial z_{kq}}\Big)^2 \le \frac{1}{\|\boldsymbol{f}\|^2}\sum_{k,q}\Big\|\frac{\partial \boldsymbol{f}}{\partial z_{kq}}\Big\|^2 \le 2\Xi.$$

The same argument gives $\|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|^2 \le 2\widetilde{\Xi}$.

Now we claim the following deterministic inequality for all $\boldsymbol{u}, \widetilde{\boldsymbol{u}}, \boldsymbol{a}, \widetilde{\boldsymbol{a}}, \boldsymbol{b}, \widetilde{\boldsymbol{b}} \in \mathbb{R}^K$ with $\|\boldsymbol{u}\| \vee \|\widetilde{\boldsymbol{u}}\| \le 1$:

$$\begin{aligned}\big||K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top\widetilde{\boldsymbol{a}}| - |K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{b}^\top\widetilde{\boldsymbol{b}}|\big| &\le (\|\boldsymbol{a}-\boldsymbol{b}\|^2 + \|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|^2) + \sqrt{K}(\|\boldsymbol{a}-\boldsymbol{b}\| + \|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|) \\ &\quad + 2^{-1}(|K\|\boldsymbol{u}\|^2 - \|\boldsymbol{b}\|^2| + |K\|\widetilde{\boldsymbol{u}}\|^2 - \|\widetilde{\boldsymbol{b}}\|^2|) \qquad (37)\end{aligned}$$

We prove this inequality later. Applying this inequality with $\boldsymbol{u} = \boldsymbol{\rho}/\|\boldsymbol{f}\|$ and $(\boldsymbol{a}, \boldsymbol{b})$ defined above, using $\|\boldsymbol{a}-\boldsymbol{b}\|^2 \le 2\Xi$, we get

$$\begin{aligned}|K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top\widetilde{\boldsymbol{a}}| &\le 2(\Xi + \widetilde{\Xi}) + \sqrt{2K}(\Xi^{1/2} + \widetilde{\Xi}^{1/2}) \\ &\quad + 2^{-1}(|K\|\boldsymbol{u}\|^2 - \|\boldsymbol{b}\|^2| + |K\|\widetilde{\boldsymbol{u}}\|^2 - \|\widetilde{\boldsymbol{b}}\|^2|) \\ &\quad + |K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{b}^\top\widetilde{\boldsymbol{b}}|\end{aligned}$$

Taking the expectation, using the moment bound (36), we are left with

$$\begin{aligned}\mathbb{E}|K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top\widetilde{\boldsymbol{a}}| &\lesssim \mathbb{E}[\Xi + \widetilde{\Xi}] + \sqrt{K}\mathbb{E}[(\Xi^{1/2} + \widetilde{\Xi}^{1/2})] \\ &\quad + \{1 + \mathbb{E}[\Xi]\}^{1/2} + \mathbb{E}[\Xi] + \{1 + \mathbb{E}[\widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\widetilde{\Xi}] \\ &\quad + \{1 + \mathbb{E}[\Xi + \widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\Xi + \widetilde{\Xi}]\end{aligned}$$

Using Jensen's inequality $\mathbb{E}[X^{1/2}] \le \sqrt{\mathbb{E}[X]}$ for any non-negative random variable X and $\sqrt{a}+\sqrt{b} \le \sqrt{2}\sqrt{a+b}$ for any non-negative scalars a, b, the RHS is bounded from above by $\sqrt{K}\{1 + \mathbb{E}[\Xi + \widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\Xi + \widetilde{\Xi}]$ up to some universal constant. This finishes the proof.

Below we prove the deterministic inequality (37). By multiple applications of the triangle inequality and Cauchy–Schwarz inequality,

$$\begin{aligned}\big||K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top\widetilde{\boldsymbol{a}}| - |K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{b}^\top\widetilde{\boldsymbol{b}}|\big| &\le |\boldsymbol{a}^\top\widetilde{\boldsymbol{a}} - \boldsymbol{b}^\top\widetilde{\boldsymbol{b}}| \\ &\le |(\boldsymbol{a}-\boldsymbol{b})^\top(\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}) + (\boldsymbol{a}-\boldsymbol{b})^\top\widetilde{\boldsymbol{b}} + \boldsymbol{b}^\top(\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}})| \\ &\le \|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\| + \|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{b}}\| + \|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|\|\boldsymbol{b}\| \\ &\le \frac{\|\boldsymbol{a}-\boldsymbol{b}\|^2 + \|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|^2}{2} + \|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{b}}\| + \|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|\|\boldsymbol{b}\|\end{aligned}$$

Using $\|\boldsymbol{b}\| \le \sqrt{|\|\boldsymbol{b}\|^2 - K\|\boldsymbol{u}\|^2|} + \sqrt{K\|\boldsymbol{u}\|^2}$ and $\|\boldsymbol{u}\| \le 1$, $\|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{b}}\|$ can be bounded from above as

$$\|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{b}}\| \le \|\boldsymbol{a}-\boldsymbol{b}\|\sqrt{|\|\widetilde{\boldsymbol{b}}\|^2 - K\|\widetilde{\boldsymbol{u}}\|^2|} + \sqrt{K}\|\boldsymbol{a}-\boldsymbol{b}\| \le \frac{\|\boldsymbol{a}-\boldsymbol{b}\|^2}{2} + \frac{|\|\widetilde{\boldsymbol{b}}\|^2 - K\|\widetilde{\boldsymbol{u}}\|^2|}{2} + \sqrt{K}\|\boldsymbol{a}-\boldsymbol{b}\|.$$

The same argument gives $\|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|\boldsymbol{b}\| \le \frac{\|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|^2}{2} + \frac{\|\boldsymbol{b}\|^2 - K\|\boldsymbol{u}\|^2|}{2} + \sqrt{K}\|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|$. Putting them all together, we obtained the desired upper bound of $\big||K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top\widetilde{\boldsymbol{a}}| - |K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{b}^\top\widetilde{\boldsymbol{b}}|\big|$.

A.4 Proof of trace convergence

We assume without loss of generality that $I = [n]$. Throughout this section, we denote $\boldsymbol{h} = \widehat{\boldsymbol{h}}_{\mu,K}$ and $\boldsymbol{\psi} = \boldsymbol{\psi}_{\mu,K}$ for simplicity.

Lemma 17. *For any $\mu \in (0,1]$ with $\mu^{-1} = o(n^{1/4})$,*

$$\begin{aligned}\operatorname{tr}[\boldsymbol{V}] &= \|\boldsymbol{h}\|^{-2}\big(\|\boldsymbol{\psi}\|^2 \operatorname{tr}[\boldsymbol{A}] - \boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}\big) + \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1}) \\ \operatorname{tr}[\boldsymbol{A}]^2 &= \frac{(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2)}{\|\boldsymbol{\psi}\|^4} + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-2})\end{aligned}$$

and $\mathbb{P}(\operatorname{tr}[\boldsymbol{A}]^2 \le C) \to 1$ for a constant $C > 0$ that only depend on (α, β, δ).

Proof. First we show the stochastic representation of $\operatorname{tr}[\boldsymbol{V}]$ by $(\boldsymbol{h}, \boldsymbol{\psi}, \operatorname{tr}[\boldsymbol{A}])$.

Lemma 18 ([BS22])**.** *Let $\boldsymbol{\rho} : \mathbb{R}^{K\times Q} \to \mathbb{R}^Q$ and $\boldsymbol{\zeta} : \mathbb{R}^{K\times Q} \to \mathbb{R}^K$ be two locally Lipschitz functions with differentiable components. If $\boldsymbol{Z} \in \mathbb{R}^{K\times Q}$ has i.i.d. $\mathcal{N}(0,1)$ entries, we have*

$$\mathbb{E}\Big[\Big(\frac{\boldsymbol{\zeta}^\top\boldsymbol{Z}\boldsymbol{\rho} - \sum_{kq}\frac{\partial(\zeta_k\rho_q)}{\partial g_{kq}}}{\|\boldsymbol{\zeta}\|^2 + \|\boldsymbol{\rho}\|^2}\Big)^2\Big] \le C_{11}(1 + \mathbb{E}[\Xi])$$

where $\Xi := \frac{1}{\|\boldsymbol{\rho}\|^2+\|\boldsymbol{\zeta}\|^2}\sum_{k\in[K]}\sum_{q\in[Q]}\Big(\big\|\frac{\partial\boldsymbol{\rho}}{\partial g_{kq}}\big\|^2 + \big\|\frac{\partial\boldsymbol{\zeta}}{\partial g_{kq}}\big\|^2\Big)$.

By Lemma 18 with $(\boldsymbol{\zeta}, \boldsymbol{\rho}) = (p^{-1/2}\boldsymbol{\psi}, \boldsymbol{h})$ and $\boldsymbol{Z} = \boldsymbol{G}$, we have

$$\mathbb{E}\Big[\Big(\frac{\frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} - \frac{1}{\sqrt{p}}\sum_{ij}\frac{\partial(\psi_i h_j)}{\partial g_{ij}}}{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2}\Big)^2\Big] \le C_{12}(1 + \mathbb{E}[\Xi])$$

where Ξ is the same estimate as in (33). Using the upper estimate $\Xi \le C_{13}\|\boldsymbol{A}\|_{\mathrm{op}}^2\|\boldsymbol{G}\|_{\mathrm{op}}^4$ in (34), $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ and $\mathbb{E}[\|\boldsymbol{G}\|^4] \le C(\delta)n^2$, we find that the RHS in the above display is bounded from away by $C(\delta)(1+\mu^{-2})$. Since the denominator $\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2$ is $\mathcal{O}_{\mathbb{P}}(1)$, we get

$$\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} - \sum_{ij}\frac{\partial\psi_i h_j}{\partial g_{ij}} = \sqrt{p}\cdot\mathcal{O}_{\mathbb{P}}(1)\cdot\mathcal{O}_{\mathbb{P}}(\mu^{-1}) = \mathcal{O}_{\mathbb{P}}(\sqrt{n}\mu^{-1}).$$

For the sum of derivative $\sum_{ij}\frac{\partial(\psi_i h_j)}{\partial g_{ij}}$, using $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ and $\|\boldsymbol{G}\|_{\mathrm{op}} = \mathcal{O}_{\mathbb{P}}(\sqrt{n})$, we have

$$\begin{aligned}\sum_{ij}\frac{\partial(\psi_i h_j)}{\partial g_{ij}} &= \|\boldsymbol{\psi}\|^2\operatorname{tr}[\boldsymbol{A}] - \boldsymbol{h}^\top\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{\psi} - \boldsymbol{\psi}^\top\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{h} - \|\boldsymbol{h}\|^2\operatorname{tr}[\boldsymbol{V}] \\ &= \|\boldsymbol{\psi}\|^2\operatorname{tr}[\boldsymbol{A}] + \mathcal{O}_{\mathbb{P}}(n^{1/2}) + \mathcal{O}_{\mathbb{P}}(\mu^{-1}) - \|\boldsymbol{h}\|^2\operatorname{tr}[\boldsymbol{V}]\end{aligned}$$

so that we are left with

$$\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} - \|\boldsymbol{\psi}\|^2\operatorname{tr}[\boldsymbol{A}] + \|\boldsymbol{h}\|^2\operatorname{tr}[\boldsymbol{V}] = \mathcal{O}_{\mathbb{P}}(\sqrt{n}\mu^{-1}).$$

Dividing by $\|\boldsymbol{h}\|^2$, noting $\|\boldsymbol{h}\|^{-1} = \mathcal{O}_{\mathbb{P}}(1)$ from $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2 > 0$, we obtain the representation of $\mathrm{tr}[\boldsymbol{V}]$.

Next, we show the stochastic representation of $\mathrm{tr}[\boldsymbol{A}^2]$ by $(\boldsymbol{G}, \boldsymbol{h}, \boldsymbol{\psi})$. By the stochastic representation of $\mathrm{tr}[\boldsymbol{V}]$,

$$\begin{aligned}\big|\|\boldsymbol{\psi}\|^4 \mathrm{tr}[\boldsymbol{A}]^2 - (\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} + \mathrm{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2)^2\big| &\le \mathcal{O}_{\mathbb{P}}(\sqrt{n}\mu^{-1})\big|\|\boldsymbol{\psi}\|^2 \mathrm{tr}[\boldsymbol{A}] + \boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} + \mathrm{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2\big| \\ &\le \mathcal{O}_{\mathbb{P}}(\sqrt{n}\mu^{-1})\mathcal{O}_{\mathbb{P}}(n\mu^{-1} + n + n) \\ &= \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2}).\end{aligned}$$

By Lemma 16 with $\boldsymbol{\rho} = \widetilde{\boldsymbol{\rho}} = \boldsymbol{\psi}/\sqrt{p}$ and $\boldsymbol{\zeta} = \widetilde{\boldsymbol{\zeta}} = \boldsymbol{h}$, we have

$$\mathbb{E}\Big[\frac{\big|p\|\frac{\boldsymbol{\psi}}{\sqrt{p}}\|^2 - \sum_{j\in[p]}(\frac{\boldsymbol{\psi}^\top}{\sqrt{p}}\boldsymbol{G}\boldsymbol{e}_j - \frac{1}{\sqrt{p}}\sum_{i\in[n]}\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}})^2\big|}{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2}\Big] \le C_{14}(\sqrt{n}(1 + \mathbb{E}[\Xi]^{1/2}) + \mathbb{E}[\Xi]).$$

where Ξ is the same estimate as in (33). Using $\mathbb{E}[\Xi] \le C_{15}\mu^{-2}$ again, we get

$$\|\boldsymbol{\psi}\|^2 - \frac{1}{p}\sum_{j\in[p]}(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j - \sum_{i\in[n]}\frac{\partial \psi_i}{\partial g_{ij}})^2 = \mathcal{O}_{\mathbb{P}}(1)\cdot\mathcal{O}_{\mathbb{P}}(\sqrt{n}(1+\mu^{-1}) + \mu^{-2}) = (n^{1/2}\mu^{-1})$$

Using $\sum_{i\in[n]}\frac{\partial \psi_i}{\partial g_{ij}} = -\boldsymbol{\psi}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j - \mathrm{tr}[\boldsymbol{V}]h_j$ by the derivative formula (29), it holds that

$$\begin{aligned}&\Big\|\sum_{j\in[p]}(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j - \sum_{i\in[n]}\frac{\partial \psi_i}{\partial g_{ij}})^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi} + \mathrm{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2\Big\| \\ &= \Big|\|\boldsymbol{G}^\top\boldsymbol{\psi} + \boldsymbol{A}^\top\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{\psi} + \mathrm{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi} + \mathrm{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2\Big| \\ &\le \|\boldsymbol{A}^\top\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{\psi}\|\Big(\|\boldsymbol{A}^\top\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{\psi}\| + 2\|\boldsymbol{G}^\top\boldsymbol{\psi} + \mathrm{tr}[\boldsymbol{V}]\boldsymbol{h}\|\Big) \\ &= \mathcal{O}_{\mathbb{P}}(\mu^{-1})(\mathcal{O}_{\mathbb{P}}(\mu^{-1}) + \mathcal{O}_{\mathbb{P}}(n)) = \mathcal{O}_{\mathbb{P}}(\mu^{-1}n).\end{aligned}$$

Combining the above displays, we are left with

$$\|\boldsymbol{\psi}\|^2 - p^{-1}\|\boldsymbol{G}^\top\boldsymbol{\psi} + \mathrm{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2 = \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1}) + \mathcal{O}_{\mathbb{P}}(\mu^{-1}n^{-1}) = \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1}).$$

Therefore,

$$\begin{aligned}\|\boldsymbol{\psi}\|^4 \mathrm{tr}[\boldsymbol{A}]^2 &= (\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} + \mathrm{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2)^2 + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2}) \\ &= (\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(2\,\mathrm{tr}[\boldsymbol{V}]\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} + \mathrm{tr}[\boldsymbol{V}]^2\|\boldsymbol{h}\|^2) + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2})) \\ &= (\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(\|\boldsymbol{G}^\top\boldsymbol{\psi} + \mathrm{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2) + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2})) \\ &= (\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2) + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-1}) + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2}).\end{aligned}$$

Multiplying by $\|\boldsymbol{\psi}\|^{-4}$, which is $\mathcal{O}_{\mathbb{P}}(n^2)$ since $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\mathrm{p}} \beta^2 > 0$, we obtain this representation of $\mathrm{tr}[\boldsymbol{A}]^2$. The upper bound $\mathbb{P}(\mathrm{tr}[\boldsymbol{A}]^2 \le C_{16})$ follows from the stochastic representation and the convergences: $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\mathrm{p}} \beta^2 > 0$, $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2 > 0$, $\|\boldsymbol{G}\|_{\mathrm{op}}/\sqrt{n} \xrightarrow{\mathrm{p}} 1 + \sqrt{\delta} > 0$. □

Lemma 19. *Letting $\overline{\mathbb{E}}[\cdot] = \mathbb{E}[\cdot|\boldsymbol{z}, \boldsymbol{\theta}]$ be the conditional expectation with respect to the design matrix $\boldsymbol{G}$, for any $\mu \in (0, 1]$ with $\mu^{-2} = o(n)$, we have*

$$\frac{(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2)}{\|\boldsymbol{\psi}\|^4} = \frac{(\overline{\mathbb{E}}[\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h}])^2 + \overline{\mathbb{E}}[\|\boldsymbol{h}\|]^2(p\overline{\mathbb{E}}[\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2])}{\overline{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2} + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1}).$$

Proof. By the Gaussian Poincaré inequality, we claim the following:

$$\begin{aligned}\bar{\mathbb{E}}[(\|\boldsymbol{\psi}\|^2 - \bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2])^2] &\le C_{17} n\mu^{-2},\\ \bar{\mathbb{E}}[(\|\boldsymbol{h}\|^2 - \bar{\mathbb{E}}[\|\boldsymbol{h}\|^2])^2] &\le C_{18} n^{-1}\mu^{-2},\\ \bar{\mathbb{E}}[(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} - \bar{\mathbb{E}}[\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h}])^2] &\le C_{19} n\mu^{-2},\\ \bar{\mathbb{E}}[(\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2 - \bar{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2])^2] &\le C_{20} n^3\mu^{-2}.\end{aligned}$$

First and second moment inequalities immediately follow from Lemma 13 with $\widetilde{I} = I = [n]$. For the third moment inequality, the Gaussian Poincaré inequality gives the upper bound $\mathbb{E}[(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} - \bar{\mathbb{E}}[\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h}])^2] \le \sum_{ij} \mathbb{E}\Big(\frac{\partial \boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h}}{\partial g_{ij}}\Big)^2$, where

$$\begin{aligned}\mathrm{RHS}_{ij} := \frac{\partial \boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h}}{\partial g_{ij}} &= \Big(-\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j\psi_i - \boldsymbol{V}\boldsymbol{e}_i h_j\Big)^\top \boldsymbol{G}\boldsymbol{h} + \psi_i h_j + \boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{A}\Big(\boldsymbol{e}_j\psi_i - \boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i h_j\Big)\\ &= \boldsymbol{e}_j^\top \boldsymbol{A}^\top \Big(-\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{h} + \boldsymbol{G}^\top\boldsymbol{\psi}\Big)\psi_i - \boldsymbol{e}_i^\top \Big(\boldsymbol{V}^\top \boldsymbol{G}\boldsymbol{h} + \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}^\top \boldsymbol{G}^\top\boldsymbol{\psi}\Big)h_j + \psi_i h_j\end{aligned}$$

By $\boldsymbol{V} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}$ and $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ in (30), we have $\sum_{i,j}\mathbb{E}[\mathrm{RHS}_{ij}^2] \le C_{21} n\mu^{-2}$. Next,

$$\mathbb{E}\Big[\Big(\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2 - \bar{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2]\Big)^2\Big] \le \sum_{i,j}\mathbb{E}\Big(\frac{\partial\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2}{\partial g_{ij}}\Big)^2 = 4\sum_{i,j}\mathbb{E}(\boldsymbol{\psi}^\top\boldsymbol{G}\frac{\partial \boldsymbol{G}^\top\boldsymbol{\psi}}{\partial g_{ij}})^2$$

where

$$\begin{aligned}\boldsymbol{\psi}^\top\boldsymbol{G}\frac{\partial \boldsymbol{G}^\top\boldsymbol{\psi}}{\partial g_{ij}} &= \boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{e}_j\psi_i + \boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{G}^\top(-\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j\psi_i - \boldsymbol{V}\boldsymbol{e}_i h_j)\\ &= \boldsymbol{\psi}^\top\boldsymbol{G}(\boldsymbol{I}_p - \boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{G}\boldsymbol{A})\boldsymbol{e}_j\psi_i - \boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{G}^\top\boldsymbol{V}\boldsymbol{e}_i h_j\end{aligned}$$

so $\sum_{ij}\mathbb{E}[\mathrm{RHS}_{ij}^2] \le C_{22} n^3\mu^{-2}$. Thus, we get

$$\begin{aligned}\|\boldsymbol{\psi}\|^2 &= \bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2] + \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1})\\ \|\boldsymbol{h}\|^2 &= \bar{\mathbb{E}}[\|\boldsymbol{h}\|^2] + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1})\\ \boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} &= \bar{\mathbb{E}}[\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}] + \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1})\\ \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2 &= \bar{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2] + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-1})\end{aligned}$$

Since $\mu^{-2} = o(n)$, the concentration of $\|\boldsymbol{\psi}\|^2$ on the conditional expectation $\mathbb{E}[\|\boldsymbol{\psi}\|^2]$ and the convergence $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\mathrm{p}} \beta^2$ yield $\mathbb{E}[\|\boldsymbol{\psi}\|^2]/p \xrightarrow{\mathrm{p}} \beta^2 > 0$. This implies that $1/\|\boldsymbol{\psi}\|^2$ and $1/\mathbb{E}[\|\boldsymbol{\psi}\|^2]$ are $\mathcal{O}_{\mathbb{P}}(n^{-1})$. Then, the error from replacing the denominator $\|\boldsymbol{\psi}\|^4$ by $\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2$ is estimated as

$$\begin{aligned}&(\|\boldsymbol{\psi}\|^{-4} - \bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2)\big((\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2)\big)\\ &= \frac{(\|\boldsymbol{\psi}\|^2 - \bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2])(\|\boldsymbol{\psi}\|^2 + \bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2])}{\|\boldsymbol{\psi}\|^4\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2}\cdot\mathcal{O}_{\mathbb{P}}(n^2)\\ &= \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1})\mathcal{O}_{\mathbb{P}}(n\cdot n^{-4})\mathcal{O}_{\mathbb{P}}(n^2) = \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1}).\end{aligned}$$

For the error from replacing the numerator with the conditional one, the error of replacing each term with the conditional one is given by

$$(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2 = (\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2 + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-1})$$

$$\|\boldsymbol{h}\|^2\|\boldsymbol{\psi}\|^2 = \bar{\mathbb{E}}[\|\boldsymbol{h}\|^2]\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2] + \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1})$$
$$\|\boldsymbol{h}\|^2\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2 = \bar{\mathbb{E}}[\|\boldsymbol{h}\|^2]\bar{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2] + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-1})$$

Thus, noting that $\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^{-1} = \mathcal{O}_{\mathbb{P}}(n^{-1})$, we get

$$\frac{(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2) - (\bar{\mathbb{E}}[\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}])^2 - \bar{\mathbb{E}}[\|\boldsymbol{h}\|^2](p\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2] - \bar{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2])}{\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2}$$
$$= \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1}).$$

This finishes the proof. □

Lemma 20. *Suppose $\mu \in (0,1]$ and $\mu^{-1} = O(n^{1/8})$. Then, there exist a non-negative random variables $\widehat{\kappa} \geq 0$, which is independent of $\boldsymbol{G}$, such that*

$$\mathbb{P}\Big(|\operatorname{tr}[\boldsymbol{A}] - \widehat{\kappa}| \leq 2n^{-1/16} \quad \textit{and} \quad |\widehat{\kappa}| \leq C\Big) \to 1$$

where C is a constant depending on (α, β, δ) only.

Proof. By Lemma 17 and Lemma 19, there exists a random variable A_n, which is independent from $\boldsymbol{G}$, such that

$$\operatorname{tr}[\boldsymbol{A}]^2 = A_n + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-2}) + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1}) = A_n + \mathcal{O}_{\mathbb{P}}(n^{-1/4}), \quad \text{and} \quad \mathbb{P}(|A_n| \leq C) \to 1$$

Noting $\operatorname{tr}[\boldsymbol{A}]^2 \geq 0$ and $\mathcal{O}_{\mathbb{P}}(n^{-1/4}) = o_{\mathbb{P}}(n^{-1/8})$, this implies that the event

$$\Omega := \{|\operatorname{tr}[\boldsymbol{A}]^2 - A_n| \leq n^{-1/8}\} \cap \{-n^{-1/8} \leq A_n \leq C\}$$

holds with high probability. Let us take $\widehat{\kappa} := \sqrt{(A_n + 2n^{-1/8})_+}$. Note in passing that under the event Ω, we have $\widehat{\kappa} = \sqrt{A_n + 2n^{-1/8}}$ and $\widehat{\kappa} \leq \sqrt{C+1}$. By the non-negativeness $\operatorname{tr}[\boldsymbol{A}] \geq 0$ and the $(1/2)$-Hölder continuity of the square root $\mathbb{R}_{\geq 0} \ni x \mapsto \sqrt{x}$, under the event Ω, it holds that

$$|\operatorname{tr}[\boldsymbol{A}] - \widehat{\kappa}| = |\sqrt{\operatorname{tr}[\boldsymbol{A}]^2} - \sqrt{A_n + 2n^{-1/8}}| \leq |\operatorname{tr}[\boldsymbol{A}]^2 - A_n - 2n^{-1/8}|^{1/2}$$

and the RHS is less than $|n^{-1/8} + 2n^{-1/8}|^{1/2} \leq 2n^{-1/16}$ by the triangle inequality. □

A.4.1 Proof of of Lemma 12

Applying Lemma 14 with $M = 1$ and $(\boldsymbol{z}, \boldsymbol{F}(\boldsymbol{z})) = (\boldsymbol{g}_i, \boldsymbol{h})$ for each $i \in [n]$, using $\sum_{ij} \bar{\mathbb{E}}[\|\partial_{ij}\boldsymbol{h}\|^2] = \mathcal{O}_{\mathbb{P}}(\mu^{-2}) = o_{\mathbb{P}}(n)$ we get

$$\sum_{i\in[n]} (\boldsymbol{g}_i^\top\boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\psi_i - \|\boldsymbol{h}\|\widehat{u}_i)^2 = o_{\mathbb{P}}(n), \quad \widehat{u}_i|\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{G}_{-i} \overset{\mathrm{d}}{=} \mathcal{N}(0,1)$$

By Lemma 20, there exists a non-negative random variable $\widehat{\kappa}$ independent from $\boldsymbol{G}$ such that

$$\operatorname{tr}[\boldsymbol{A}] = \widehat{\kappa} + o_{\mathbb{P}}(1), \quad \mathbb{P}(\widehat{\kappa} \leq C) \to 1$$

for a positive constant C. Then, combined with $\|\boldsymbol{h}\| = \alpha + o_{\mathbb{P}}(1)$, we get

$$\frac{1}{2n}\sum_{i\in[n]} (\boldsymbol{g}_i^\top\boldsymbol{h} - \widehat{\kappa}\psi_i - \alpha\widehat{u}_i)^2 = o_{\mathbb{P}}(1) + (\widehat{\kappa} - \operatorname{tr}[\boldsymbol{A}])^2\frac{\|\boldsymbol{\psi}\|^2}{n} + (\|\boldsymbol{h}\| - \alpha)^2\frac{\sum_{i\in[n]}\widehat{\boldsymbol{u}}_i^2}{n} = o_{\mathbb{P}}(1).$$

Furthermore, using the independence of $(\widehat{\kappa}, \boldsymbol{G})$ and the upper bound $\widehat{\kappa} \le C$, by the same argument in the proof of Theorem 2, we can easily show that the conditional expectation $\bar{\mathbb{E}}[\cdot] = \mathbb{E}[\cdot|\boldsymbol{z}, \boldsymbol{\theta}]$ of the square of LHS is bounded by C with high probability for some constant C. This implies that the conditional expectation $\bar{\mathbb{E}}$ of LHS is also $o(1)$, i.e.,

$$\frac{1}{n}\sum_{i\in[n]} \bar{\mathbb{E}}[(\boldsymbol{g}_i^\top \boldsymbol{h} - \widehat{\kappa}\psi_i - \alpha\widehat{u}_i)^2] = o_{\mathbb{P}}(1).$$

Let us define $\Xi_i \coloneqq \boldsymbol{g}_i^\top \boldsymbol{h} - \widehat{\kappa}\psi_i - \alpha\widehat{u}_i$ so that the above display reads $n^{-1}\sum_i \bar{\mathbb{E}}[\Xi_i^2] = o_{\mathbb{P}}(1)$. Noting $\psi_i = \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h})$ for all $i \in [n]$, the residual can be written as

$$z_i - \boldsymbol{g}_i^\top \boldsymbol{h} = \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i - \Xi_i; \widehat{\kappa})$$

for all $i \in [n]$. Since $\mathsf{prox}_f(\cdot)$ is 1-Lipschitz for any convex function, we have

$$\frac{1}{n}\sum_{i\in[n]} \bar{\mathbb{E}}\Big[\Big(z_i - \boldsymbol{g}_i^\top \boldsymbol{h} - \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \widehat{\kappa})\Big)^2\Big] \le \frac{1}{n}\sum_{i\in[n]} \bar{\mathbb{E}}[\Xi_i^2] = o_{\mathbb{P}}(1).$$

Since loss' is Lipschitz, the above display lets us approximate $\psi_i = \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h})$ by $\mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \widehat{\kappa})$:

$$\frac{1}{n}\sum_{i\in[n]} \bar{\mathbb{E}}\Big[\Big(\psi_i - \mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \widehat{\kappa})\Big)^2\Big] = o_{\mathbb{P}}(1).$$

Applying this approximation to the concentration $\beta^2 = \|\boldsymbol{\psi}\|^2/p + o_{\mathbb{P}}(1) = \bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]/p + o_{\mathbb{P}}(1)$ by Lemma 13 with $I = \widetilde{I} = [n]$, we get

$$\begin{aligned}\beta^2 &= \frac{n}{p}\frac{1}{n}\sum_{i\in[n]} \bar{\mathbb{E}}\Big[\psi_i^2\Big] + o_{\mathbb{P}}(1) = \delta\frac{1}{n}\sum_{i\in[n]} \bar{\mathbb{E}}\Big[\mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \widehat{\kappa})^2\Big] + o_{\mathbb{P}}(1) \\ &= \delta\frac{1}{n}\sum_{i\in[n]} \int_{-\infty}^{\infty} \varphi(x)\mathsf{env}'_{\mathsf{loss}}(z_i + \alpha x; \widehat{\kappa})^2 \,\mathrm{d}x + o_{\mathbb{P}}(1). \end{aligned} \tag{38}$$

where $\varphi(x)$ is the pdf of $\mathcal{N}(0,1)$. Let us define the functions $F, \widehat{F} : [0, +\infty) \to \mathbb{R}$ by

$$F(\tau) \coloneqq \beta^2 - \delta\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(Z + \alpha G; \tau)^2], \quad \widehat{F}(\tau) \coloneqq \beta^2 - \delta\frac{1}{n}\sum_{i\in[n]} \int_{-\infty}^{\infty} \varphi(x)\mathsf{env}'_{\mathsf{loss}}(z_i + \alpha x; \tau)^2 \,\mathrm{d}x$$

so that $F(\kappa) = 0$ by (5b) in System 1a (with $c = 1$), while (38) reads $\widehat{F}(\widehat{\kappa}) = o_{\mathbb{P}}(1)$. Note in passing that the weak law of large numbers implies $\widehat{F}(\tau) \xrightarrow{\mathrm{p}} F(\tau)$ pointwise, and F and $\widehat{F}$ are strictly increasing functions in τ since $-2^{-1}\mathsf{env}'_{\mathsf{loss}}(x;\tau)^2$ is the derivative of the convex function $\tau \mapsto \mathsf{env}_{\mathsf{loss}}(x;\tau)$, which is strictly convex under Assumption D-(4) (see [TAH18, Lemma 4.4]). Then, for any $\epsilon > 0$, we have

$$F(\kappa + \epsilon) > 0 = F(\kappa) > F(\kappa - \epsilon).$$

By the pointwise convergence $\widehat{F}(\tau) \xrightarrow{\mathrm{p}} F(\tau)$, it holds that

$$\widehat{F}(\kappa + \epsilon) > 2^{-1}F(\kappa + \epsilon) > 0 > 2^{-1}F(\kappa - \epsilon) > \widehat{F}(\kappa - \epsilon).$$

with high probability. Then, combined with $\widehat{F}(\widehat{\kappa}) = o_{\mathbb{P}}(1)$, we have

$$\widehat{F}(\kappa+\epsilon) > 2^{-1}F(\kappa+\epsilon) > \widehat{F}(\widehat{\kappa}) > 2^{-1}F(\kappa-\epsilon) > \widehat{F}(\kappa-\epsilon).$$

with high probability. Since $\widehat{F}$ is non-decreasing with probability 1, this gives $\mathbb{P}(|\widehat{\kappa}-\kappa| \le \epsilon) \to 1$. Since we took $\epsilon > 0$ arbitrarily, we obtain $\widehat{\kappa} \xrightarrow{\mathrm{p}} \kappa$.

Going back to the place where we replaced $\operatorname{tr}[\boldsymbol{A}]$ by $\widehat{\kappa}$, now replacing $\operatorname{tr}[\boldsymbol{A}]$ by κ instead, we obtain

$$\frac{1}{n}\sum_{i\in[n]}(\boldsymbol{g}_i^\top \boldsymbol{h} - \kappa\psi_i - \alpha\widehat{u}_i)^2 = o_{\mathbb{P}}(1),$$

and this approximation also holds in $\overline{\mathbb{E}}$. Thus, we get the approximation of residual and ψ_i

$$\frac{1}{n}\sum_{i\in[n]}\overline{\mathbb{E}}\Big[\Big(z_i - \boldsymbol{g}_i^\top\boldsymbol{h} - \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa)\Big)^2\Big] = o_{\mathbb{P}}(1), \quad \frac{1}{n}\sum_{i\in[n]}\overline{\mathbb{E}}\Big[\Big(\psi_i - \mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa)\Big)^2\Big] = o_{\mathbb{P}}(1)$$

Let us show $\operatorname{tr}[\boldsymbol{V}]/p \xrightarrow{\mathrm{p}} \nu$. Using the concentration $\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\|\boldsymbol{\psi}\|^2 + \operatorname{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2 = o_{\mathbb{P}}(n)$ from Lemma 17 and $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2 > 0$, $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\mathrm{p}} \beta^2$, $\operatorname{tr}[\boldsymbol{A}] \xrightarrow{\mathrm{p}} \kappa$, we have

$$p^{-1}\operatorname{tr}[\boldsymbol{V}] = \alpha^{-2}\kappa\beta^2 - \alpha^{-2}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}/p + o_{\mathbb{P}}(1).$$

Recall that we have shown the concentration $\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} = \overline{\mathbb{E}}[\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}] + o_{\mathbb{P}}(n)$ in the proof of Lemma 19. Applying the proximal approximation of the residual $z_i - \boldsymbol{g}_i^\top\boldsymbol{h}$ and ψ_i to this, noting $\overline{\mathbb{E}}[\|\boldsymbol{\psi}\|^2] = \mathcal{O}_{\mathbb{P}}(n)$ and $\overline{\mathbb{E}}[\|\boldsymbol{G}\boldsymbol{h}\|^2] = \mathcal{O}_{\mathbb{P}}(n)$, we are left with

$$\begin{aligned}
\frac{1}{n}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} &= \frac{1}{n}\sum_{i\in[n]}\overline{\mathbb{E}}[\psi_i\cdot\boldsymbol{g}_i^\top\boldsymbol{h}] + o_{\mathbb{P}}(1)\\
&= \frac{1}{n}\sum_{i\in[n]}\overline{\mathbb{E}}\Big[\mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa)(z_i - \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa))\Big] + o_{\mathbb{P}}(1)\\
&= \frac{1}{n}\sum_{i\in[n]}\int_{-\infty}^{\infty}\varphi(x)\mathsf{env}'_{\mathsf{loss}}(z_i + \alpha x;\kappa)(z_i - \mathsf{prox}_{\mathsf{loss}}(z_i + \alpha x;\kappa))\,\mathrm{d}x + o_{\mathbb{P}}(1)
\end{aligned}$$

so that the weak law of large numbers yields

$$\begin{aligned}
\frac{1}{n}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} &= \mathbb{E}[\mathsf{env}_{\mathsf{loss}'}(Z + \alpha G;\kappa)(Z - \mathsf{prox}_{\mathsf{loss}}(Z + \alpha G;\kappa))] + o_{\mathbb{P}}(1)\\
&= \mathbb{E}[\mathsf{env}_{\mathsf{loss}'}(Z + \alpha G;\kappa)(\alpha G + Z - \mathsf{prox}_{\mathsf{loss}}(Z + \alpha G;\kappa) - \alpha G)] + o_{\mathbb{P}}(1)\\
&= \kappa\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z;\kappa)^2] - \alpha\mathbb{E}[G\cdot\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z;\kappa)]\\
&= \kappa\cdot\beta^2/\delta - \alpha\cdot\nu\alpha/\delta.
\end{aligned}$$

where we have used (5b) and (5d) in System 1a (with $c = 1$) for the last equation. Combined with $p^{-1}\operatorname{tr}[\boldsymbol{V}] = \alpha^{-2}\kappa\beta^2 - \alpha^{-2}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}/p + o_{\mathbb{P}}(1)$, this gives $p^{-1}\operatorname{tr}[\boldsymbol{V}] = \nu + o_{\mathbb{P}}(1)$ and finishes the proof.

A.5 Convergence of error vector norm squared under Assumption D

In this section we verify that Assumption C and Assumption D-(1)-(3) are sufficient for [TAH18, Theorem 4.1] to hold. Comparing our assumptions and the condition assumed in the theorem, it suffices to show that the conditions (10) and (12) in [TAH18] can be omitted.

Indeed, the authors used the condition (10) to show that any $\widehat{\boldsymbol{u}} \in \partial \mathsf{loss}(\boldsymbol{z} - \boldsymbol{G}\widehat{\boldsymbol{h}}^B)/\sqrt{p}$ belongs to a compact set with high probability, where $\widehat{\boldsymbol{h}}^B$ is a "bounded" estimator. More precisely, $\widehat{\boldsymbol{h}}^B$ is a solution to the constrained optimization problem $\min_{\|\boldsymbol{h}\| \le K_\alpha} \mathsf{obj}(\boldsymbol{h})$ for a positive constant $K_\alpha > 0$ where $\mathsf{obj}(\boldsymbol{h}) = \sum_{i\in[n]} \mathsf{loss}(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \sum_{j\in[p]} \mathsf{reg}(\sqrt{p} h_j + \theta_j)$ is the objective function for the original unconstrained M-estimator. Since loss is differentiable with Lipschitz derivative loss' by Assumption C, using the same argument in Lemma 10, the norm of $\widehat{\boldsymbol{u}}$ is bounded as $\|\widehat{\boldsymbol{u}}\| \le (\|\mathsf{loss}'(\boldsymbol{z})\| + \|\mathsf{loss}'\|_{\mathrm{Lip}} \|\boldsymbol{G}\|_{\mathrm{op}} K_\alpha)/\sqrt{p}$. Since $\boldsymbol{G}$ has i.i.d. $\mathcal{N}(0,1)$ entries while $\mathsf{loss}'(z_i)$ has a finite second moment by Assumption D-(1), this gives $\|\widehat{\boldsymbol{u}}\|^2 \le C$ with probability approaching to 1 for a positive constant C.

Next, we claim that the condition (12) in [TAH18] is not necessary. Noting that the condition (12) is used for [TAH18, Assumption 2-(b)], it suffices to show that the assumption 2-(b) is satisfied given our assumptions. Here we restate the assumption 2-(b) for convenience:

$$\forall \tau > 0, \quad \lim_{c \to +\infty} c^2/(2\tau) - \mathbb{E}\big[\mathsf{env}_{\mathsf{reg}}(cH + \Theta) - \mathsf{reg}(\Theta)\big] = +\infty.$$

By the condition $\mathbb{P}(\Theta \neq 0) > 0$ in Assumption D-(3), either $\mathbb{P}(\Theta > 0) > 0$ or $\mathbb{P}(\Theta < 0) > 0$ holds. Let us consider the case $\mathbb{P}(\Theta > 0) > 0$. Define a measurable function $u(H, \Theta)$ of (H, Θ) as follows:

$$u(H, \Theta) := -\min(\Theta, 1) I\{\Theta > 0 \text{ and } H < 0\}.$$

Note that $u(H, \Theta)$ is always bounded as $|u(H, \Theta)| \le 1$. By the definition of Moreau envelope, i.e., $\mathsf{env}_{\mathsf{reg}}(cH + \Theta) := \mathrm{argmin}_{p \in \mathbb{R}} (cH + \Theta - p)^2/(2\tau) + \mathsf{reg}(p)$, taking the point $p = \Theta + u(H, \Theta)$, we get the lower estimate as follows:

$$\begin{aligned}
\frac{c^2}{2\tau} - \mathbb{E}\big[\mathsf{env}_{\mathsf{reg}}(cH + \Theta) - \mathsf{reg}(\Theta)\big] &\ge \frac{c^2}{2\tau} - \mathbb{E}\Big[\frac{(cH - u(H, \Theta))^2}{2\tau} + \mathsf{reg}(\Theta + u(H, \Theta)) - \mathsf{reg}(\Theta)\Big] \\
&= \frac{\mathbb{E}[uH]}{\tau} \cdot c - \frac{\mathbb{E}[u^2]}{2\tau} - \mathbb{E}\Big[\mathsf{reg}(\Theta + u) - \mathsf{reg}(\Theta)\Big].
\end{aligned}$$

Thus it suffices to show that $\mathbb{E}[uH] > 0$ and $\mathbb{E}[\mathsf{reg}(\Theta + u) - \mathsf{reg}(\Theta)]$ is finite for the RHS to diverge as $c \to +\infty$. By the definition of $u = u(\Theta, H)$, the expectation $\mathbb{E}[uH]$ can be written as

$$\mathbb{E}[uH] = \mathbb{E}[-H \min(\Theta, 1) I\{\Theta > 0, H < 0\}] = \mathbb{E}[|H| \min(\Theta, 1) I\{\Theta > 0, H < 0\}].$$

Here $|H| \min(\Theta, 1)$ is always strictly positive under the event $\{\Theta > 0, H > 0\}$, and this event has positive probability $\mathbb{P}(\Theta > 0, H > 0) = \mathbb{P}(\Theta > 0)\mathbb{P}(H > 0) = \mathbb{P}(\Theta > 0) \cdot 2^{-1}$ by the assumption $\mathbb{P}(\Theta > 0) > 0$ and the independence of Θ and $H \sim \mathcal{N}(0, 1)$. This means that $\mathbb{E}[uH]$ is strictly positive. Regarding $\mathbb{E}[\mathsf{reg}(\Theta + u) - \mathsf{reg}(\Theta)]$, we have

$$\mathbb{E}[\mathsf{reg}(\Theta + u) - \mathsf{reg}(\Theta)] = \mathbb{E}\Big[\big\{\mathsf{reg}(\Theta - \min(\Theta, 1)) - \mathsf{reg}(\Theta)\big\} I\{\Theta > 0, H < 0\}\Big].$$

Note that $0 < \Theta - \min(\Theta, 1) < \Theta$ for all $\Theta > 0$. Then, by the convexity of reg and the condition $\mathsf{reg}(0) = \min_x \mathsf{reg}(x)$ in Assumption C, under the event $I\{\Theta > 0, H < 0\}$ it holds that

$$0 > \mathsf{reg}(\Theta - \min(\Theta, 1)) - \mathsf{reg}(\Theta) > -d_\Theta \min(\Theta, 1) > -d_\Theta \quad \text{for all } d_\Theta \in \partial \mathsf{reg}(\Theta)$$

By Assumption D-(1), d_Θ has a finite second moment for any choice of sub-derivative d_Θ. Therefore, we have $0 > \mathbb{E}[\mathsf{reg}(\Theta + u) - \mathsf{reg}(\Theta)] > -\mathbb{E}[d_\Theta I\{\Theta > 0, H < 0\}] > -\infty$ so that $\mathbb{E}[\mathsf{reg}(\Theta + u) - \mathsf{reg}(\Theta)]$ is finite.

In the other case $\mathbb{P}(\Theta < 0) > 0$, we may take $u(H, \Theta) := \min(-\Theta, 1) I\{\Theta < 0, H > 0\}$. Then the same argument leads to $\mathbb{E}[uH] > 0$ and $|\mathbb{E}[\mathsf{reg}(\Theta + u) - \mathsf{reg}(\Theta)]| < +\infty$.

A.6 Convergence of loss gradient norm squared under Assumption D

Let $\boldsymbol{\psi} = \mathsf{loss}'(\boldsymbol{z} - \boldsymbol{G}\boldsymbol{h})$ and loss^* be the conjugate of loss. By the same argument in [TAH18], restricting the range of $\boldsymbol{\psi}$ to a compact set so that the strong duality holds, we observe that $\boldsymbol{\psi}$ is a solution to the following min-max problem with probability approaching to 1:

$$\max_{\boldsymbol{\psi}\in\mathbb{R}^n}\min_{\boldsymbol{h}\in\mathbb{R}^p} \boldsymbol{\psi}^\top(\boldsymbol{z}-\boldsymbol{G}\boldsymbol{h}) - \mathsf{loss}^*(\boldsymbol{\psi}) + \mathsf{reg}(\sqrt{p}\boldsymbol{h}+\boldsymbol{\theta}) = \max_{\boldsymbol{\psi}} \boldsymbol{\psi}^\top\boldsymbol{z} - \mathsf{loss}^*(\boldsymbol{\psi}) + \boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{\theta}/\sqrt{p} - \mathsf{reg}^*(\boldsymbol{G}^\top\boldsymbol{\psi}/\sqrt{p})$$

If we write $\widehat{\boldsymbol{u}} = \boldsymbol{\psi}/\sqrt{p} \in \mathbb{R}^n$ then $\widehat{\boldsymbol{u}}$ is the M-estimator of the form:

$$\widehat{\boldsymbol{u}} \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^n} \mathsf{F}(-\boldsymbol{G}^\top\boldsymbol{u}) + \mathsf{L}(\sqrt{n}\boldsymbol{u}), \quad \text{where} \quad \begin{aligned} &\mathsf{F}:\mathbb{R}^p\mapsto\mathbb{R}, \quad \boldsymbol{v}\mapsto \mathsf{reg}^*(-\boldsymbol{v}) + \boldsymbol{\theta}^\top\boldsymbol{v} \\ &\mathsf{L}:\mathbb{R}^n\mapsto\mathbb{R}, \quad \boldsymbol{u}\mapsto \mathsf{loss}^*(\sqrt{\frac{p}{n}}\boldsymbol{w}) - \sqrt{\frac{p}{n}}\boldsymbol{z}^\top\boldsymbol{w} \end{aligned}$$

For any $\boldsymbol{H}\sim\mathcal{N}(\boldsymbol{0}_p,\boldsymbol{I}_p)$, $\boldsymbol{G}\sim\mathcal{N}(\boldsymbol{0}_n,\boldsymbol{I}_n)$, for all $c\in\mathbb{R}$ and $\tau>0$ it holds that

$$\begin{aligned} \frac{1}{p}\big(\mathsf{env}_{\mathsf{F}}(c\boldsymbol{H};\tau) - \mathcal{F}(\boldsymbol{0})\big) &\xrightarrow{\mathrm{p}} \frac{c^2}{2\tau} - \mathbb{E}[\mathsf{env}_{\mathsf{reg}}(\frac{c}{\tau}H+\Theta_0;\frac{1}{\tau}) - \mathsf{reg}(0)] \\ \frac{1}{n}\big(\mathsf{env}_{\mathsf{L}}(c\boldsymbol{G};\tau) - \mathcal{L}(\boldsymbol{0})\big) &\xrightarrow{\mathrm{p}} \frac{c^2}{2\tau} - \mathbb{E}[\mathsf{env}_{\mathsf{loss}}(\sqrt{\delta}\frac{c}{\tau}G+Z;\frac{\delta}{\tau}) - \mathsf{loss}(0)] \end{aligned}$$

Thus, letting $F(c,\tau) := \mathbb{E}[\mathsf{env}_{\mathsf{reg}}(cH+\Theta;\tau)]$ and $L(c,\tau) := \mathbb{E}[\mathsf{env}_{\mathsf{loss}}(cG+\Theta;\tau)]$, [TAH18] implies that $\|\widehat{\boldsymbol{u}}\|^2 \xrightarrow{\mathrm{p}} \widetilde{\alpha}^2$ where $\widetilde{\alpha}$ is the minimizer of the min-max optimization problem

$$\inf_{\widetilde{\alpha},\widetilde{\tau}_g}\sup_{\widetilde{\beta},\widetilde{\tau}_h} \frac{\widetilde{\beta}\widetilde{\tau}_g}{2} + \delta^{-1}\Big(\frac{\widetilde{\alpha}^2\widetilde{\beta}}{2\widetilde{\tau}_g} - F(\frac{\widetilde{\alpha}\widetilde{\beta}}{\widetilde{\tau}_g},\frac{\widetilde{\beta}}{\widetilde{\tau}_g})\Big) - \frac{\widetilde{\alpha}\widetilde{\tau}_h}{2} - \frac{\widetilde{\alpha}\widetilde{\beta}^2}{2\widetilde{\tau}_h} + \Big(\frac{\widetilde{\alpha}\widetilde{\beta}^2}{2\widetilde{\tau}_h} - L(\sqrt{\delta}\widetilde{\beta},\delta\frac{\widetilde{\tau}_h}{\widetilde{\alpha}})\Big)$$

Thus, by the change of variables $(\widetilde{\alpha},\widetilde{\beta},\widetilde{\tau}_g,\widetilde{\tau}_h)\mapsto(\beta,\alpha/\sqrt{\delta},\tau_h/\sqrt{\delta},\tau_g/\delta)$ and multiplying the potential by $-\delta$, we are left with $\|\widehat{\boldsymbol{u}}\|^2 \xrightarrow{\mathrm{p}} \beta_*^2$ where β_* is the minimizer of

$$\sup_{\beta\tau_h}\inf_{\alpha,\tau_g} -\frac{\alpha\tau_h}{2} - \frac{\beta^2\alpha}{2\tau_g} + F(\frac{\beta\alpha}{\tau_h},\frac{\alpha}{\tau_h}) + \frac{\beta\tau_g}{2} + \delta L(\alpha,\frac{\tau_g}{\beta}),$$

which is the same potential in [TAH18].

A.7 Explicit expressions for degrees of freedom

Table 5: Explicit formulae for the degrees of freedom and residual degrees of freedom of the estimator $\widehat{\boldsymbol{\theta}}_I$ defined using (3). Here the loss functions $\mathsf{loss}(r)$ are: 1) square loss: $r^2/2$, 2) Huber loss: $r^2/2$ for $|r| \le 1$ and $|r| - 1/2$ for $|r| > 1$; and the regularization functions $\mathsf{reg}(b)$ are: 1) ridge penalty: $\frac{\lambda_1}{2} b^2$, 2) lasso penalty: $\lambda_2 |b|$, and 3) elastic net penalty: $\frac{\lambda_1}{2} b^2 + \lambda_2 |b|$. And other quantities are: 1) $\widehat{S}_I = \{j \in [p] \colon \widehat{\boldsymbol{\theta}}_I(j) \neq 0\}$ is the set of active variables of $\widehat{\boldsymbol{\theta}}_I$ and $\boldsymbol{X}_{\widehat{S}_I}$ is the submatrix of $\boldsymbol{X}_I$ made of columns indexed in $\widehat{S}_I$, 2) $\widehat{T}_I = \{i \in I \colon \mathsf{loss}''(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}_I) > 0\}$ is the set of detected inliers (active observations), and 3) the matrix $\boldsymbol{D}_I = \mathrm{diag}(\mathsf{loss}''(\boldsymbol{y}_I - \boldsymbol{X}_I \widehat{\boldsymbol{\theta}}_I))$.

Loss (loss)	**Regularizer** (reg)	**Degrees of freedom** (df_I)	**Residual degrees of freedom** ($\mathrm{tr}[\boldsymbol{V}_I]$)
Square	Ridge	$\mathrm{tr}\left[\left(\boldsymbol{X}_I^\top \boldsymbol{X}_I + \lambda_1 \boldsymbol{I}\right)^{-1} \boldsymbol{X}_I^\top \boldsymbol{X}_I\right]$	$\lvert I\rvert - \mathsf{df}_I$
Square	Lasso	$\lvert\widehat{S}_I\rvert$	$\lvert I\rvert - \mathsf{df}_I$
Square	Elastic net	$\mathrm{tr}\left[\left(\boldsymbol{X}_{\widehat{S}_I}^\top \boldsymbol{X}_{\widehat{S}_I} + \lambda_1 \boldsymbol{I}\right)^{-1} \boldsymbol{X}_{\widehat{S}_I}^\top \boldsymbol{X}_{\widehat{S}_I}\right]$	$\lvert I\rvert - \mathsf{df}_I$
Huber	Ridge	$\mathrm{tr}\left[\left(\boldsymbol{X}_I^\top \boldsymbol{D}_I \boldsymbol{X}_I + \lambda_1 \boldsymbol{I}\right)^{-1} \boldsymbol{X}_I^\top \boldsymbol{D}_I \boldsymbol{X}_I\right]$	$\lvert\widehat{T}_I\rvert - \mathsf{df}_I$
Huber	Lasso	$\lvert\widehat{S}_I\rvert$	$\lvert\widehat{T}_I\rvert - \mathsf{df}_I$
Huber	Elastic net	$\mathrm{tr}\left[\left(\boldsymbol{X}_{\widehat{S}_I}^\top \boldsymbol{D}_I \boldsymbol{X}_{\widehat{S}_I} + \lambda_1 \boldsymbol{I}\right)^{-1} \boldsymbol{X}_{\widehat{S}_I}^\top \boldsymbol{D}_I \boldsymbol{X}_{\widehat{S}_I}\right]$	$\lvert\widehat{T}_I\rvert - \mathsf{df}_I$
Convex	Convex	$\mathrm{tr}[(\partial/\partial \boldsymbol{y}_I) \boldsymbol{X}_I \widehat{\boldsymbol{\theta}}_I]$	$\mathrm{tr}[(\partial/\partial \boldsymbol{y}_I)\mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I \widehat{\boldsymbol{\theta}}_I)]$

B Proofs in Section 4

B.1 Proof of Proposition 7

Let us fix $\psi = (c\delta)^{-1}$ and take the derivative of $\mathcal{R}_M = M^{-1}\alpha^2 + (1 - M^{-1})\alpha^2 \eta_G$ with respect to $\phi = \delta^{-1}$. Note that $(\alpha, \beta, \kappa, \nu)$ are all fixed. Below, we derive the partial derivative of η_G with respect to ϕ. Using ψ and ϕ, we observe that η_G is the unique solution to the fixed point equation

$$\eta_G = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G; \phi).$$

Here, F_{reg} does not depend on ϕ but F_{loss} depends on ϕ as:

$$F_{\mathsf{loss}}(\eta_G; \phi) = \frac{\phi}{\psi^2 \beta^2} \cdot \mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa) \cdot \mathsf{env}'_{\widetilde{\mathsf{loss}}}(\alpha \widetilde{G} + Z; \kappa)],$$

Since the map $\eta_G \mapsto F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G; \phi)$ is differentiable and c-Lipschitz with $c < 1$ (see Theorem 1), the implicit function theorem implies that η_G is differentiable with respect to ϕ and the derivative satisfies:

$$\frac{\partial \eta_G}{\partial \phi} = (F_{\mathsf{reg}} \circ F_{\mathsf{loss}})'(\eta_G) \frac{\partial \eta_G}{\partial \phi} + F'_{\mathsf{reg}}(\eta_H) \frac{F_{\mathsf{loss}}(\eta_G)}{\phi}.$$

Rearranging the above display, we get

$$\frac{\partial \eta_G}{\partial \phi} = \frac{F'_{\mathsf{reg}}(\eta_H)}{\phi} \frac{F_{\mathsf{loss}}(\eta_G)}{1 - (F_{\mathsf{reg}} \circ F_{\mathsf{loss}})'(\eta_G)} = \frac{F'_{\mathsf{reg}}(\eta_H)}{\phi} \frac{\eta_H}{1 - (F_{\mathsf{reg}} \circ F_{\mathsf{loss}})'(\eta_G)} \ge 0.$$

where the last inequality follows from the fact that F_{reg} is non-decreasing and η_H is non-negative (see Theorem 1 for a homogeneous case). Combined with $\partial_\psi \mathcal{R}_M = (1 - M^{-1})\alpha^2 \frac{\partial \eta_G}{\partial \phi}$, we observe that $\mathcal{R}_M$ is non-decreasing in ϕ. Therefore, for any two $\phi_1 \le \phi_2$, we have that

$$\mathcal{R}_M(\phi_2, \psi) \ge \mathcal{R}_M(\phi_1, \psi) \ge \inf_{\psi \ge \phi_2} \mathcal{R}_M(\phi_1, \psi) \ge \inf_{\psi \ge \phi_1} \mathcal{R}_M(\phi_1, \psi) \quad \text{for all all } \psi \ge \phi_2.$$

This gives $\inf_{\psi \geq \phi_2} \mathcal{R}_M(\phi_2, \psi) \geq \inf_{\psi \geq \phi_1} \mathcal{R}_M(\phi_1, \psi)$ for any $\phi_1 \leq \phi_2$, which means the map $\phi \mapsto \inf_{\psi \geq \phi} \mathcal{R}_M(\phi, \psi)$ is non-decreasing.

B.2 Derivation of System 2

We will first reformulate Systems 1a and 1b in a slightly different set of parameters. Since the purpose of this reparameterization is to match with existing work, we will also consider regularizer reg with an explicit regularization level λ. The mapping with respect to the parameters in Systems 1a and 1b is as follows: $a = \frac{\lambda}{\beta}$, $\tau = \sqrt{c\delta}\frac{\beta}{\nu}$. Under this parameterization, note that $\frac{\beta}{\nu} = \frac{\tau}{\sqrt{c\delta}}$ and $\frac{\lambda}{\nu} = \frac{a\tau}{\sqrt{c\delta}}$.

Remark 9 (Scaling differences in design)**.** It is worth remarking that in the literature on risk characterization of regularized M-estimator under proportional asymptotics, the scaling of λ can be slightly different, up to a factor of $\sqrt{c\delta}$. One of the reasons for the differences is how the design matrix $\boldsymbol{X}$ is scaled. We assume that the entries of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ each have variance $1/p$ and thus each row has a unit average norm squared. It is also common to assume that the entries of $\boldsymbol{X}$ each have variance $1/n$ and thus each column has a unit average norm squared. This brings in a factor of $\sqrt{\delta}$. For the subsampled design $\boldsymbol{X}_I \in \mathbb{R}^{k \times p}$, we get an additional factor of $\sqrt{c}$. Consequently, some expressions may appear different up to this scaling.

Derivation of System 2. The derivation is straightforward. We will use some simple relationships between proximal operators and Moreau envelopes. Recall from (2) that $\mathsf{env}'_f(x;\tau) = \frac{1}{\tau}(x - \mathsf{prox}_f(x;\tau))$ so that the derivative identity $\mathsf{env}''_f(x;\tau) = \frac{1}{\tau}(1 - \mathsf{prox}'_f(x;\tau))$ holds for almost every x by the non-expansiveness of the proximal operator.

(1) Case of $m = \ell$. The original system of equations is given by System 1a. For regularizers of the form $\lambda\mathsf{reg}$, from (5a), we have

$$\begin{aligned}
\alpha^2 &= \mathbb{E}\big[\big(\tfrac{\lambda}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\beta}{\nu}H + \Theta; \tfrac{\lambda}{\nu}\big) - \tfrac{\beta}{\nu}H\big)^2\big] \\
&= \mathbb{E}\big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\beta}{\nu}H + \Theta; \tfrac{\lambda}{\nu}\big) - \Theta\big)^2\big] \\
&= \mathbb{E}\big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)^2\big]
\end{aligned} \tag{39}$$

where the variables τ and a are defined as:

$$\tau = \sqrt{c\delta}\frac{\beta}{\nu} \overset{(5b),(5d)}{=\!=} \frac{\alpha\sqrt{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z;\kappa)^2]}}{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z;\kappa) \cdot G]}, \tag{40}$$

$$a = \frac{\lambda}{\beta} \overset{(5b)}{=\!=} \frac{\lambda}{\sqrt{c\delta}\sqrt{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z;\kappa)^2]}}. \tag{41}$$

For squared loss $\mathsf{loss}(x) = x^2/2$, since $\mathsf{env}'_{\mathsf{loss}}(x;\tau) = x/(1+\tau)$ from Table 6, squaring the final expression in (40) yields

$$\tau^2 = \alpha^2 + \sigma^2. \tag{42}$$

Thus, τ^2 and α^2 are the limiting total and excess risks, respectively. Combining (39) and (42) gives the first desired equation (18a).

Similarly, (5c) after diving by τ yields

$$\tfrac{\kappa\beta}{\tau} = \tfrac{\beta}{\nu\tau} - \tfrac{\lambda}{\nu\tau}\mathbb{E}\big[\mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big) \cdot H\big].$$

Multiplying both sides by $\sqrt{c\delta}a\tau$ and noting that $a\beta = \lambda$, $\frac{\beta}{\nu} = \frac{\tau}{\sqrt{c\delta}}$, $\frac{\lambda}{\nu} = \frac{a\tau}{\sqrt{c\delta}}$ then gives

$$\sqrt{c\delta}\kappa\lambda = a\tau - \tfrac{(a\tau)^2}{\tau}\mathbb{E}\big[\mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big) \cdot H\big]. \tag{43}$$

From Stein's lemma, we have

$$\mathbb{E}[\mathsf{env}'_{\mathsf{reg}}(\tau H + \Theta; \kappa) \cdot H] = \tau\mathbb{E}[\mathsf{env}''_{\mathsf{reg}}(\tau H + \Theta; \kappa)], \tag{44}$$

and so (43) reduces to

$$\begin{aligned}\sqrt{c\delta}\kappa\lambda &\overset{(44)}{=\!=} a\tau - \tfrac{(a\tau)^2}{\sqrt{c\delta}}\mathbb{E}\big[\mathsf{env}''_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big] \\ &= a\tau - a\tau\mathbb{E}\big[1 - \mathsf{prox}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big] \\ &= a\tau\mathbb{E}\big[\mathsf{prox}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big].\end{aligned} \tag{45}$$

Using similar manipulations as above, (40) reduces to

$$\tau = \frac{\kappa\sqrt{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)^2]}}{\mathbb{E}[1 - \mathsf{prox}'_{\mathsf{loss}}(\alpha G + Z; \kappa)]}. \tag{46}$$

Combining (45) and (46), we get

$$0 = \kappa\lambda\Big(1 - \tfrac{a\tau}{\sqrt{c\delta}\kappa\lambda}\mathbb{E}\big[\mathsf{prox}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big]\Big). \tag{47}$$

From (41), under squared loss, we also have

$$1 + \kappa = \tfrac{\sqrt{c\delta}a\tau}{\lambda}. \tag{48}$$

Combining (47) with (48) then leads to the second desired equation (18b).

(2) Case of $m \neq \ell$. Step 2. (η_G, η_H) is the solution to the following 2-scalar fix-point equations:

$$\begin{aligned}\eta_G &= \frac{\mathbb{E}\big[\big(\tfrac{\lambda}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\beta}{\nu}H + \Theta; \tfrac{\lambda}{\nu}\big) - \tfrac{\beta}{\nu}H\big) \cdot \big(\tfrac{\lambda}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\beta}{\nu}\widetilde{H} + \Theta; \tfrac{\lambda}{\nu}\big) - \tfrac{\beta}{\nu}\widetilde{H}\big)\big]}{\mathbb{E}\big[\big(\tfrac{\lambda}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\beta}{\nu}H + \Theta; \tfrac{\lambda}{\nu}\big) - \tfrac{\beta}{\nu}H\big)^2\big]} \\ \eta_H &= c \cdot \frac{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa) \cdot \mathsf{env}'_{\mathsf{loss}}(\alpha\widetilde{G} + Z; \kappa)]}{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)^2]},\end{aligned}$$

where $\begin{pmatrix} G \\ \widetilde{G} \end{pmatrix} \sim N\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \eta_G \\ \eta_G & 1 \end{pmatrix}\Big)$ and $\begin{pmatrix} H \\ \widetilde{H} \end{pmatrix} \sim N\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \eta_H \\ \eta_H & 1 \end{pmatrix}\Big)$. Similarly, by variable substitution, we can rewrite the equations as:

$$\eta_G\alpha^2 = \mathbb{E}\big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)\big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}\widetilde{H} + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)\big] \tag{49a}$$

$$\eta_H = c \cdot \frac{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa) \cdot \mathsf{env}'_{\mathsf{loss}}(\alpha\widetilde{G} + Z; \kappa)]}{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)^2]}. \tag{49b}$$

Since $\mathsf{env}'_{\mathsf{loss}}(x; \tau) = x/(1 + \tau)$ for squared loss, we obtain the third desired equation (19a). □

B.3 Proximal operators and Moreau envelopes

Table 6: Proximal operators, Moreau envelopes, and their derivatives for ridge (first row) and lasso (second row) regularizers and Huber (third row) loss considered in Sections 4.2.1 and 4.2.2.

$f(x)$	$\mathsf{prox}_f(x;\tau)$	$\mathsf{prox}'_f(x;\tau)$	$\mathsf{env}_f(x;\tau)$	$\mathsf{env}'_f(x;\tau)$
$\frac{1}{2}x^2$	$\frac{x}{1+\tau}$	$\frac{1}{1+\tau}$	$\frac{1}{2}\frac{x^2}{1+\tau}$	$\frac{x}{1+\tau}$
$\lvert x\rvert$	$(\lvert x\rvert-\tau)_+\,\mathrm{sign}(x)$	$\mathbb{1}\{\lvert x\rvert\ge\tau\}$	$\begin{cases}\frac{1}{2\tau}x^2 & \lvert x\rvert<\tau\\ \lvert x\rvert-\frac{1}{2}\tau & \lvert x\rvert\ge\tau\end{cases}$	$\min\left\{\frac{\lvert x\rvert}{\tau},1\right\}\mathrm{sign}(x)$
$\begin{cases}\frac{x^2}{2} & \lvert x\rvert\le 1\\ \lvert x\rvert-\frac{1}{2} & \lvert x\rvert>1\end{cases}$	$\begin{cases}\frac{x}{1+\tau} & \lvert x\rvert\le 1+\tau\\ x-\tau\,\mathrm{sign}(x) & \lvert x\rvert>1+\tau\end{cases}$	$\begin{cases}\frac{1}{1+\tau} & \lvert x\rvert\le 1+\tau\\ 1 & \lvert x\rvert>1+\tau\end{cases}$	$\begin{cases}\frac{1}{2}\frac{x^2}{1+\tau} & \lvert x\rvert\le 1+\tau\\ \lvert x\rvert-\tau-\frac{1}{2} & \lvert x\rvert>1+\tau\end{cases}$	$\left(\frac{\lvert x\rvert}{1+\tau}-1\right)_+\mathrm{sign}(x)$

C Proofs in Section 5

C.1 Proof of Equation (26)

Applying Stein's lemma to (23b) and Cauchy–Schwarz inequality, we have

$$\begin{aligned}
1 &= \frac{1}{c\delta}\mathbb{E}\Big[\mathsf{prox}'_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}H;\tfrac{a\tau}{\sqrt{c\delta}}\big)\Big]\\
&= \frac{1}{\sqrt{c\delta}}\frac{1}{\tau}\mathbb{E}\Big[H\cdot\Big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}H;\tfrac{a\tau}{\sqrt{c\delta}}\big)-\Theta\Big)\Big]\\
&\le \frac{1}{\sqrt{c\delta}\tau}\mathbb{E}\Big[\Big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}H;\tfrac{a\tau}{\sqrt{c\delta}}\big)-\Theta\Big)^2\Big]^{1/2}\\
&= \frac{1}{\sqrt{c\delta}\tau}\sqrt{\tau^2-\sigma^2} \qquad \text{(by (23a)).}
\end{aligned}$$

Taking the square of both sides and rearranging the resulting τ^2 term, we get the lower estimate $\tau^2\ge\sigma^2/(1-c\delta)$ as desired.

C.2 Derivation of System 4 as a limit of System 2 as $\lambda\to 0^+$

Fix $\mathsf{reg}(x)=|x|^q$ for $q\in\{1,2\}$. Let $(a_\lambda,\tau_\lambda,\xi_\lambda)$ be the solution to System 2 for any $\lambda>0$ and let (a_*,τ_*,ξ_*) be the solution to System 4. Note that previous papers showed that (a_λ,τ_λ) converges to the solution (a_*,τ_*) as $\lambda\to(0)^+$; $q=1$ case is given by the proof of Lemma A.1 in [LW21] while the $q=2$ case immediately follows from the explicit formulae of (a_λ,τ_λ) and (a_*,τ_*). Below we claim the continuity of ξ, i.e., $\xi_\lambda\to\xi_*$. Denoting $\eta_\lambda=c\xi_\lambda^2/\tau_\lambda^2$ and $\eta_*=c\xi_*^2/\tau_*^2$, what we want to show is $\eta_\lambda\to\eta_*$.

Observe that the systems for ξ_λ and ξ_* read to $\eta_\lambda=F(\eta_\lambda;\tau_\lambda,a_\lambda)$ and $\eta_*=F(\eta_*;\tau_*,a_*)$ where

$$F(\eta;\tau,a):=\frac{c}{\tau^2}\Big\{\mathbb{E}\Big[\Big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}H;\tfrac{a\tau}{\sqrt{c\delta}}\big)-\Theta\Big)\cdot\Big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}\widetilde{H};\tfrac{a\tau}{\sqrt{c\delta}}\big)-\Theta\Big)\Big]+\sigma^2\Big\}$$

with $(H,\widetilde{H})$ being the mean zero jointly normals such that $\mathbb{E}[H^2]=\mathbb{E}[\widetilde{H}^2]=1$ and $\mathbb{E}[H\widetilde{H}]=\eta$. Note that the map $\eta\mapsto F(\eta;\tau_\lambda,a_\lambda)$ is c-Lipschitz over $[-1,1]$ by the same argument in Appendix A.1, while the map $(\tau,a)\mapsto F(\eta;\tau,a)$ is continuous over $(0,\infty)^2$ for any $\eta\in[-1,1]$ by the moment assumption in Assumption D-1 and the dominated convergence theorem. Then, $|\eta_\lambda-\eta_*|$ is bounded from above as

$$|\eta_\lambda-\eta_*|=|F(\eta_\lambda;\tau_\lambda,a_\lambda)-F(\eta_*,\tau_*,a_*)|$$

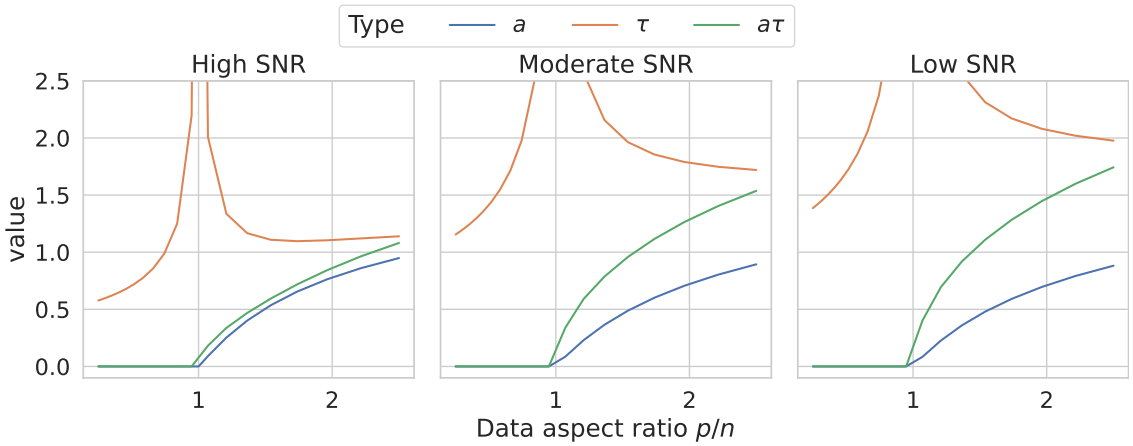

Figure 8: Fixed-point quantities for lassoless at different data aspect ratios p/n. The data model is given by (50) with signal strength $\rho = 0.5$ and sparsity levels $s = 0.2$ at different noise levels σ. *Left*: High SNR $\sigma = 0.5$. *Middle*: Modest SNR $\sigma = 1$. *Right*: Low SNR $\sigma = 1.2$.

$$\begin{aligned}
&\le |F(\eta_\lambda; \tau_\lambda, a_\lambda) - F(\eta_*, \tau_\lambda, a_\lambda)| + |F(\eta_*, \tau_\lambda, a_\lambda) - F(\eta_*, \tau_*, a_*)| \\
&\le c|\eta_\lambda - \eta_*| + |F(\eta_*, \tau_\lambda, a_\lambda) - F(\eta_*, \tau_*, a_*)|
\end{aligned}$$

so that $|\eta_\lambda - \eta_*| \le (1-c)^{-1}|F(\eta_*, \tau_\lambda, a_\lambda) - F(\eta_*, \tau_*, a_*)|$ holds. This upper bounds converges to 0 as $\lambda \to 0$ by the continuity of $F(\eta_*, \tau, a)$ in (τ, a) and the convergence $(\tau_\lambda, a_\lambda) \to (\tau_*, a_*)$.

C.3 Additional details for Remark 8

Expanding the product term in the fixed point equation (24a) by

$$\mathsf{prox}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta = \tfrac{\tau}{\sqrt{c\delta}}H - \tfrac{a\tau}{\sqrt{c\delta}}\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big),$$

we have

$$\begin{aligned}
\xi^2 - \sigma^2 = \frac{\tau^2}{c\delta}\mathbb{E}[H\widetilde{H}] - 2\frac{a\tau^2}{c\delta}\mathbb{E}\Big[\widetilde{H}\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] \\
+ \frac{(a\tau)^2}{c\delta}\mathbb{E}\Big[\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big) \cdot \mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}\widetilde{H}; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big],
\end{aligned}$$

where $\frac{\tau^2}{c\delta}\mathbb{E}[H\widetilde{H}] = \frac{\tau^2}{c\delta}\eta_H = \frac{\xi^2}{\delta}$ by the definition $\eta_H = c\xi^2/\tau^2$. For the second term, realizing $\widetilde{H} = \eta_H H + \sqrt{1-\eta_H^2}\overline{H}$ for a standard normal $\overline{H} \sim \mathcal{N}(0,1)$ independent of $(H, \widetilde{H}, \Theta)$, noting that $\mathsf{env}'_{\mathsf{reg}}(\Theta + \frac{\tau}{\sqrt{c\delta}}H)$ is bounded in second moment by Assumption D-(1), we have

$$\begin{aligned}
\mathbb{E}\Big[\widetilde{H} \cdot \mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] &= \eta_H \mathbb{E}\Big[H \cdot \mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] \\
&= \eta_H \tfrac{\tau}{\sqrt{c\delta}}\mathbb{E}\Big[\mathsf{env}''_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] && \text{(by Stein's lemma)} \\
&= \eta_H \tfrac{\tau}{\sqrt{c\delta}}\frac{\sqrt{c\delta}}{a\tau}\mathbb{E}\Big[1 - \mathsf{prox}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] \\
&= \frac{\eta_H}{a}(1 - c\delta) && \text{(by (23b))}.
\end{aligned}$$

Combined with $\eta_H = c\xi^2/\tau^2$, we have

$$\frac{a\tau^2}{c\delta}\mathbb{E}\Big[\widetilde{H}\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H\big)\Big] = \frac{\xi^2}{\delta}(1 - c\delta).$$

Therefore, we get

$$\xi^2\Big(1-\frac{1}{\delta}+2\frac{(1-c\delta)}{\delta}\Big)=\sigma^2+\frac{(a\tau)^2}{c\delta}\mathbb{E}\Big[\mathsf{env}'_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}H;\tfrac{a\tau}{\sqrt{c\delta}}\big)\cdot\mathsf{env}'_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}\widetilde{H};\tfrac{a\tau}{\sqrt{c\delta}}\big)\Big].$$

This means that $\xi^2\to\frac{\delta}{\delta-1}\sigma^2$ holds if and only if the rightmost term converges to 0 as $c\to(\delta^{-1})^-$. This is true if reg is Lipschitz and $\lim_{c\to(\delta^{-1})^-}a\tau=0$, as $|\mathsf{env}'_{\mathsf{reg}}(x;\tau)|\le\|\mathsf{reg}\|_{\mathrm{Lip}}$ for all $x\in\mathbb{R}$ and $\tau>0$. However, we are not able to provably establish that $a\tau\to0$ as $c\to\delta^{-1}$. For the lasso regularizer, we observe in Figure 8 that $a\tau\to0$ appears to hold as $c\to\delta^{-1}$.

D Additional numerical simulations

D.1 Minimum ℓ_1-norm interpolator

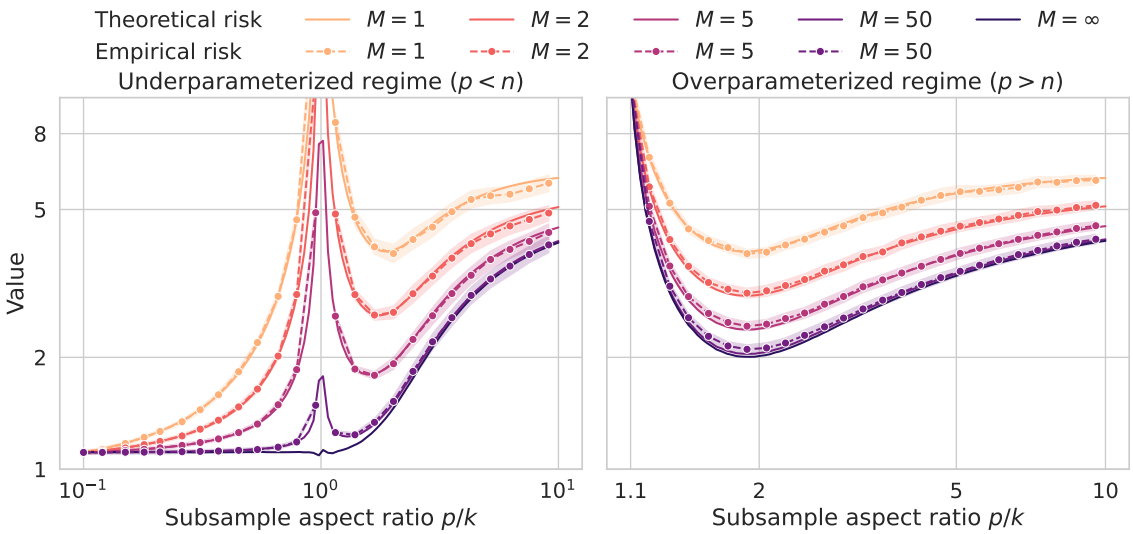

Figure 9: The prediction risk for lasso ensemble at different subsample aspect ratios p/k with regularization parameter $\lambda = 0.001$ and varying ensemble size M. The solid lines represent the theoretical risks, the dashed lines represent the empirical risks averaged over 50 simulations, and the shaded regions represent the standard errors. The data model is given by (50) with signal strength $\rho = 2$, noise level $\sigma = 1$, and support proportion $s = 0.1$. *Left*: underparameterized regime when $p/n = 0.1$ and $n = 2000$. *Right*: overparameterized regime when $p/n = 1.1$ and $n = 500$.

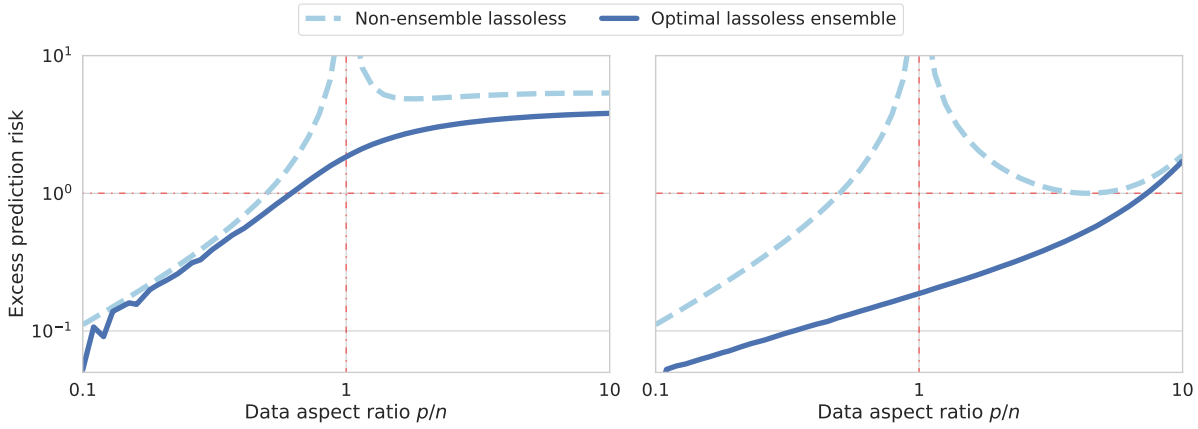

Figure 10: **Optimal subsample risk of the lassoless ensemble is monotonic in the data aspect ratio.** Excess risk of the lasso and optimal lasso ensemble, at different data aspect ratios p/n ranging from 0.1 to 10. The data model is given by (50) with signal strength $\rho = 2$, noise level $\sigma = 1$, data aspect ratio $p/n = 0.1$, feature size $p = 500$, and varying support proportion s. *Left*: dense regime with $s = 0.9$. *Right*: sparse regime with $s = 0.01$.

D.2 Optimal subagging versus optimal regularization

D.2.1 Square loss, lasso regularizer

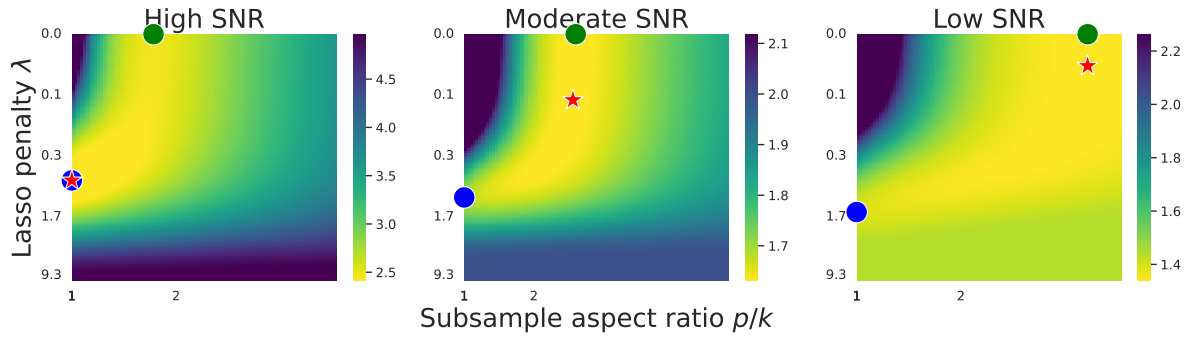

Figure 11: Heatmaps of theoretical prediction risk in λ and p/k of full lasso ensemble in the overparameterized regime ($p/n = 1.1$) and sparse regime ($s = 0.2$). The data model is given by (50) with support proportion $s = 0.2$ and noise level $\sigma = 1$ at different signal levels ρ. *Left*: high SNR $\rho = 2$ (optimal lasso is better than optimal subsample lassoless). *Middle*: moderate SNR $\rho = 1$ (optimal subsample lasso is better than optimal lasso and optimal subsample lassoless). *Right*: low SNR $\rho = 0.67$ (optimal subsample lassoless is better than optimal lasso).

D.2.2 Huber loss, unregularized

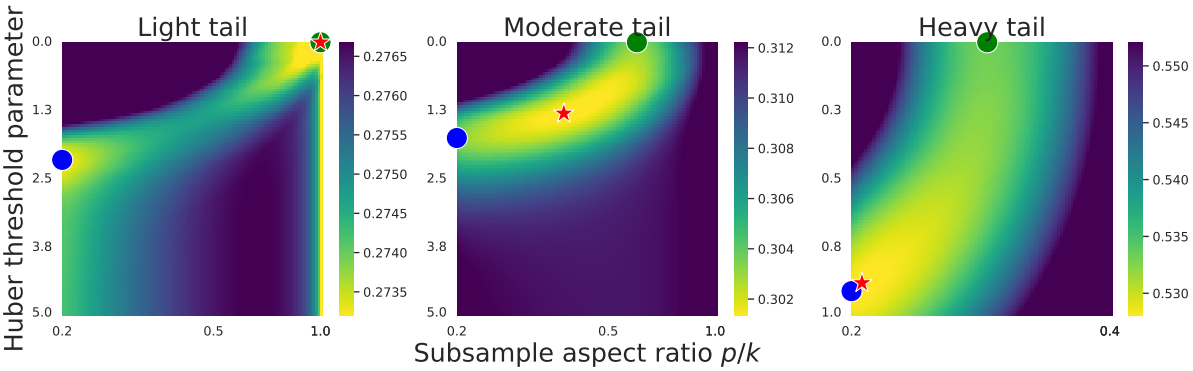

Figure 12: Heatmaps of theoretical prediction risk in Huber threshold parameter and subsample ratio k/n of full Huber ensemble in the underparameterized regime ($p/n = 0.2$). *Left*: noise follows Student's t distribution t_{20} (optimal subsample Huberless is better than optimal Huber). *Middle*: noise follows Student's t distribution t_{10} (optimal subsample Huber is better than both optimal subsample Huberless and optimal Huber). *Right*: noise follows Student's t distribution t_2 (optimal Huber is better than optimal subsample Huberless).

D.3 Huber regression with ℓ_1-regularization

D.3.1 Varying threshold parameter

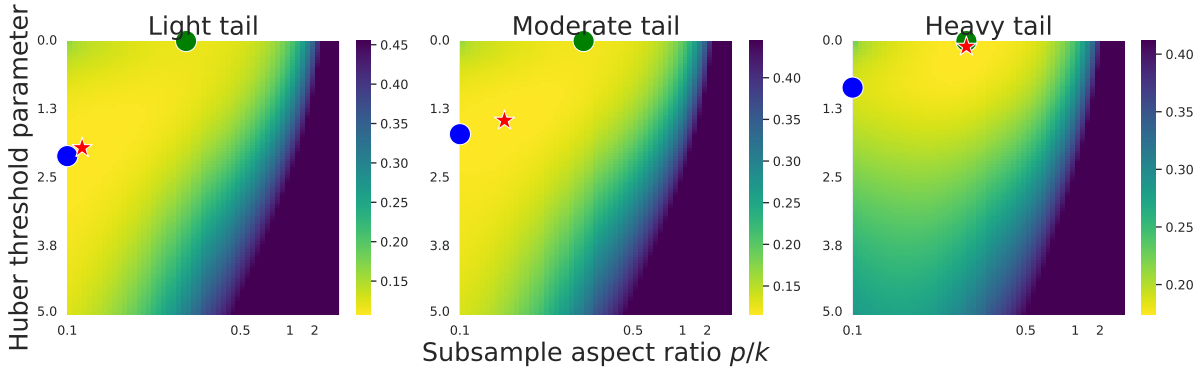

Figure 13: Heatmaps of theoretical prediction risk in Huber threshold parameter and subsample ratio k/n of full ℓ_1-regularized Huber ensemble with lasso regularization level 0.5 in the underparameterized regime ($p/n = 0.1$). *Left*: noise follows Student's t distribution t_{20}. *Middle*: noise follows Student's t distribution t_{10}. *Right*: noise follows Student's t distribution t_2.

D.3.2 Varying regularization level

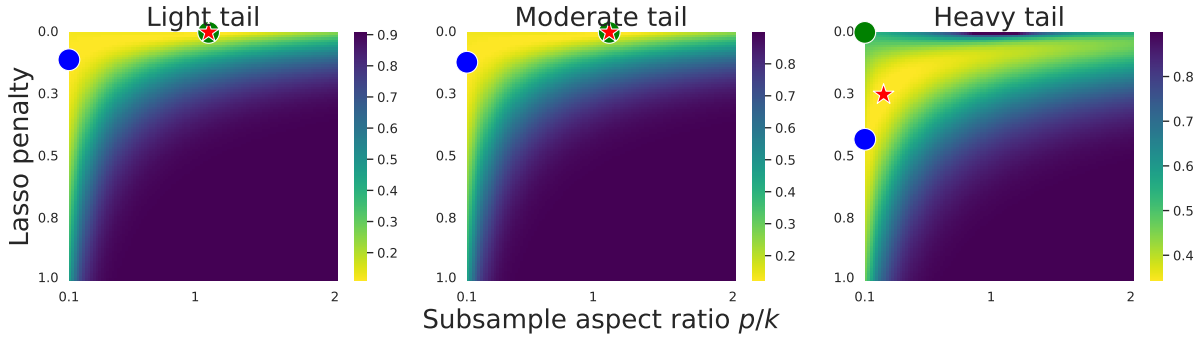

Figure 14: Heatmaps of theoretical prediction risk in lasso regularization level and subsample ratio k/n of full ℓ_1-regularized Huber ensemble with Huber threshold parameter 10 in the underparameterized regime ($p/n = 0.1$). *Left*: noise follows Student's t distribution t_{20}. *Middle*: noise follows Student's t distribution t_{10}. *Right*: noise follows Student's t distribution t_2.

D.4 Details of numerical experiments

D.4.1 Data model for lasso experiments

We consider a linear model whose signal variables are generated from two-point distribution:

$$y = \boldsymbol{x}^\top \boldsymbol{\theta} + z, \qquad \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, p^{-1}\boldsymbol{I}_p), \quad z \sim \mathcal{N}(0, \sigma^2), \quad \boldsymbol{\theta} \overset{i.i.d.}{\sim} s\mathcal{P}_{\rho/\sqrt{s}} + (1-s)\mathcal{P}_0, \tag{50}$$

where $\mathcal{P}_c$ denotes the Dirac measure at point $c \in \mathbb{R}$, and $\rho > 0$ is some given quantity that determines the signal magnitude. Here, $s \in (0,1)$ is the support proportion. The signal-to-noise-ratio (SNR) under the above model obeys

$$\text{SNR} = \frac{\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{\theta})^2]}{\sigma^2} = \frac{s \cdot (\rho^2/s)}{\sigma^2} = \frac{\rho^2}{\sigma^2}.$$

D.4.2 Data model for ℓ_1-regularized Huber experiments

We consider the ℓ_1-regularized Huber regression

$$\widehat{\boldsymbol{\theta}}_I \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i \in I} \mathsf{Huber}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta}; \rho) + \lambda \|\boldsymbol{\theta}\|_1$$

where $\lambda \geq 0$ is a regularization parameter and $\mathrm{Huber}(\cdot;\rho)$ is the Huber loss with threshold parameter $\rho > 0$, which is defined as follows [Hub64]:

$$\text{For all } x \in \mathbb{R}, \qquad \mathsf{Huber}(x;\rho) := \mathsf{env}_{|\cdot|}(x;\rho) = \begin{cases} \frac{1}{2\rho}x^2 & \text{if } |x| \leq \rho \\ |x| - \frac{1}{2}\rho & \text{if } |x| > \rho. \end{cases} \tag{51}$$

Observe that $\mathsf{Huber}(\cdot;\rho)$ behaves like the squared loss for large ρ and like the absolute loss for small ρ. More precisely, for all $x \in \mathbb{R}$ it holds that

$$\lim_{\rho \to 0^+} \mathsf{Huber}(x;\tau) = |x| \quad \text{and} \quad \lim_{\rho \to +\infty} \rho \cdot \mathsf{Huber}(x;\rho) = x^2/2.$$

We consider the linear model $y = \boldsymbol{x}^\top \boldsymbol{\theta} + z$ where the marginal distribution of the signal is set to the mixture $\epsilon\mathcal{N}(0,1) + (1-\epsilon)\mathcal{P}_0$ (recall that $\mathcal{P}_0$ denotes the Dirac measure at 0) with support proportion $\epsilon = 0.1$, while the noise z follows Student's t-distribution t_d for some degree of freedom $d \geq 2$, which will be specified for each numerical simulation.