THE DISTRIBUTION OF RIDGELESS LEAST SQUARES INTERPOLATORS

QIYANG HAN AND XIAOCONG XU

Abstract. The Ridgeless minimum ℓ_2-norm interpolator in overparametrized linear regression has attracted considerable attention in recent years. While it seems to defy the conventional wisdom that overfitting leads to poor prediction, recent research reveals that its norm minimizing property induces an 'implicit regularization' that helps prediction in spite of interpolation. This renders the Ridgeless interpolator a theoretically tractable proxy that offers useful insights into the mechanisms of modern machine learning methods.

This paper takes a different perspective that aims at understanding the precise stochastic behavior of the Ridgeless interpolator as a statistical estimator. Specifically, we characterize the distribution of the Ridgeless interpolator in high dimensions, in terms of a Ridge estimator in an associated Gaussian sequence model with positive regularization, which plays the role of the prescribed implicit regularization observed previously in the context of prediction risk. Our distributional characterizations hold for general random designs and extend uniformly to positively regularized Ridge estimators.

As a demonstration of the analytic power of these characterizations, we derive approximate formulae for a general class of weighted ℓ_q risks ($0 < q < \infty$) for Ridge(less) estimators that were previously available only for ℓ_2. Our theory also provides certain further conceptual reconciliation with the conventional wisdom: given any (regular) data covariance, for all but an exponentially small proportion of the signals, a certain amount of regularization in Ridge regression remains beneficial across various statistical tasks including (in-sample) prediction, estimation and inference, as long as the noise level is non-trivial. Surprisingly, optimal tuning can be achieved simultaneously for all the designated statistical tasks by a single generalized or k-fold cross-validation scheme, despite being designed specifically for tuning prediction risk.

The proof follows a two-step strategy that first proceeds under a Gaussian design using Gordon's comparison principles, and then lifts the Gaussianity via universality arguments. Our analysis relies on uniform localization and stability properties of the Gordon's optimization problem, along with uniform delocalization of the Ridge(less) estimators, both of which remain valid down to the interpolation regime.

Date: July 6, 2023.

2000 *Mathematics Subject Classification.* 60E15, 60G15.

Key words and phrases. comparison inequality, cross validation, minimum norm interpolator, random matrix theory, ridge regression, universality.

The research of Q. Han is partially supported by NSF grants DMS-1916221 and DMS-2143468. The research of X. Xu is partially supported by Hong Kong RGC grants GRF 16305421 and GRF 16303922.

CONTENTS

1. INTRODUCTION

1.1. **Overview.**

Consider the standard linear regression model

$$Y_i = X_i^\top \mu_0 + \xi_i, \quad 1 \le i \le m, \tag{1.1}$$

where we observe i.i.d. feature vectors $X_i \in \mathbb{R}^n$ and responses $Y_i \in \mathbb{R}$, and ξ_i's are unobservable errors. For notational simplicity, we write $X = [X_1 \cdots X_m]^\top \in \mathbb{R}^{m\times n}$ as the design matrix that collects all the feature vectors, and $Y = (Y_1, \ldots, Y_m)^\top \in \mathbb{R}^m$ as the response vector. The feature vectors X_i's are assumed to satisfy $\mathbb{E}\, X_1 = 0$ and $\mathrm{Cov}(X_1) = \Sigma$, and the errors satisfy $\mathbb{E}\, \xi_1 = 0$ and $\mathrm{Var}(\xi_1) = \sigma_\xi^2$.

Throughout this paper, we reserve m for the sample size, and n for the signal dimension. The aspect ratio ϕ, i.e., the number of samples per dimension, is then defined as $\phi \equiv m/n$. Accordingly, we refer to $\phi^{-1} > 1$ as the *overparametrized regime*, and $\phi^{-1} < 1$ as the *underparametrized regime*.

Within the linear model (1.1), the main object of interest is to recover/estimate the unknown signal $\mu_0 \in \mathbb{R}^n$. While a large class of regression techniques can be used for the purpose of signal recovery under various structural assumptions on μ_0, here we will focus our attention on one widely used class of regression estimators, namely, the *Ridge estimator* (cf. [HK70]) with regularization $\eta > 0$,

$$\widehat{\mu}_\eta = \mathop{\arg\min}_{\mu\in\mathbb{R}^n} \left\{\frac{1}{2n}\|Y - X\mu\|^2 + \frac{\eta}{2}\|\mu\|^2\right\} = \frac{1}{n}\Big(\frac{1}{n}X^\top X + \eta I_n\Big)^{-1} X^\top Y, \tag{1.2}$$

and the *Ridgeless estimator* (also known as the *minimum-norm interpolator*),

$$\widehat{\mu}_0 = \mathop{\arg\min}_{\mu\in\mathbb{R}^n} \{\|\mu\|^2 : Y = X\mu\} = (X^\top X)^- X^\top Y, \tag{1.3}$$

which is almost surely (a.s.) well-defined in the overparametrized regime $\phi^{-1} > 1$. Here A^- is the Moore-Penrose pseudo-inverse of A. The notation $\widehat{\mu}_0$ is justified since for $\phi^{-1} > 1, \widehat{\mu}_\eta \to \widehat{\mu}_0$ a.s. as $\eta \downarrow 0$.

From a conventional statistical point of view, the Ridgeless estimator seems far from an obviously good choice: As $\widehat{\mu}_0$ perfectly interpolates the data, it is susceptible to high variability due to the widely recognized bias-variance tradeoff inherent in 'optimal' statistical estimators [JWHT21, DSH23]. On the other hand, as the Ridgeless estimator $\widehat{\mu}_0$ is the limit point of the gradient descent algorithm run on the squared loss in the overparametrized regime $\phi^{-1} > 1$, it provides a simple yet informative test case for understanding one major enigma of modern machine learning methods: these methods typically interpolate training data perfectly; still, they enjoy good generalization properties [JGH18, DZPS18, AZLS19, BHMM19, COB19, ZBH+21].

Inspired by this connection, recent years have witnessed a surge of interest in understanding the behavior of the Ridgeless estimator $\widehat{\mu}_0$ and its closely related Ridge estimator $\widehat{\mu}_\eta$, with a particular focus on their prediction risks, cf. [TV+04, EK13, HKZ14, Dic16, DW18, EK18, ASS20, BHX20, WX20, BLLT20, BMR21, RMR21, HMRT22, TB22, CM22]. An emerging picture from these works is that, the norm minimizing property of the Ridgeless estimator $\widehat{\mu}_0$ inherited from the gradient descent algorithm, induces certain 'implicit regularization' that helps prediction accuracy despite interpolation. More interestingly, in certain scenarios of (Σ, μ_0), the Ridgeless estimator $\widehat{\mu}_0$ has a smaller (limiting) prediction risk compared to any Ridge estimator $\widehat{\mu}_\eta$ with $\eta > 0$, and therefore in such scenarios interpolation becomes optimal for the task of prediction, cf. [KLS20, HMRT22, TB22]. As a result, the Ridgeless estimator $\widehat{\mu}_0$ can be viewed, at least to some extent, as a (substantially) simplified yet theoretically tractable proxy that captures some important features of modern machine learning methods. Moreover, understanding for the prediction risk of $\widehat{\mu}_0$ serves as a basis for more complicated interpolation methods in e.g., kernel ridge regression [LR20, BMR21], random features model [MM22, MMM22], neural tangent model [MZ22], among others.

Despite these encouraging progress, there remains limited understanding of the behavior of the Ridgeless estimator $\widehat{\mu}_0$ as a statistical estimator. From a statistical perspective, a 'good' estimator usually possesses multiple desirable properties, and therefore it is natural to ask whether the Ridgeless estimator $\widehat{\mu}_0$ is also 'good' in other senses beyond the prediction accuracy. This is particularly relevant, if we aim to consider $\widehat{\mu}_0$ also as a 'good' estimator that can be applied in broader statistical contexts, rather than viewing it solely as a theoretical proxy designed to provide insights into the mechanisms of modern machine learning methods.

From a different angle, while much of the aforementioned research has focused on identifying favorable scenarios of (Σ, μ_0) in which prediction via $\widehat{\mu}_0$ can be accurate or even optimal, this line of theory does not automatically imply that in the practical statistical applications of overparametrized linear regression, the Ridgeless estimator $\widehat{\mu}_0$ remains the optimal choice for the task of prediction for 'typical' scenarios of (Σ, μ_0). The prevailing conventional statistical wisdom, which strongly advocates the use of regularization to strike a balance between bias and variance, may be still at work. From both conceptual and practical standpoints, it is therefore essential to understand the extent to which the optimality phenomenon of interpolation, as observed in certain scenarios of (Σ, μ_0) in the literature, can be considered 'generic' within the context of Ridge regression.

The main goal of this paper is to make a further step in understanding the precise stochastic behavior of the Ridge(less) estimator $\widehat{\mu}_\eta$, by developing a high-dimensional *distributional* characterization in the so-called proportional regime, where m and n is of the same order. As will be clear from below, the distributional characterization of the Ridge(less) estimator $\widehat{\mu}_\eta$ offers a detailed understanding of its statistical properties across various statistical tasks, extending beyond the prediction accuracy. Notable examples include a characterization for a general class of weighted ℓ_q risks ($0 < q < \infty$) associated with $\widehat{\mu}_\eta$, as well as its capacity for statistical inference in terms of constructing confidence intervals.

In addition, the distributional characterization also leads to new insights on the (sub-)optimality of the Ridgeless estimator $\widehat{\mu}_0$, which, interestingly, further reconciles with the conventional statistical wisdom: given any covariance structure Σ, except for an exponentially small proportion of signal μ_0's within the ℓ_2 norm ball, interpolation is optimal in prediction, estimation and inference, if and only if the noise level $\sigma_\xi^2 = 0$. In other words, in the common statistical setting where the noise level is nontrivial $\sigma_\xi^2 > 0$, a certain amount of regularization in Ridge regression remains beneficial for 'most' signal μ_0's across a number of different statistical tasks in the linear model (1.1)—including the most intensively studied prediction task in the literature.

Of course, in practice, the optimal regularization is unknown and may differ significantly for each statistical task. Surprisingly, our distributional characterization reveals that two widely used adaptive tuning methods, the *generalized cross-validation* scheme [CW79, GHW79] and the *k-fold cross-validation* scheme [GKKW02, JWHT21]—both of which designed for tuning the prediction risk—are actually simultaneously optimal for all the aforementioned statistical tasks, at least for 'most' signal μ_0's. In particular, for 'most' signal μ_0's, these two popular adaptive tuning methods automatically lead to optimal prediction, estimation and in-sample risks. Moreover, when combined with the debiased Ridge estimator, they produce the shortest confidence intervals for the coordinates of μ_0 with asymptotically valid coverage.

1.2. **Distribution of Ridge(less) estimators.** For simplicity of discussion, we will focus here on the overparametrized regime $\phi^{-1} > 1$ that is of our main interest.

1.2.1. *The Gaussian sequence model.* We will describe the distribution of the Ridge(less) estimator $\widehat{\mu}_\eta$ in the linear model (1.1), in terms of a corresponding Ridge estimator in a simpler *Gaussian sequence model*, defined for a given pair of (Σ, μ_0) and a noise level $\gamma > 0$ as

$$y^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma) \equiv \Sigma^{1/2}\mu_0 + \frac{\gamma g}{\sqrt{n}}, \quad g \sim \mathcal{N}(0, I_n). \tag{1.4}$$

The Ridge estimator $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)$ with regularization $\tau \geq 0$ in the Gaussian sequence model (1.4) is defined as

$$\begin{aligned}\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) &\equiv \arg\min_{\mu\in\mathbb{R}^n}\Big\{\frac{1}{2}\|\Sigma^{1/2}\mu - y^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma)\|^2 + \frac{\tau}{2}\|\mu\|^2\Big\}\\ &= (\Sigma+\tau I_n)^{-1}\Sigma^{1/2}\Big(\Sigma^{1/2}\mu_0 + \frac{\gamma g}{\sqrt{n}}\Big). \end{aligned}\tag{1.5}$$

1.2.2. *Distributional characterization of $\widehat{\mu}_\eta$ via $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}$.* Under standard assumptions on (i) the design matrix $X = \Sigma^{1/2}Z$, where Z consists of independent mean 0, unit-variance and light-tailed entries, and (ii) the error vector ξ with light-tailed components, we show that the distribution $\widehat{\mu}_\eta$ can be characterized via $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}$ as follows. For any $\eta \geq 0$, there exists a unique pair $(\gamma_{\eta,*},\tau_{\eta,*}) \in (0,\infty)^2$ determined via an implicit fixed point equation (cf. Eqn. (2.1) in Section 2 ahead), such that the distribution of $\widehat{\mu}_\eta$ is about the same as that of $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})$. Formally, for any 1-Lipschitz function $\mathsf{g}:\mathbb{R}^n\to\mathbb{R}$ and any $K>0$, we show in Theorems 2.3 and 2.4 that with high probability,

$$\sup_{\eta\in[0,K]}\Big|\mathsf{g}(\widehat{\mu}_\eta) - \mathbb{E}\,\mathsf{g}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}))\Big| \approx 0. \tag{1.6}$$

A particularly important technical aspect of (1.6) is that the distributional approximation (1.6) holds uniformly down to the interpolation regime $\eta = 0$ for $\phi^{-1} > 1$. This uniform guarantee will prove essential in the results to be detailed ahead.

From (1.6), the quantities $\gamma_{\eta,*}$ and $\tau_{\eta,*}$ can be naturally interpreted as the *effective noise* and *effective regularization* for the Ridge estimator $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}$ in the Gaussian sequence model (1.4). Moreover, $\tau_{0,*} > 0$ can be regarded as the aforementioned 'implicit regularization'. Interestingly, while this interpretation has been previously noted in the special context of prediction risk of the Ridge(less) estimator $\widehat{\mu}_\eta$ (cf. [BMR21, Remark 4.15], [CM22, Eqns. (16)-(19)]), our distributional characterization (1.6) reveals that such effective/implicit regularization is an inherently general phenomenon that persists at the level of distributional properties of $\widehat{\mu}_\eta$.

1.2.3. *Approximate formulae for general weighted ℓ_q risks.* As a first, yet nontrivial demonstration of the analytic power of (1.6), we show in Theorem 2.5 that for 'most' signals μ_0's, the weighted ℓ_q risk ($0<q<\infty$) of $\widehat{\mu}_\eta$, namely, $\|\mathsf{A}(\widehat{\mu}_\eta - \mu_0)\|_q$ with a well-behaved p.s.d. matrix A, concentrates around some deterministic quantity in the following sense: with high probability,

$$\sup_{\eta\in[0,K]}\bigg|\frac{\|\mathsf{A}(\widehat{\mu}_\eta-\mu_0)\|_q}{n^{-1/2}\|\mathrm{diag}(\Gamma^{\mathsf{A}}_{\eta;(\Sigma,\|\mu_0\|)})\|^{1/2}_{q/2}M_q} - 1\bigg| \approx 0. \tag{1.7}$$

Here $M_q \equiv \mathbb{E}^{1/q}|\mathcal{N}(0,1)|^q$, and $\mathrm{diag}(\Gamma^{\mathsf{A}}_{\eta;(\Sigma,\|\mu_0\|)}) \in \mathbb{R}^n$ is the vector that collects all diagonal elements of a p.s.d. matrix $\Gamma^{\mathsf{A}}_{\eta;(\Sigma,\|\mu_0\|)} \in \mathbb{R}^{n\times n}$, defined explicitly (in Eqn. (2.7)) via the weight matrix A, the data covariance Σ, the effective regularization $\tau_{\eta,*}$, and the signal energy $\|\mu_0\|$.

To the best of our knowledge, results of the form (1.7) are available in the literature (cited above) only for the special case $q = 2$, where the quantity $\|\mathsf{A}(\widehat{\mu}_\eta - \mu_0)\|_2$ admits a closed form expression in terms of the spectral statistics of the sample covariance that allows for direct applications of random matrix methods. Moving beyond this explicit closed form presents a notable analytic advantage of our theory (1.6) over existing random matrix approaches in analyzing ℓ_2 risks of $\widehat{\mu}_\eta$.

1.3. **Phase transitions for the optimality of interpolation.** While our theory (1.6) is strong enough to characterize all ℓ_q risks of the Ridge(less) estimator $\widehat{\mu}_\eta$ in the sense of (1.7), the uniform nature of (1.6) also illuminates novel insights into certain global, qualitative behavior of the most commonly studied ℓ_2 risks for finite samples. To fix notation, we define

- (*prediction risk*) $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) \equiv \|\Sigma^{1/2}(\widehat{\mu}_\eta - \mu_0)\|^2$,
- (*estimation risk*) $R^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) \equiv \|\widehat{\mu}_\eta - \mu_0\|^2$,
- (*in-sample risk*) $R^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta) \equiv n^{-1}\|X(\widehat{\mu}_\eta - \mu_0)\|^2$.

Using our uniform distributional characterization in (1.6), we show in Theorem 3.4 that for 'most' μ_0's and all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$, with high probability,

$$\Big|R^{\#}_{(\Sigma,\mu_0)}(0) - \min_{\eta\in[0,K]} R^{\#}_{(\Sigma,\mu_0)}(\eta)\Big| \begin{cases} \gtrsim 1, & \sigma_\xi^2 > 0, \\ \approx 0, & \sigma_\xi^2 = 0. \end{cases} \tag{1.8}$$

In fact, we prove in Theorem 3.4 a much stronger statement: for 'most' μ_0's, the (random) global optimum of $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}(\eta)$ for all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$ will be achieved approximately at $\eta_* = \mathsf{SNR}_{\mu_0}^{-1}$ with high probability. Here $\mathsf{SNR}_{\mu_0} = \|\mu_0\|^2/\sigma_\xi^2$ is the usual notion of signal-to-noise ratio; when $\mu_0 \neq 0$ and $\sigma_\xi^2 = 0$, we shall interpret $\mathsf{SNR}_{\mu_0}^{-1} = 0$.

It must be stressed that, for different $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$, the empirical risk curves $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}(\eta)$ concentrate on genuinely different deterministic counterparts $\eta \mapsto \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ with different mathematical expressions (cf. Theorem 3.1). As such, there are no apriori reasons to expect that these risk curves share approximately the same global minimum. Remarkably, as a consequence of the approximate formulae for the deterministic risk curves $\eta \mapsto \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ (cf. Theorem 3.2), we show that the curves $\eta \mapsto \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ are qualitatively similar, in that they approximately behave locally like a quadratic function centered around $\eta_* = \mathsf{SNR}_{\mu_0}^{-1}$ (cf. Proposition 3.3), at least for 'most' signal μ_0's.

From a broader perspective, the phase transition (1.8) aligns closely with the conventional statistical wisdom: for 'most' signal μ_0's, certain amount of regularization is necessary to achieve optimal performance for all the prediction, estimation and in-sample risks, as long as the noise level $\sigma_\xi^2 > 0$.

1.4. **Cross-validation: optimality beyond prediction.** The phase transition in (1.8) naturally raises the question of how one can choose the optimal regularization in a data-driven manner. Here we study two widely used adaptive tuning methods, namely,

(1) the generalized cross-validation scheme $\widehat{\eta}^{\mathsf{GCV}}$, and
(2) the k-fold cross-validation scheme $\widehat{\eta}^{\mathsf{CV}}$.

The readers are referred to (4.3) and (4.5) for precise definitions of $\widehat{\eta}^{\mathsf{GCV}}, \widehat{\eta}^{\mathsf{CV}}$ in the context of Ridge regression. Both methods have a long history in the literature; see, e.g., [Sto74, Sto77, CW79, GHW79, Li85, Li86, Li87, DvdL05] for some historical references. In essence, both methods $\widehat{\eta}^{\mathsf{GCV}}, \widehat{\eta}^{\mathsf{CV}}$ are designed to estimate the prediction risk, so it is natural to expect that they perform well for the task of prediction. Rigorous theoretical justifications along this line in Ridge regression in high dimensional settings can be found in, e.g., [LD19, PWRT21, HMRT22].

Interestingly, the insight from (1.8) suggests a far broader utility of these adaptive tuning methods. As all the empirical risk curves $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}(\eta)$ are approximately minimized at the same point $\eta_* = \mathsf{SNR}^{-1}_{\mu_0}$ with high probability, it is plausible to conjecture that the aforementioned cross-validation methods $\widehat{\eta}^{\mathsf{GCV}}, \widehat{\eta}^{\mathsf{CV}}$ could also yield optimal performance for estimation and in-sample risks, at least for 'most' signal μ_0's. We show in Theorems 4.2 and 4.3 that this is indeed the case: for 'most' signal μ_0's and all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$, with high probability,

$$R^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}), R^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{CV}}) \approx \min_{\eta\in[0,K]} R^{\#}_{(\Sigma,\mu_0)}(\eta). \tag{1.9}$$

Even more surprisingly, the optimality of $\widehat{\eta}^{\mathsf{GCV}}, \widehat{\eta}^{\mathsf{CV}}$ extends to the much more difficult task of statistical inference. In fact, we show in Theorem 4.4 that in the so-called debiased Ridge scheme, these two adaptive tuning methods $\widehat{\eta}^{\mathsf{GCV}}, \widehat{\eta}^{\mathsf{CV}}$ yield an asymptotically valid construction of confidence intervals for the coordinates of μ_0 with the shortest possible length.

To the best of our knowledge, theoretical optimality properties for the cross-validation schemes beyond the realm of prediction accuracy has not been observed in the literature, either for Ridge regression or for other regularized regression estimators. Our findings here in the context of Ridge regression can therefore be viewed as a first step in understanding the potential broader merits of cross validation schemes for a larger array of statistical inference problems.

1.5. **Proof techniques.** As previously mentioned, a significant body of recent research (cited above) has concentrated on various facets of the precise asymptotics of the prediction risk for the Ridgeless estimator $\widehat{\mu}_0$. These studies extensively utilize the explicit form of (1.3), and rely almost exclusively on techniques from random matrix theory (RMT) to relate the behavior of the bias and variance terms in the prediction risk and the spectral properties of the sample covariance.

Here, as (1.6) lacks a direct connection to the spectrum of the sample covariance, we adopt a different, two-step strategy for its proof:

(1) In the first step, we establish (1.6) under a Gaussian design X via the so-called convex Gaussian min-max theorem (CGMT) approach [Gor85, Gor88].
(2) In the second step, we prove universality that lifts the Gaussianity X via leveraging the recently developed comparison inequalities in [HS22].

Under a Gaussian design, the proof method for establishing distributional properties of regularized regression estimators via the CGMT has been executed for the closely related Lasso estimator with strict non-vanishing regularization, cf. [MM21, CMW22][1]. Here in our setting, the major technical hurdle is to handle the vanishing strong convexity in the optimization problem (1.2) as $\eta \downarrow 0$. We overcome this technical issue by establishing uniform localization of Gordon's min-max optimization, valid down to $\eta = 0$. The localization property, in a certain sense, allows us to conclude that the distributional properties of $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})$ are stable as $\eta \downarrow 0$. The readers are referred to Section 6.3 for a proof outline.

For the universality problem, the key step of our arguments is to reduce, in a proper sense, the difficult problem on the universality of $\widehat{\mu}_0$ to an easier problem on the universality of $\widehat{\mu}_\eta$ for small $\eta > 0$, so that the comparison inequalities in [HS22] can be applied. This reduction is achieved by proving that (i) $\widehat{\mu}_\eta$ and related quantities are uniformly delocalized down to the interpolation regime $\eta = 0$, and (ii) both the primal and Gordon optimization problems are 'stable' in suitable senses as $\eta \downarrow 0$. A more detailed proof outline is contained in Section 6.4.

1.6. **Organization.** The rest of the paper is organized as follows. In Section 2, we present our main results on the distributional characterizations (1.6) of the Ridge(less) estimator $\widehat{\mu}_\eta$, and the approximate ℓ_q risk formulae (1.7). In Section 3, we provide a number of approximate ℓ_2 risk formulae via RMT, and establish the phase transition (1.8). In Section 4, we give a formal validation for the two cross validation schemes mentioned above, both in terms of (1.9) and statistical inference via the debiased Ridge estimator. A set of illustrative simulation results are presented in Section 5 to collaborate (some of) the theoretical results. Due to the high technicalities involved in the proof of (1.6), a proof outline will be given in Section 6. All the proof details are then presented in Sections 7-12.

1.7. **Notation.** For any positive integer n, let $[n] = [1 : n]$ denote the set $\{1, \ldots, n\}$. For $a, b \in \mathbb{R}$, $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. For $a \in \mathbb{R}$, let $a_\pm \equiv (\pm a) \vee 0$. For $x \in \mathbb{R}^n$, let $\|x\|_p$ denote its p-norm ($0 \le p \le \infty$), and $B_{n;p}(R) \equiv \{x \in \mathbb{R}^n : \|x\|_p \le R\}$. We simply write $\|x\| \equiv \|x\|_2$ and $B_n(R) \equiv B_{n;2}(R)$. For a matrix $M \in \mathbb{R}^{m\times n}$, let $\|M\|_{\mathrm{op}}, \|M\|_F$ denote the spectral and Frobenius norm of M, respectively. I_n is reserved for an $n \times n$ identity matrix, written simply as I (in the proofs) if no confusion arises. For a square matrix $M \in \mathbb{R}^{n\times n}$, we let $\mathrm{diag}(M) \equiv (M_{ii})_{i=1}^n \in \mathbb{R}^n$.

We use C_x to denote a generic constant that depends only on x, whose numeric value may change from line to line unless otherwise specified. $a \lesssim_x b$ and $a \gtrsim_x b$ mean $a \le C_x b$ and $a \ge C_x b$, abbreviated as $a = O_x(b), a = \Omega_x(b)$ respectively; $a \asymp_x b$ means $a \lesssim_x b$ and $a \gtrsim_x b$, abbreviated as $a = \Theta_x(b)$. $\mathcal{O}$ and $\mathfrak{o}$ (resp. $\mathcal{O}_{\mathbf{P}}$ and $\mathfrak{o}_{\mathbf{P}}$) denote the usual big and small O notation (resp. in probability). For a random variable X, we use $\mathbb{P}_X, \mathbb{E}_X$ (resp. $\mathbb{P}^X, \mathbb{E}^X$) to indicate that the probability and expectation are taken with respect to X (resp. conditional on X).

For a measurable map $f : \mathbb{R}^n \to \mathbb{R}$, let $\|f\|_{\mathrm{Lip}} \equiv \sup_{x\neq y}|f(x) - f(y)|/\|x - y\|$. f is called *L-Lipschitz* iff $\|f\|_{\mathrm{Lip}} \le L$. For a proper, closed convex function f defined on $\mathbb{R}^n$, its *Moreau envelope* $\mathsf{e}_f(\cdot;\tau)$ and *proximal operator* $\mathsf{prox}_f(\cdot;\tau)$ for

[1]More literature on other statistical applications of the CGMT can be found in Section 6.

any $\tau > 0$ are defined by $\mathsf{e}_f(x;\tau) \equiv \min_{z\in\mathbb{R}^n} \{\frac{1}{2\tau}\|x-z\|^2 + f(z)\}$ and $\mathsf{prox}_f(x;\tau) \equiv \arg\min_{z\in\mathbb{R}^n} \{\frac{1}{2\tau}\|x-z\|^2 + f(z)\}$.

Throughout this paper, for an invertible covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$, we write $\mathcal{H}_\Sigma \equiv \mathrm{tr}(\Sigma^{-1})/n$ as the harmonic mean of the eigenvalues of Σ.

2. Distribution of Ridge(less) estimators

2.1. **Some definitions.**

For $K > 1$, let

$$\Xi_K \equiv [\mathbf{1}_{\phi^{-1}<1+1/K}K^{-1}, K].$$

This notation will be used throughout the paper for uniform-in-η statements. In particular, in the overparametrized regime $\phi^{-1} \geq 1 + 1/K$, we have $\Xi_K = [0, K]$.

Next, for $\gamma, \tau \geq 0$, recall $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)$ in (1.5), and we define its associated estimation error $\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau)$ and the degrees-of-freedom $\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau)$ as

$$\begin{cases} \mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau) \equiv \|\Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) - \mu_0)\|^2, \\ \mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau) \equiv \big\langle \frac{\gamma g}{\sqrt{n}}, \Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) - \mu_0)\big\rangle. \end{cases}$$

We note that the $\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau)$ defined above is naturally related to the usual notion of degrees-of-freedom (cf. [Ste81, Efr04]) for $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)$, in the sense that $\mathrm{df}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)) \equiv \sum_{j=1}^n \frac{1}{\gamma^2/n} \mathrm{Cov}\,\big((\Sigma^{1/2}\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)})_j, y^{\mathsf{seq}}_{(\Sigma,\mu_0),j}\big) = \frac{n}{\gamma^2}\,\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau)$.

2.2. **Working assumptions.**

Assumption A. $X = Z\Sigma^{1/2}$, where (i) $Z \in \mathbb{R}^{m\times n}$ has independent, mean-zero, unit variance, uniformly sub-gaussian entries, and (ii) $\Sigma \in \mathbb{R}^{n\times n}$ is an invertible covariance matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n > 0$.

Here 'uniform sub-gaussianity' means $\sup_{i\in[m],j\in[n]}\|Z_{ij}\|_{\psi_2} \leq C$ for some universal $C > 0$, where ψ_2 is the Orlicz 2-norm (cf. [vdVW96, Section 2.2, pp. 95]).

We shall often write the Gaussian design as $Z = G$, where $G \in \mathbb{R}^{m\times n}$ consists of i.i.d. $\mathcal{N}(0,1)$ entries.

Assumption B. $\xi = \sigma_\xi \cdot \xi_0$ for some ξ_0 with i.i.d. mean zero, unit variance and uniform sub-gaussian entries.

The requirement on the noise level σ_ξ^2 will be specified in concrete results below.

2.3. **The fixed point equation.**

Fix $\eta \geq 0$. Consider the following fixed point equation in (γ,τ):

$$\begin{cases} \phi\gamma^2 = \sigma_\xi^2 + \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau), \\ \phi - \frac{\eta}{\tau} = \frac{1}{n}\,\mathrm{tr}\,\big((\Sigma+\tau I_n)^{-1}\Sigma\big) = \frac{1}{\gamma^2}\,\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau). \end{cases} \tag{2.1}$$

We first establish some qualitative properties for the solution of (2.1).

Proposition 2.1. *Recall $\mathcal{H}_\Sigma = \mathrm{tr}(\Sigma^{-1})/n$. The following hold.*

(1) The fixed point equation (2.1) admits a unique solution $(\gamma_{\eta,*},\tau_{\eta,*}) \in (0,\infty)^2$, for all $(m,n) \in \mathbb{N}^2$ when $\eta > 0$ and $m < n$ when $\eta = 0$.

(2) *Suppose* $1/K \leq \phi^{-1} \leq K$ *and* $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \leq K$ *for some* $K > 1$. *Then there exists some* $C = C(K) > 1$ *such that uniformly in* $\eta \in \Xi_K$,

$$1/C \leq \tau_{\eta,*} \leq C, \quad 1/C \leq (-1)^{q+1}\partial_\eta^q \tau_{\eta,*} \leq C, \quad q \in \{1,2\}.$$

If furthermore $1/K \leq \sigma_\xi^2 \leq K$ *and* $\|\mu_0\| \leq K$, *then uniformly in* $\eta \in \Xi_K$,

$$1/C \leq \gamma_{\eta,*} \leq C, \quad |\partial_\eta \gamma_{\eta,*}| \leq C.$$

(3) *Suppose* $1/K \leq \phi^{-1} \leq K$ *and* $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \leq K$ *for some* $K > 1$. *Then there exists some* $C = C(K) > 1$ *such that the following hold. For any* $\varepsilon \in (0, 1/2]$, *we may find some* $\mathcal{U}_\varepsilon \subset B_n(1)$ *with* $\mathrm{vol}(\mathcal{U}_\varepsilon)/\mathrm{vol}(B_n(1)) \geq 1 - C\varepsilon^{-1}e^{-n\varepsilon^2/C}$,

$$\sup_{\mu_0 \in \mathcal{U}_\varepsilon} \sup_{\eta \in \Xi_K} \left|\gamma_{\eta,*}^2 - \widetilde{\gamma}_{\eta,*}^2(\|\mu_0\|)\right| \leq \varepsilon,$$

where $\widetilde{\gamma}_{\eta,*}^2(\|\mu_0\|) \equiv \sigma_\xi^2 \partial_\eta \tau_{\eta,*} + \|\mu_0\|^2(\tau_{\eta,*} - \eta\partial_\eta\tau_{\eta,*}) > 0$. *When* $\Sigma = I_n$, *we may take* $\mathcal{U}_\varepsilon = B_n(1)$ *and the above inequality holds with* $\varepsilon = 0$.

The above proposition combines parts of Propositions 8.1 and 11.2.

As an important qualitative consequence of (2), under the condition $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \leq K$, the effective regularization $\eta \mapsto \tau_{\eta,*}$ is a strictly increasing and concave function of η. Moreover, in the overparametrized regime $\phi^{-1} > 1$, the quantity $\tau_{0,*}$—also known as '*implicit regularization*' in the literature [BLLT20, BMR21, CM22, HMRT22, TB22]—is strictly bounded away from zero.

The claim in (3) offers a useful approximate representation of the effective noise $\gamma_{\eta,*}^2$ in terms of the original noise σ_ξ^2, the effective regularization $\tau_{\eta,*}$ and the signal energy $\|\mu_0\|$ without explicitly dependence of Σ. This representation will prove useful in the approximate ℓ_q risk formulae in Theorem 2.5, as well as in understanding some qualitative aspects of the risk curves in Section 3 ahead.

2.4. **Some connections of (2.1) to RMT.** The second equation of (2.1) has a natural connection to random matrix theory (RMT). To detail this connection, let $\widehat{\Sigma} \equiv \Sigma^{1/2}G^\top G\Sigma^{1/2}/m \in \mathbb{R}^{n\times n}$ and $\check{\Sigma} \equiv G\Sigma G^\top/m \in \mathbb{R}^{m\times m}$ be the sample covariance matrix and its dimension flipped, companion matrix. For $z \in \mathbb{C}^+ \equiv \{z \in \mathbb{C} : \Im z > 0\}$, let $\mathfrak{m}_n(z) \equiv m^{-1}\,\mathrm{tr}\,(\check{\Sigma} - zI_m)^{-1}$ and $\mathfrak{m}(z)$ be the Stieltjes transforms of the empirical spectral distribution and the asymptotic eigenvalue density (cf. [KY17, Definition 2.3]) of $\check{\Sigma}$, respectively. It is well-known that $\mathfrak{m}(z)$ can be determined uniquely via the fixed point equation

$$z = -\frac{1}{\mathfrak{m}(z)} + \frac{1}{\phi}\cdot\frac{1}{n}\,\mathrm{tr}\Big((I_n + \Sigma\mathfrak{m}(z))^{-1}\Sigma\Big). \tag{2.2}$$

See, e.g., [KY17, Lemma 2.2] for more technical details and historical references. We also note that while the above equation is initially defined for $z \in \mathbb{C}^+$, it can be straightforwardly extended to the real axis provided that z lies outside the support of the asymptotic spectrum of $\check{\Sigma}$.

The following proposition provides a precise connection between the effective regularization $\tau_{\eta,*}$ defined via the second equation of (2.1), and the Stieltjes transform $\mathfrak{m}$. This connection will prove important in some of the results ahead.

Proposition 2.2. *For any $\eta > 0$ and $\eta = 0$ with $\phi^{-1} > 1$,*

$$n^{-1}\mathrm{tr}\big((\Sigma + \tau_{\eta,*}I_n)^{-1}\Sigma\big) = \phi - \eta \cdot \mathfrak{m}(-\eta/\phi). \tag{2.3}$$

Proof. By comparing (2.2) and the second equation of (2.1), we may identify the two equations by setting $\tau_{\eta,*} \equiv 1/\mathfrak{m}(-z_\eta)$ with $z_\eta \equiv \eta/\phi$, as claimed. □

While (2.3) appears somewhat purely algebraic, it actually admits a natural statistical interpretation. Suppose ξ is also Gaussian. We may then compute

$$\mathrm{df}(\widehat{\mu}_\eta) = \sum_{j=1}^{n} \frac{\mathrm{Cov}^X\big((X\widehat{\mu}_\eta)_j, Y_j\big)}{\sigma_\xi^2} = \mathrm{tr}\big((\widehat{\Sigma} + z_\eta I_n)^{-1}\widehat{\Sigma}\big) = n(\phi - \eta \cdot \mathfrak{m}_n(-z_\eta)). \tag{2.4}$$

Now comparing the above display with (2.3), we arrive at the following intriguing equivalence between the averaged law in RMT, and the proximity of $\widehat{\mu}_\eta$ and $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})$ in terms of "degrees-of-freedom":

$$\mathfrak{m}_n(-z_\eta) \overset{\mathbb{P}}{\approx} \mathfrak{m}(-z_\eta) \ \Leftrightarrow\ \mathrm{df}(\widehat{\mu}_\eta) \overset{\mathbb{P}}{\approx} \mathrm{df}\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})\big).$$

Below we will show the above proximity of $\widehat{\mu}_\eta$ and $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})$ can be taken as far as the distributions of the two themselves.

2.5. **Distribution of Ridge(less) estimators.** In addition to $\widehat{\mu}_\eta$, we will also consider the distribution of the (scaled) residual $\widehat{r}_\eta$, defined by

$$\widehat{r}_\eta \equiv \frac{1}{\sqrt{n}}(Y - X\widehat{\mu}_\eta). \tag{2.5}$$

We define the 'population' version of $\widehat{r}_\eta$ as

$$r_{\eta,*} \equiv \frac{\eta}{\phi\tau_{\eta,*}}\Big(- \sqrt{\phi\gamma_{\eta,*}^2 - \sigma_\xi^2} \cdot \frac{h}{\sqrt{n}} + \frac{\xi}{\sqrt{n}}\Big). \tag{2.6}$$

Here $h \sim \mathcal{N}(0, I_m)$ is independent of ξ.

We are now in a position to state our main results on the distributional results for the Ridge(less) estimator $\widehat{\mu}_\eta$ and the residual $\widehat{r}_\eta$.

First we work under the Gaussian design $Z = G$, and we write $\widehat{\mu}_\eta = \widehat{\mu}_{\eta;G}, \widehat{r}_\eta = \widehat{r}_{\eta;G}$. Recall $\mathcal{H}_\Sigma = \mathrm{tr}(\Sigma^{-1})/n$.

Theorem 2.3. *Suppose Assumption A holds with $Z = G$ and the following hold for some $K > 0$.*

- $1/K \le \phi^{-1} \le K$, $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$.
- *Assumption B holds with $\sigma_\xi^2 \in [1/K, K]$.*

Then there exists some constant $C = C(K) > 0$ such that the following hold.

(1) *For any 1-Lipschitz function $\mathsf{g} : \mathbb{R}^n \to \mathbb{R}$ and $\varepsilon \in (0, 1/2]$,*

$$\sup_{\mu_0 \in B_n(1)} \mathbb{P}\Big(\sup_{\eta \in \Xi_K} \big|\mathsf{g}(\widehat{\mu}_{\eta;G}) - \mathbb{E}\,\mathsf{g}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}))\big| \ge \varepsilon\Big) \le Cne^{-n\varepsilon^4/C}.$$

(2) *For any* $\varepsilon \in (0,1/2]$, $\xi \in \mathbb{R}^m$ *satisfying* $|\,\|\xi\|^2/m - \sigma_\xi^2| \leq \varepsilon^2/C$, *and* 1*-Lipschitz function* $\mathsf{h}: \mathbb{R}^m \to \mathbb{R}$ *(which may depend on* ξ*),*

$$\sup_{\mu_0 \in B_n(1)} \mathbb{P}^\xi\Big(\sup_{\eta \in [1/K, K]} |\mathsf{h}(\widehat{r}_{\eta;G}) - \mathbb{E}^\xi \mathsf{h}(r_{\eta,*})| \geq \varepsilon \Big) \leq Cne^{-n\varepsilon^4/C}.$$

The choice $\mu_0 \in B_n(1)$ is made merely for simplicity of presentation; it can be replaced by $\mu_0 \in B_n(R)$ with another constant C that depends further on R. The assumption $\mathcal{H}_\Sigma \lesssim 1$ is quite common in the literature of Ridge(less) regression; see, e.g., [BMR21, Assumption 4.12] or a slight variant in [MRSY23, Assumption 1]. The major assumption in the above theorem is the Gaussianity on the design X. This may be lifted at the cost of a set of slightly stronger conditions.

Theorem 2.4. *Suppose Assumption A holds and the following hold for some* $K > 0$.

- $1/K \leq \phi^{-1} \leq K$, $\|\Sigma\|_{\mathrm{op}} \vee \|\Sigma^{-1}\|_{\mathrm{op}} \leq K$.
- *Assumption B holds with* $\sigma_\xi^2 \in [1/K, K]$.

Fix $\vartheta \in (0, 1/18)$. *There exist some* $C = C(K,\vartheta) > 0$ *and two measurable sets* $\mathcal{U}_\vartheta \subset B_n(1), \mathcal{E}_\vartheta \subset \mathbb{R}^m$ *with* $\min\{\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)), \mathbb{P}(\xi \in \mathcal{E}_\vartheta)\} \geq 1 - Ce^{-n^{2\vartheta}/C}$, *such that the following hold.*

(1) *For any* 1*-Lipschitz function* $\mathsf{g}: \mathbb{R}^n \to \mathbb{R}$, *and* $\varepsilon \in (0,1/2]$,

$$\sup_{\mu_0 \in \mathcal{U}_\vartheta} \mathbb{P}\Big(\sup_{\eta \in \Xi_K} \big| \mathsf{g}(\widehat{\mu}_\eta) - \mathbb{E}\, \mathsf{g}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*}; \tau_{\eta,*})) \big| \geq \varepsilon \Big) \leq C\varepsilon^{-13} n^{-1/6+3\vartheta}.$$

(2) *For any* $\varepsilon \in (0,1/2]$, $\xi \in \mathcal{E}_\vartheta$ *and* 1*-Lipschitz function* $\mathsf{h}: \mathbb{R}^m \to \mathbb{R}$ *(which may depend on* ξ*),*

$$\sup_{\mu_0 \in \mathcal{U}_\vartheta} \mathbb{P}^\xi\Big(\sup_{\eta \in [1/K, K]} |\mathsf{h}(\widehat{r}_\eta) - \mathbb{E}^\xi \mathsf{h}(r_{\eta,*})| \geq \varepsilon \Big) \leq C\varepsilon^{-7} n^{-1/6+3\vartheta}.$$

Concrete forms of $\mathcal{U}_\vartheta, \mathcal{E}_\vartheta$ *are specified in Proposition 10.3.*

Theorems 2.3 and 2.4 are proved in Section 9 and Section 10, respectively. Due to the high technicalities in the proof, a sketch is outlined in Section 6.

We mention two particular important features on the theorems above:

(1) The distributional characterizations for $\widehat{\mu}_\eta$ in both theorems above are uniformly valid down to the interpolation regime $\eta = 0$ for $\phi^{-1} > 1$. This uniform control will play a crucial role in our non-asymptotic analysis of cross-validation methods to be studied in Section 4 ahead.
(2) The distribution of the residual $\widehat{r}_\eta$ in (2) is formulated *conditional on the noise* ξ. A fundamental reason for adopting this formulation is that the distribution of $\widehat{r}_\eta$ is *not* universal with respect to the law of ξ. In other words, one cannot simply assume Gaussianity of ξ in Theorem 2.3 in hope of proving universality of $\widehat{r}_\eta$ in Theorem 2.4.

In the context of distributional characterizations for regularized regression estimators in the proportional regime, results in similar vein to Theorem 2.3 have been obtained in the closely related Lasso setting for isotropic $\Sigma = I_n$ in [MM21], and for general Σ in [CMW22], both under Gaussian designs and with strictly non-vanishing regularization. A substantially simpler, isotropic ($\Sigma = I_n$) version of

Theorem 2.4 is obtained in [HS22] that holds pointwise in non-vanishing regularization level $\eta > 0$. As will be clear from the proof sketch in Section 6, in addition to the complications due to the implicit nature of the solution to the fixed point equation (2.1) for general Σ, the major difficulty in proving Theorems 2.3 and 2.4 rests in handling the singularity of the optimization problem (1.2) as $\eta \downarrow 0$.

2.6. **Weighted ℓ_q risks of $\widehat{\mu}_\eta$.** As a demonstration of the analytic power of the above Theorems 2.3 and 2.4, we compute below the weighted ℓ_q risk $\|\mathsf{A}(\widehat{\mu}_\eta - \mu_0)\|_q$ for a well-behaved matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$ and $q \in (0,\infty)$.

Theorem 2.5. *Suppose the same conditions in Theorem 2.4 hold for some $K > 0$. Fix $q \in (0,\infty)$ and a p.s.d. matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$ with $\|\mathsf{A}\|_{\mathrm{op}} \vee \|\mathsf{A}^{-1}\|_{\mathrm{op}} \leq K$. Then there exist constants $C > 1, \vartheta \in (0, 1/50)$ depending on K, q, and a measurable set $\mathcal{U}_\vartheta \subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \geq 1 - Ce^{-n^\vartheta/C}$, such that*

$$\sup_{\mu_0\in\mathcal{U}_\vartheta} \mathbb{P}\bigg(\sup_{\eta\in\Xi_K}\bigg|\frac{\|\mathsf{A}(\widehat{\mu}_\eta - \mu_0)\|_q}{\bar{R}^{\mathsf{A}}_{(\Sigma,\mu_0);q}(\eta)} - 1\bigg| \geq n^{-\vartheta}\bigg) \leq Cn^{-1/7}.$$

Here $\bar{R}^{\mathsf{A}}_{(\Sigma,\mu_0);q}(\eta) \in \{\mathbb{E}\|\mathsf{A}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0)\|_q, n^{-1/2}\|\mathrm{diag}(\Gamma^{\mathsf{A}}_{\eta;(\Sigma,\|\mu_0\|)})\|^{1/2}_{q/2}M_q\}$, where $M_q \equiv \mathbb{E}^{1/q}|\mathcal{N}(0,1)|^q = 2^{1/2}\{\Gamma((q+1)/2)/\sqrt{\pi}\}^{1/q}$,

$$\Gamma^{\mathsf{A}}_{\eta;(\Sigma,\|\mu_0\|)} \equiv \mathsf{A}(\Sigma + \tau_{\eta,*}I_n)^{-1}\Big(\widetilde{\gamma}^2_{\eta,*}(\|\mu_0\|)\Sigma + \tau^2_{\eta,*}\|\mu_0\|^2 I_n\Big)(\Sigma + \tau_{\eta,*}I_n)^{-1}\mathsf{A}, \tag{2.7}$$

and $\widetilde{\gamma}^2_{\eta,*}(\|\mu_0\|)$ is defined in Proposition 2.1-(3).

The proof of the above theorem can be found in Section 10.7. To the best of our knowledge, general weighted ℓ_q risks for the Ridge(less) estimator $\widehat{\mu}_\eta$ have not be available in the literature except for the special case $q = 2$, for which $\|\mathsf{A}(\widehat{\mu}_\eta - \mu_0)\|_2$ admits a closed-form expression in terms of the spectral statistics of X that facilitates direct applications of RMT techniques, cf. [TV+04, EK13, Dic16, DW18, EK18, ASS20, WX20, BMR21, RMR21, HMRT22, CM22].

Here, obtaining ℓ_q risks for $q \in (0,2)$ via our Theorems 2.3 and 2.4 is relatively easy, as $x \mapsto \|x\|_q/n^{1/q-1/2}$ is 1-Lipschitz with respect to $\|\cdot\|$ for $q \in (0,2)$. The stronger norm case $q \in (2,\infty)$ is significantly harder. In fact, we need additionally the following delocalization result for $\widehat{\mu}_\eta$.

Proposition 2.6. *Suppose the same conditions as in Theorem 2.5 hold for some $K > 0$. Fix $\vartheta \in (0, 1/2]$. Then there exist some constant $C = C(K,\vartheta) > 0$ and a measurable set $\mathcal{U}_\vartheta \subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \geq 1 - Ce^{-n^{2\vartheta}/C}$, such that*

$$\sup_{\mu_0\in\mathcal{U}_\vartheta} \mathbb{P}\Big(\sup_{\eta\in\Xi_K}\|\mathsf{A}(\widehat{\mu}_\eta - \mu_0)\|_\infty \geq Cn^{-1/2+\vartheta}\Big) \leq Cn^{-100}.$$

The above proposition is a simplified version of Proposition 10.3, proved via the anisotropic local laws developed in [KY17]. In essence, delocalization allows us to apply Theorems 2.3 and 2.4 with a truncated version of the ℓ_q norm ($q > 2$) with a well-controlled Lipschitz constant with respect to ℓ_2. Moreover, delocalization of $\widehat{\mu}_\eta$ also serves as a key technical ingredient in proving the universality Theorem 2.4; the readers are referred to Section 6 for a detailed account on the technical connection between delocalization and universality.

Convention on probability estimates:

(1) When $Z = G$, $n^{-1/7}$ in Theorem 2.5 can be replaced by n^{-D} for any $D > 0$.
(2) n^{-100} in Proposition 2.6 can be replaced by n^{-D} for any $D > 0$.

The cost will be a possibly enlarged constant $C > 0$ that depends further on D. This convention applies to other statements in the following sections in which the probability estimates $n^{-1/7}, n^{-100}$ appear.

3. ℓ_2 RISK FORMULAE AND PHASE TRANSITIONS

In this section, we will study in some detail the behavior of various ℓ_2 risks associated with $\widehat{\mu}_\eta$. Compared to the general ℓ_q risks as in Theorem 2.5, the major additional analytic advantage of working with ℓ_2 risks is its close connection to techniques from random matrix theory (RMT). As will be clear from below, this allows us to characterize certain global, qualitative behavior of these ℓ_2 risk curves with respect to the regularization level η.

3.1. **Definitions of various ℓ_2 risks.** Recall the notation $R^{\#}_{(\Sigma,\mu_0)}(\eta)$ defined in Section 1.3. Let their 'theoretical' versions be defined as follows:

- $\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) \equiv \tau_{\eta,*}^2 \|(\Sigma + \tau_{\eta,*} I_n)^{-1}\Sigma^{1/2}\mu_0\|^2 + \frac{\gamma_{\eta,*}^2}{n}\,\mathrm{tr}\,(\Sigma^2(\Sigma + \tau_{\eta,*} I_n)^{-2})$.
- $\bar{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) \equiv \tau_{\eta,*}^2 \|(\Sigma + \tau_{\eta,*} I_n)^{-1}\mu_0\|^2 + \frac{\gamma_{\eta,*}^2}{n}\,\mathrm{tr}\,(\Sigma(\Sigma + \tau_{\eta,*} I_n)^{-2})$.
- $\bar{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta) \equiv (\frac{\eta\gamma_{\eta,*}}{\tau_{\eta,*}})^2 + \phi\sigma_\xi^2 \cdot (1 - \frac{2\eta}{\phi\tau_{\eta,*}})$.

We also define the residual and its theoretical version as

- $R^{\mathsf{res}}_{(\Sigma,\mu_0)}(\eta) \equiv n^{-1}\|Y - X\widehat{\mu}_\eta\|^2$, $\bar{R}^{\mathsf{res}}_{(\Sigma,\mu_0)}(\eta) \equiv (\frac{\eta\gamma_{\eta,*}}{\tau_{\eta,*}})^2$.

The following theorem follows easily from Theorems 2.3 and 2.4. The proof of this and all other results in this section can be found in Section 11.

Theorem 3.1. *Suppose Assumption A holds and the following hold for some $K > 0$.*

- $1/K \le \phi^{-1} \le K$, $\|\Sigma^{-1}\|_{\mathrm{op}} \vee \|\Sigma\|_{\mathrm{op}} \le K$.
- *Assumption B holds with $\sigma_\xi^2 \in [1/K, K]$.*

Fix a small enough $\vartheta \in (0, 1/50)$. Then there exist a constant $C = C(K, \vartheta) > 1$, and a measurable set $\mathcal{U}_\vartheta \subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \ge 1 - Ce^{-n^\vartheta/C}$, such that for any $\varepsilon \in (0, 1/2]$, and $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}, \mathsf{res}\}$,

$$\sup_{\mu_0 \in \mathcal{U}_\vartheta} \mathbb{P}\Big(\sup_{\eta \in \Xi^{\#}} |R^{\#}_{(\Sigma,\mu_0)}(\eta) - \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)| \ge \varepsilon\Big) \le C \cdot \begin{cases} ne^{-n\varepsilon^4/C}, & Z = G; \\ \varepsilon^{-c_0} n^{-1/6.5}, & \textit{otherwise}. \end{cases}$$

Here $\Xi^{\#} = \Xi_K$ for $\# \in \{\mathsf{pred}, \mathsf{est}\}$ and $\Xi^{\#} = [1/K, K]$ for $\# \in \{\mathsf{in}, \mathsf{res}\}$, and $c_0 > 0$ is universal. Moreover, when $Z = G$, the supremum in the above display extends to $\mu_0 \in B_n(1)$, and the constant $C > 0$ does not depend on ϑ.

Remark 1. For $\# \in \{\mathsf{in}, \mathsf{res}\}$, we may take $\Xi^{\#} = \Xi_K$ at the cost of an worsened probability estimate $C(ne^{-n\varepsilon^{c_0}/C} + \varepsilon^{-c_0} n^{-1/6.5} \mathbf{1}_{Z \ne G})$, cf. Lemma 11.4.

A substantial recent line of research has focused on understanding the behavior of $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$, with a special emphasis for the interpolating regime $\eta \approx 0$ when $\phi^{-1} > 1$, cf. [BLLT20, KLS20, WX20, BMR21, KZSS21, RMR21, CM22, HMRT22, TB22, ZKS+22]. We refer the readers to [TB22, Section 1.2 and Section 9] for a thorough account and a state-of-art review on various results on $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(0)$ and their comparisons.

Here, the closest non-asymptotic results on exact risk characterizations related to our Theorem 3.1, appear to be those presented in (i) [HMRT22, Theorems 2 and 5], which proved non-asymptotic additive approximations $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) = \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) + \mathfrak{o}_{\mathbf{P}}(1)$, and (ii) [CM22, Theorems 1 and 2], which provided substantially refined, multiplicative approximations $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)/\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) = 1 + \mathfrak{o}_{\mathbf{P}}(1)$ that hold beyond the proportional regime. Both works [HMRT22, CM22] leverage the closed form of the Ridge(less) estimator $\widehat{\mu}_\eta$ to analyze the bias and variance terms in $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$, by means of calculus for the resolvent of the sample covariance. Their analysis works under $\eta \gg n^{-c_0}$ for some suitable $c_0 > 0$. For the case $\# = \mathsf{pred}$, Theorem 3.1 above complements the results in [HMRT22, CM22] by providing uniform control in η when $\phi^{-1} > 1$ (under a set of different conditions).

3.2. **Approximate representation of $\bar{R}^{\#}_{(\Sigma,\mu_0)}$ via RMT.** Recall the Stieltjes transformation $\mathfrak{m}$ defined via (2.2), and the signal-to-noise ratio $\mathsf{SNR}_{\mu_0} = \|\mu_0\|^2/\sigma_\xi^2$ (for $\sigma_\xi^2 = 0$, we interpret $\sigma_\xi^2 \cdot \mathsf{SNR}_{\mu_0} = \|\mu_0\|^2$). The following theorem provides an efficient RMT representation of $\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ that holds for 'most' μ_0's.

Theorem 3.2. *Suppose $1/K \le \phi^{-1} \le K$, $\sigma_\xi^2 \in [0,K]$ and $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ for some $K > 0$. There exists some constant $C = C(K) > 0$ such that for any $\varepsilon \in (0,1/2]$, we may find a measurable set $\mathcal{U}_\varepsilon \subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_\varepsilon)/\mathrm{vol}(B_n(1)) \ge 1 - C\varepsilon^{-1}e^{-n\varepsilon^2/C}$,*

$$\sup_{\mu_0\in\mathcal{U}_\varepsilon}\sup_{\eta\in\Xi_K} |\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)| \le \varepsilon. \tag{3.1}$$

Here with $\mathfrak{m}_\eta \equiv \mathfrak{m}(-\eta/\phi)$ and $\mathfrak{m}'_\eta \equiv \mathfrak{m}'(-\eta/\phi)$,

- $\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) \equiv \sigma_\xi^2 \cdot \Big\{\frac{1}{\mathfrak{m}_\eta^2}\Big(\phi \cdot \mathsf{SNR}_{\mu_0}\mathfrak{m}_\eta - (\eta\cdot \mathsf{SNR}_{\mu_0} - 1)\mathfrak{m}'_\eta\Big) - 1\Big\}$,
- $\mathscr{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) \equiv \sigma_\xi^2 \cdot \Big\{\mathsf{SNR}_{\mu_0}(1-\phi) + \mathfrak{m}_\eta + \frac{\eta}{\phi}(\eta\cdot\mathsf{SNR}_{\mu_0} - 1)\mathfrak{m}'_\eta\Big\}$,
- $\mathscr{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta) \equiv \sigma_\xi^2 \cdot \frac{\eta^2}{\phi}\Big(\phi\cdot\mathsf{SNR}_{\mu_0}\mathfrak{m}_\eta - (\eta\cdot\mathsf{SNR}_{\mu_0} - 1)\mathfrak{m}'_\eta\Big) + \sigma_\xi^2 \cdot (\phi - 2\eta\mathfrak{m}_\eta)$,
- $\mathscr{R}^{\mathsf{res}}_{(\Sigma,\mu_0)}(\eta) \equiv \sigma_\xi^2 \cdot \frac{\eta^2}{\phi}\Big(\phi\cdot\mathsf{SNR}_{\mu_0}\mathfrak{m}_\eta - (\eta\cdot\mathsf{SNR}_{\mu_0} - 1)\mathfrak{m}'_\eta\Big)$.

When $\Sigma = I_n$, we may take $\mathcal{U}_\varepsilon = B_n(1)$ and (3.1) holds with $\varepsilon = 0$.

The RMT representation above yields a crucial insight into the extremal behavior of the risk maps $\eta \mapsto \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$. In fact, the following derivative formulae hold.

Proposition 3.3. *For $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$,*

$$\partial_\eta \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) = \sigma_\xi^2 \cdot \mathfrak{M}^{\#}(\eta)\cdot(\eta\cdot\mathsf{SNR}_{\mu_0} - 1).$$

Here with ρ denoting the asymptotic eigenvalue density of $\check{\Sigma}$, and $z_\eta \equiv \eta/\phi$,

$$\mathfrak{M}^{\#}(\eta) = 2\cdot\begin{cases}(\phi \mathfrak{m}_\eta^3)^{-1}\Big\{\int \frac{\rho(\mathrm{d}x)}{(x+z_\eta)}\int \frac{\rho(\mathrm{d}x)}{(x+z_\eta)^3} - \Big(\int \frac{\rho(\mathrm{d}x)}{(x+z_\eta)^2}\Big)^2\Big\}, & \# = \mathsf{pred};\\ \int \frac{x}{(x+z_\eta)^3}\,\rho(\mathrm{d}x), & \# = \mathsf{est};\\ \int \frac{x^2}{(x+z_\eta)^3}\,\rho(\mathrm{d}x), & \# = \mathsf{in}.\end{cases}$$

As $\mathrm{supp}(\rho) \subset [0,\infty)$ (cf. [KY17, Lemma 2.2]), it is easy to see $\mathfrak{M}^{\#} \geq 0$ using the above integral representation via ρ. In Proposition 11.3 ahead, we will give a different representation of $\mathfrak{M}^{\#}$ via the effective regularization $\tau_{\eta,*}$. With the help of the stability estimates of $\tau_{\eta,*}$ in Proposition 2.1, this representation allows us to derive a much stronger estimate

$$1/C \leq \mathfrak{M}^{\#}(\eta) \leq C, \quad \forall \eta \in \Xi_K, \quad \# \in \{\mathsf{pred},\mathsf{est},\mathsf{in}\}, \tag{3.2}$$

under the condition $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \leq K$.

A particular important consequence of (3.2) is that, the maps $\eta \mapsto \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ are approximately (locally) quadratic functions centered *at the same point* $\eta_* = \mathsf{SNR}_{\mu_0}^{-1}$ *for all* $\# \in \{\mathsf{pred},\mathsf{est},\mathsf{in}\}$. Therefore, in view of Theorems 3.1 and 3.2, approximately so do the risk maps $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}(\eta), \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ for 'most' μ_0's. Moreover, the approximate local quadraticity of $\eta \mapsto \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ due to (3.2) allows one to relate tightly the change of value in η and that of the actual risks $R^{\#}_{(\Sigma,\mu_0)}(\eta)$. This will play an important technical role in validating the optimality of cross-validation schemes beyond prediction errors in Section 4 ahead.

Remark 2. Proposition 3.3 can also be used to study certain qualitative features of the optimally tuned (theoretical) risks $\mathsf{OPT}^{\#}_{(\Sigma,\mu_0)} \equiv \min_{\eta\geq 0}\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)/\sigma_\xi^2 = \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta_*)/\sigma_\xi^2$ for $\# \in \{\mathsf{pred},\mathsf{est},\mathsf{in}\}$. Some further results in this direction are detailed in Appendix A. In particular, Proposition A.1 therein proves the monotonicity of $\mathsf{OPT}^{\#}_{(\Sigma,\mu_0)}$ with respect to the aspect ratio ϕ for $\# \in \{\mathsf{pred},\mathsf{est},\mathsf{in}\}$.

3.3. **Phase transitions on the optimality of interpolation.** The fact that the maps $\eta \mapsto \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ admit the same global minimizer $\eta_* = \mathsf{SNR}_{\mu_0}^{-1} = \sigma_\xi^2/\|\mu_0\|^2$ due to Proposition 3.3 may be viewed from a different perspective: $\eta_* = 0$ if and only if $\sigma_\xi^2 = 0$ for 'most' signal μ_0's. Combined with Theorems 3.1 and 3.2, it naturally suggests that for 'most' μ_0's, interpolation is optimal simultaneously for prediction, estimation and in-sample risks, if and only if the noise level is (nearly) zero. The following theorem makes this precise.

Theorem 3.4. *Suppose Assumptions A-B hold, and* $\|\Sigma^{-1}\|_{\mathrm{op}} \vee \|\Sigma\|_{\mathrm{op}} \leq K$ *for some* $K > 0$. *Fix a small enough* $\vartheta \in (0, 1/50)$. *The following hold for all* $\# \in \{\mathsf{pred},\mathsf{est},\mathsf{in}\}$.

(1) *(**Noisy case**). Suppose* $1/K \leq \phi^{-1} \leq K$ *and* $1/K \leq \sigma_\xi^2 \leq K$. *Fix* $\delta \in (0, 1/2]$ *and* $L \geq K/\delta^2$. *There exist a constant* $C = C(K, L, \delta, \vartheta) > 0$ *and a measurable set* $\mathcal{U}_{\delta,\vartheta} \subset B_n(1) \setminus B_n(\delta)$ *with* $\mathrm{vol}(\mathcal{U}_{\delta,\vartheta})/\mathrm{vol}(B_n(1) \setminus B_n(\delta)) \geq 1 - Ce^{-n^{\vartheta}/C}$,

such that

$$\sup_{\mu_0\in\mathcal{U}_{\delta,\vartheta}} \mathbb{P}\Big(\inf_{\eta'\in\Xi_L:|\eta'-\mathsf{SNR}_{\mu_0}^{-1}|\geq\delta} \big|R^{\#}_{(\Sigma,\mu_0)}(\eta') - \min_{\eta\in\Xi_L} R^{\#}_{(\Sigma,\mu_0)}(\eta)\big| < \frac{1}{C}\Big) \leq Cn^{-1/7}.$$

(2) (***Noiseless case***). *Suppose* $1 + 1/K \leq \phi^{-1} \leq K$ *and* $\sigma_\xi^2 = 0$. *There exist a constant* $C = C(K, \vartheta) > 0$ *and a measurable set* $\mathcal{U}_\vartheta \subset B_n(1)$ *with* $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \geq 1 - Ce^{-n^{\vartheta}/C}$, *such that*

$$\sup_{\mu_0\in\mathcal{U}_\vartheta} \mathbb{P}\Big(R^{\#}_{(\Sigma,\mu_0)}(0) \geq \min_{\eta\in[0,K]} R^{\#}_{(\Sigma,\mu_0)}(\eta) + n^{-\vartheta}\Big) \leq Cn^{-1/7}.$$

We note that the sub-optimality of interpolation in the noisy setting $\sigma_\xi^2 > 0$, at least for 'most' μ_0's as described in the above theorem, does not contradict the possible benign behavior of the prediction risk $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(0)$ extensively studied in the literature (cited after Theorem 3.1). In fact, from a conceptual standpoint, the observation that interpolation may fall short while a certain amount of regularization remains advantageous in 'typical' scenarios aligns closely with the conventional statistical wisdom, which emphasizes the vital role of regularization in striking a balance between the bias and variance [JWHT21].

On the other hand, the optimality of interpolation in the noiseless setting $\sigma_\xi^2 = 0$ stated in the above theorem is not as intuitively obvious. This phenomenon can be elucidated from our distributional characterizations in Theorems 2.3 and 2.4. As the effective regularization $\tau_{\eta,*}$ decreases as $\eta \downarrow 0$, and for $\sigma_\xi^2 = 0$, the effective noise $\gamma_{\eta,*}^2 \approx \|\mu_0\|^2(\tau_{\eta,*} - \eta\partial_\eta\tau_{\eta,*})$ (cf. Proposition 2.1-(3)) also decreases as $\eta \downarrow 0$, both the bias and variance of $\widehat{\mu}_\eta$ are also expected to decrease as $\eta \downarrow 0$, at least for $\# = \mathsf{est}$ and for 'most' μ_0's. The above theorem rigorously establishes this heuristic for all the prediction, estimation and in-sample risks.

Remark 3. Some remarks on the connection of Theorem 3.4 to the literature:

(1) Theorem 3.4-(1) is related to some results in [CDK22] which considers a Bayesian setting with an isotropic prior on μ_0 in that $\mathbb{E}\mu_0 = 0$ and $\mathrm{Cov}(\mu_0) = I_n/n$. In particular, [CDK22, Theorem 3-(ii)] shows that, using our notation, in the proportional asymptotics $m/n \to \phi \in (0,\infty)$, if $\sigma_\xi^2 > 0$ and the empirical spectral distribution of Σ converges, then with $\mathbb{P}^X$-probability 1, $\lim_{n\to\infty}|\mathbb{E}[R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta_*)|X] - \min_{\eta'>0}\mathbb{E}[R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta')|X]| = 0$. Here for the case $\# = \mathsf{pred}$, our Theorem 3.4-(1) above provides a non-asymptotic, and more importantly, a non-Bayesian version that holds for 'most' μ_0's.

(2) As mentioned before, the recent work [CM22] proves an important, multiplicative characterization $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)/\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) = 1 + \mathfrak{o}_{\mathbf{P}}(1)$ beyond the proportional regime. In particular, the multiplicative formulation encompasses several important cases of data covariance Σ with decaying eigenvalues that lead to benign overfitting $\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(0) \ll 1$, cf. [CM22, Section 4.2]. We conjecture that Theorem 3.4 also remains valid for such irregular data covariance, in a similar multiplicative formulation. However, formally establishing this validity remains an interesting open question.

4. Cross-validation: optimality beyond prediction

This section is devoted to the validation of the broad optimality of two widely used cross-validation schemes beyond the prediction risk. Some consequences to statistical inference via debiased Ridge(less) estimators will also be discussed.

4.1. **Estimation of effective noise and regularization.** We shall first take a slight detour, by considering estimation of the effective regularization $\tau_{\eta,*}$ and the effective noise $\gamma_{\eta,*}$. We propose the following estimators:

$$\begin{cases}\widehat{\tau}_\eta \equiv \Big\{\frac{1}{m}\operatorname{tr}\big(\frac{1}{m}XX^\top + \frac{\eta}{\phi}I_m\big)^{-1}\Big\}^{-1} = \big\{\operatorname{tr}(XX^\top + \eta\cdot nI_m)^{-1}\big\}^{-1},\\ \widehat{\gamma}_\eta \equiv \frac{\widehat{\tau}_\eta}{\sqrt{n}}\big(\eta^{-1}\|Y - X\widehat{\mu}_\eta\|\mathbf{1}_{\phi^{-1}<1} + \|(XX^\top/n)^{-1}X\widehat{\mu}_\eta\|\mathbf{1}_{\phi^{-1}\geq 1}\big).\end{cases} \tag{4.1}$$

These estimators will not only be useful in their own rights, they will also play an important rule in understanding the generalized cross-validation scheme in the next subsection.

The following theorem provides a formal justification for $\widehat{\tau}_\eta, \widehat{\gamma}_\eta$ in (4.1). The proof of this and all other results in this section can be found in Section 12.

Theorem 4.1. *Suppose Assumption A holds, and* $1/K \leq \phi^{-1} \leq K$, $\|\Sigma^{-1}\|_{\mathrm{op}} \vee \|\Sigma\|_{\mathrm{op}} \leq K$ *hold for some* $K > 0$.

(1) For any small $\varepsilon > 0$, *there exists some* $C_1 = C_1(K, \varepsilon) > 0$ *such that*

$$\mathbb{P}\Big(\sup_{\eta\in\Xi_K}|\widehat{\tau}_\eta - \tau_{\eta,*}| \geq n^{-1/2+\varepsilon}\Big) \leq C_1 n^{-100}.$$

(2) Suppose further Assumption B holds with either (i) $\sigma_\xi^2 \in [1/K, K]$ *or (ii)* $\sigma_\xi^2 \in [0, K]$ *with* $1 + 1/K \leq \phi^{-1} \leq K$. *Fix a small enough constant* $\vartheta \in (0, 1/50)$. *Then there exist a constant* $C_2 = C_2(K, \vartheta) > 1$, *and a measurable set* $\mathcal{U}_\vartheta \subset B_n(1)$ *with* $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \geq 1 - Ce^{-n^\vartheta/C}$, *such that*

$$\sup_{\mu_0\in\mathcal{U}_\vartheta}\mathbb{P}\Big(\sup_{\eta\in\Xi_K}|\widehat{\gamma}_\eta - \gamma_{\eta,*}| \geq n^{-\vartheta}\Big) \leq C_2 n^{-1/7}.$$

Remark 4. The original noise level σ_ξ^2 can be consistently estimated when Σ is known. In particular, we may use

$$\widehat{\sigma}_\eta^2 \equiv \widehat{\gamma}_\eta^2(1 - \phi + 2\eta\widehat{\tau}_\eta^{-1}) - \widehat{\tau}_\eta^2\cdot\|\Sigma^{-1/2}\widehat{\mu}_\eta\|^2. \tag{4.2}$$

Estimators for σ_ξ^2 of a similar flavor for other convex regularized estimators under Gaussian designs can be found in [BEM13, Bel20, MM21]. An advantage of $\widehat{\sigma}_\eta^2$ in (4.2) is its validity in the interpolation regime when $\phi^{-1} > 1$. We may also replace $\widehat{\tau}_\eta$ in the above display by $\tau_{\eta,*}$, as it can be solved exactly with known Σ using the second equation of (2.1). A formal proof of $\widehat{\sigma}_\eta^2 \overset{\mathbb{P}}{\approx} \sigma_\xi^2$ can be carried out similar to that of Theorem 4.1-(2) above, so we omit the details.

4.2. **Validation of cross-validation schemes.**

4.2.1. *Generalized cross-validation.* Consider choosing η by minimizing the estimated effective noise $\widehat{\gamma}_\eta$ given in (4.1): for any $L > 0$,

$$\widehat{\eta}_L^{\mathsf{GCV}} \in \arg\min_{\eta \in \Xi_L} \widehat{\gamma}_\eta. \tag{4.3}$$

The subscript on L in $\widehat{\eta}_L^{\mathsf{GCV}}$ will usually be suppressed for notational simplicity.

The above proposal (4.3) is known in the literature as the *generalized cross validation* [CW79, GHW79] in the underparametrized regime $\phi^{-1} < 1$, with the same form of modification [HMRT22, Eqn. (48)] in the overparametrized regime $\phi^{-1} > 1$. The connection there is strongly tied to the so-called shortcut formula for leave-one-out cross validation that exists uniquely for Ridge regression, cf. [HMRT22, Eqn. (46)].

Here we take a different perspective on (4.3). From our developed theory, this tuning scheme is easily believed to "work" since

$$\widehat{\gamma}_\eta^2 \overset{\mathbb{P}}{\approx} \gamma_{\eta,*}^2 = \phi^{-1}(\sigma_\xi^2 + \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)) \overset{\mathbb{P}}{\approx} \phi^{-1}(\sigma_\xi^2 + R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)). \tag{4.4}$$

We therefore expect that minimization of $\eta \mapsto \widehat{\gamma}_\eta$ is approximately the same as that of $\eta \mapsto R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$. Moreover, in view of Theorem 3.4, the minimizer of the latter problem should be roughly the same as $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}(\eta)$ for $\# \in \{\mathsf{est}, \mathsf{in}\}$ for 'most' μ_0's. Now it is natural to expect that the tuning method $\widehat{\eta}^{\mathsf{GCV}}$ in (4.3)—which aims at minimizing the prediction risk—should simultaneously give optimal performance for all prediction, estimation and in-sample risks for 'most' signal μ_0's. We make precise the foregoing heuristics in the following theorem.

Theorem 4.2. *Suppose Assumption A holds, and the following hold some $K > 0$.*

- $1/K \le \phi^{-1} \le K$, $\|\Sigma^{-1}\|_{\mathrm{op}} \vee \|\Sigma\|_{\mathrm{op}} \le K$.
- *Assumption B holds with either (i) $\sigma_\xi^2 \in [1/K, K]$ or (ii) $\sigma_\xi^2 \in [0, K]$ with $\phi^{-1} \ge 1 + 1/K$.*

Fix $\delta \in (0, 1/2]$, $L \ge K/\delta^2$ and a small enough $\vartheta \in (0, 1/50)$. There exist a constant $C = C(K, L, \delta, \vartheta) > 0$ and a measurable set $\mathcal{U}_{\delta,\vartheta} \subset B_n(1) \setminus B_n(\delta)$ with $\mathrm{vol}(\mathcal{U}_{\delta,\vartheta})/\mathrm{vol}(B_n(1) \setminus B_n(\delta)) \ge 1 - Ce^{-n^\vartheta/C}$*, such that for $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$,*

$$\sup_{\mu_0 \in \mathcal{U}_{\delta,\vartheta}} \mathbb{P}\Big(R^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}_L^{\mathsf{GCV}}) \ge \min_{\eta \in \Xi_L} R^{\#}_{(\Sigma,\mu_0)}(\eta) + n^{-\vartheta}\Big) \le Cn^{-1/7}.$$

Earlier results for generalized cross validation in Ridge regression in low-dimensional settings include [Sto74, Sto77, CW79, Li85, Li86, Li87, DvdL05]. In the proportional high-dimensional regime, [HMRT22, Theorem 7] provides an asymptotic justification for the generalized cross validation under isotropic $\Sigma = I_n$ and an isotropic prior on μ_0. Subsequent improvement by [PWRT21, Theorem 4.1] allows for general Σ, deterministic μ_0's and a much larger range of regularization levels that include $\eta = 0$ and possibly even negative η's. Both works consider the optimality of $\widehat{\eta}^{\mathsf{GCV}}$ with respect to the prediction risk $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}$ in an asymptotic framework; the proofs therein rely intrinsically on the asymptotics.

Here in Theorem 4.2 above, we provide a non-asymptotic justification for the optimality of $\widehat{\eta}^{\mathsf{GCV}}$ that surprisingly holds simultaneously for all the three indicated risks. To the best of our knowledge, the optimality of the generalized cross validation in (4.3) beyond prediction risk has not been observed in prior literature.

4.2.2. *k-fold cross-validation.* Next we consider the widely used k-fold cross-validation. Before detailing the procedure, we need some further notation. Let m_ℓ be the sample size of batch $\ell \in [k]$, so $\sum_{\ell\in[k]} m_\ell = m$. In the standard k-fold cross validation, we choose equal sized batch with $m_\ell = m/k$ (assumed to be integer without loss of generality). Let $X^{(\ell)} \in \mathbb{R}^{m_\ell \times n}$ (resp. $Y^{(\ell)} \in \mathbb{R}^{m_\ell}$) be the submatrix of X (resp. subvector of Y) that contains all rows corresponding to the training data in batch ℓ. In a similar fashion, let $X^{(-\ell)} \in \mathbb{R}^{(m-m_\ell)\times n}$ (resp. $Y^{(-\ell)} \in \mathbb{R}^{m-m_\ell}$) be the submatrix of X (resp. subvector of Y) that removes all rows corresponding to $X^{(\ell)}$ (resp. $Y^{(\ell)}$).

The k-fold cross-validation works as follows. For $\ell \in [k]$, let $\widehat{\mu}_\eta^{(\ell)} \equiv \arg\min_{\mu\in\mathbb{R}^n} \{\frac{1}{2n}\|Y^{(-\ell)} - X^{(-\ell)}\mu\|^2 + \frac{\eta}{2}\|\mu\|^2\}$ be the Ridge estimator over $(X^{(-\ell)}, Y^{(-\ell)})$ with regularization $\eta \geq 0$. We then pick the tuning parameter that minimizes the averaged test errors of $\widehat{\mu}_\eta^{(\ell)}$ over $(X^{(\ell)}, Y^{(\ell)})$: for any $L > 0$,

$$\widehat{\eta}_L^{\mathsf{CV}} \in \arg\min_{\eta\in\Xi_L} \bigg\{\frac{1}{k}\sum_{\ell\in[k]} \frac{1}{m_\ell}\|Y^{(\ell)} - X^{(\ell)}\widehat{\mu}_\eta^{(\ell)}\|^2\bigg\} \equiv \arg\min_{\eta\in\Xi_L} R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta). \tag{4.5}$$

We shall often omit the subscript L in $\widehat{\eta}_L^{\mathsf{CV}}$. Intuitively, due to the independence between $\widehat{\mu}_\eta^{(\ell)}$ and $(X^{(\ell)}, Y^{(\ell)})$, $R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta)$ can be viewed as an estimator of the generalization error $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) + \sigma_\xi^2$. So it is natural to expect that $\widehat{\eta}^{\mathsf{CV}}$ approximately minimizes $\eta \mapsto R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$. Based on the same heuristics as for $\widehat{\eta}^{\mathsf{GCV}}$ in (4.3), we may therefore expect that $\widehat{\eta}^{\mathsf{CV}}$ in (4.5) simultaneously provides optimal prediction, estimation and in-sample risks for 'most' signal μ_0's. This is the content of the following theorem.

Theorem 4.3. *Suppose the same conditions as in Theorem 4.2 and* $\max_{\ell\in[k]} m_\ell/n \leq 1/(2K)$ *hold for some* $K > 0$*. Fix* $\delta \in (0, 1/2]$, $L \geq K/\delta^2$ *and a small enough* $\vartheta \in (0, 1/50)$*. Further assume* $\min_{\ell\in[k]} m_\ell \geq \log^{2/\delta} m$*. There exist a constant* $C = C(K, L, \delta, \vartheta) > 0$ *and a measurable set* $\mathcal{U}_{\delta,\vartheta} \subset B_n(1) \setminus B_n(\delta)$ *with* $\mathrm{vol}(\mathcal{U}_{\delta,\vartheta})/\mathrm{vol}(B_n(1) \setminus B_n(\delta)) \geq 1 - Ce^{-n^\vartheta/C}$*, such that for* $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$,

$$\sup_{\mu_0\in\mathcal{U}_{\delta,\vartheta}} \mathbb{P}\bigg(R^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}_L^{\mathsf{CV}}) \geq \min_{\eta\in\Xi_L} R^{\#}_{(\Sigma,\mu_0)}(\eta) + C\cdot\bigg\{\frac{1}{k}\sum_{\ell\in[k]}\frac{1}{m_\ell^{(1-\delta)/2}} + \frac{1}{k} + n^{-\vartheta}\bigg\}\bigg)$$
$$\leq C(1 + \mathcal{L}_{\{m_\ell\}})\cdot n^{-1/7}.$$

Here $\mathcal{L}_{\{m_\ell\}} \equiv \sum_{\ell\in[k]}(m_\ell/m)^{-1}$.

Non-asymptotic results of this type for k-fold cross validation are previously obtained for $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{CV}})$ in the Lasso setting [MM21, Proposition 4.3] under isotropic

$\Sigma = I_n$, where the range of the regularization must be strictly away from the interpolation regime. In contrast, our results above are valid down to $\eta = 0$ when $\phi^{-1} > 1$, and allow for general anisotropic Σ.

Interestingly, the error bound in the above theorem reflects the folklore tension between the bias and variance in the selection of k in the cross validation scheme (cf. [JWHT21, Chapter 5]):

- For a small number of k, $R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta)$ is biased for estimating $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$; this corresponds to the term $O(1/k)$ in the error bound, which is known to be of the optimal order in Ridge regression (cf. [LD19]).
- For a large number of k, $R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta)$ has large fluctuations; this corresponds to the term $O(k^{-1}\sum_{\ell\in[k]} m_\ell^{-(1-\delta)/2}) = O((k/m)^{(1-\delta)/2})$ in the equal-sized case. By a central limit heuristic (cf. [AZ20, KL22]), we also expect this term to be of a near optimal order.

4.3. **Implications to statistical inference via $\widehat{\mu}_\eta$.** As Ridge(less) estimators $\widehat{\mu}_\eta$ are in general biased, debiasing is necessary for statistical inference of μ_0, cf. [BZ23]. Here the debiasing scheme for $\widehat{\mu}_\eta$ can be readily read off from the distributional characterizations in Theorems 2.3 and 2.4. Assuming known covariance Σ, let the debiased Ridge(less) estimator be defined as

$$\widehat{\mu}^{\mathsf{dR}}_\eta \equiv (\Sigma + \tau_{\eta,*} I)\Sigma^{-1}\widehat{\mu}_\eta. \tag{4.6}$$

Similar to Remark 4, $\tau_{\eta,*}$ and $\widehat{\tau}_\eta$ is interchangeable in the above display due to known Σ. Using Theorems 2.3 and 2.4, we expect that $\widehat{\mu}^{\mathsf{dR}}_\eta \overset{d}{\approx} \mu_0 + \gamma_{\eta,*}\Sigma^{-1/2} g/\sqrt{n}$. This motivates the following confidence intervals for $\{\mu_{0,j}\}$:

$$\mathrm{CI}_j(\eta) \equiv \Big[\widehat{\mu}^{\mathsf{dR}}_{\eta,j} \pm \widehat{\gamma}_\eta \cdot (\Sigma^{-1})^{1/2}_{jj} \cdot \frac{z_{\alpha/2}}{\sqrt{n}}\Big], \quad j \in [n]. \tag{4.7}$$

Here z_α is the normal upper-α quantile defined via $\mathbb{P}(\mathcal{N}(0,1) > z_\alpha) = \alpha$. It is easy to see from the above definition that minimization of $\eta \mapsto \widehat{\gamma}_\eta$ is equivalent to that of the CI length. As the former minimization procedure corresponds exactly to the proposal $\widehat{\eta}^{\mathsf{GCV}}$ in (4.3), we expect that $\{\mathrm{CI}_j(\widehat{\eta}^{\mathsf{GCV}})\}$ provide the shortest (asymptotic) $(1-\alpha)$-CIs along the regularization path, and so do $\{\mathrm{CI}_j(\widehat{\eta}^{\mathsf{CV}})\}$.

Below we give a rigorous statement on the above informal discussion. Let $\mathscr{C}^{\mathsf{dR}}(\eta) \equiv n^{-1}\sum_{j=1}^n \mathbf{1}(\mu_{0,j} \in \mathrm{CI}_j(\eta))$ denote the averaged coverage of $\{\mathrm{CI}_j(\eta)\}$ for $\{\mu_{0,j}\}$. We have the following.

Theorem 4.4. *Suppose the same conditions as in Theorem 4.2 (resp. Theorem 4.3) for $\widehat{\eta}^{\mathsf{GCV}}$ (resp. $\widehat{\eta}^{\mathsf{CV}}$) hold for some $K > 0$. Fix $\alpha \in (0, 1/4]$, $\delta \in (0, 1/2]$, $L \geq K/\delta^2$ and a small enough $\vartheta \in (0, 1/50)$. There exist a constant $C = C(K, L, \delta, \vartheta) > 0$ and a measurable set $\mathcal{U}_{\delta,\vartheta} \subset B_n(1) \setminus B_n(\delta)$ with $\mathrm{vol}(\mathcal{U}_{\delta,\vartheta})/\mathrm{vol}(B_n(1) \setminus B_n(\delta)) \geq 1 - Ce^{-n^\vartheta/C}$, such that the CI length and the averaged coverage satisfy*

$$\sup_{\mu_0\in\mathcal{U}_{\delta,\vartheta}} \Big\{\mathbb{P}\Big(\sqrt{n} z^{-1}_{\alpha/2} \cdot \max_{j\in[n]} \big||\mathrm{CI}_j(\widehat{\eta}^{\#}_L)| - \min_{\eta\in\Xi_L} |\mathrm{CI}_j(\eta)|\big| \geq C\mathcal{E}^{\#}_n\Big)$$
$$\vee\, \mathbb{P}\Big(|\mathscr{C}^{\mathsf{dR}}(\widehat{\eta}^{\#}_L) - (1-\alpha)| \geq C(\mathcal{E}^{\#}_n)^{1/4}\Big)\Big\} \leq C\mathfrak{p}^{\#}_n.$$

Here for $\# \in \{\mathsf{GCV}, \mathsf{CV}\}$*, the quantities* $\mathcal{E}_n^{\#}, \mathfrak{p}_n^{\#}$ *are defined via*

	$\mathcal{E}_n^{\#}$	$\mathfrak{p}_n^{\#}$
$\# = \mathsf{GCV}$	$n^{-\vartheta}$	$n^{-1/7}$
$\# = \mathsf{CV}$	$k^{-1}\sum_{\ell\in[k]} m_\ell^{-(1-\delta)/2} + k^{-1} + n^{-\vartheta}$	$(1+\mathcal{L}_{\{m_\ell\}})\cdot n^{-1/7}$

An interesting and somewhat non-standard special case of the above theorem is the noiseless setting $\sigma_\xi^2 = 0$ in the overparametrized regime $\phi^{-1} > 1$. In this case, exact recovery of μ_0 is impossible and our CI's above provide a precise scheme for partial recovery of μ_0. Moreover, as the effective noise $\phi\gamma_{\eta,*}^2(0) = \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$, Theorem 3.2 and Proposition 3.3 suggest that $\eta \mapsto \gamma_{\eta,*}^2(0)$ is approximately minimized at $\eta = 0$ for 'most' μ_0's. This means that, in this noiseless case, the length of the adaptively tuned CIs is also approximately minimized at the interpolation regime for 'most' μ_0's.

Remark 5. The debiased Ridge estimator $\widehat{\mu}_\eta^{\mathsf{dR}}$ in (4.6) takes a slightly non-standard form, which allows for interpolation $\eta = 0$ when $\phi^{-1} > 1$. To see its equivalence to the standard debiased form, for any $\eta > 0$, using the KKT condition $\widehat{\mu}_\eta = X^\top(Y - X\widehat{\mu}_\eta)/(n\eta)$ and the calculation in (2.4),

$$\widehat{\mu}_\eta^{\mathsf{dR}} = \widehat{\mu}_\eta + \frac{\tau_{\eta,*}}{n\eta}\cdot\Sigma^{-1}\widehat{\mu}_\eta \overset{\mathbb{P}}{\approx} \widehat{\mu}_\eta + \frac{\widehat{\tau}_\eta}{n\eta}\cdot\Sigma^{-1}X^\top(Y - X\widehat{\mu}_\eta) = \widehat{\mu}_\eta + \frac{\Sigma^{-1}X^\top(Y - X\widehat{\mu}_\eta)}{m - \mathrm{df}(\widehat{\mu}_\eta)}.$$

The form in the right hand side of the above display matches the standard form of the debiased Ridge estimator, cf. [BZ23, Eqn. (3.15)].

5. Some illustrative simulations

In this section, we provide some illustrative numerical simulations to validate some of the developed theoretical results. In particular, we focus on:

(1) the phase transitions on the optimality of interpolation in Section 3;
(2) the effectiveness of cross-validation methods in Section 4.

5.1. **Common numerical settings.** We set $\Sigma = 1.99\cdot I_n + 0.01\cdot \mathsf{1}_n\mathsf{1}_n^\top$, with $\mathsf{1}_n$ representing an n-dimensional all one vector. The random design matrix Z and the error ξ are both generated by t-distribution with 10 degrees of freedom, scaled by $\sqrt{0.8}$. This scaling choice ensures that Z_{ij} and ξ_i have mean zero and variance one. The concrete choice of the signal dimension n, the sample size m, and μ_0 will be specified later.

5.2. **Validation of the phase transitions in Theorem 3.4.** First, we use $m = 100$, $n = 200$, and a unit μ_0 chosen randomly (but fixed afterwards) from the sphere $\partial B_n(1)$, and we plot both the theoretical risk curve $\eta \mapsto \bar{R}^{\#}_{(\Sigma,\mu_0)}$ and the empirical risk curve $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}$ for all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$. The left panel of Figure 1 reports the outcome of this simulation with noise level $\sigma_\xi^2 = 1$ and $\mathsf{SNR}_{\mu_0}^{-1} = 1$, while the middle panel of the same figure reports the noiseless case $\sigma_\xi^2 = 0$ with $\mathsf{SNR}_{\mu_0}^{-1} = 0$. These plots show excellent agreements with the theory in Theorem 3.4 in that

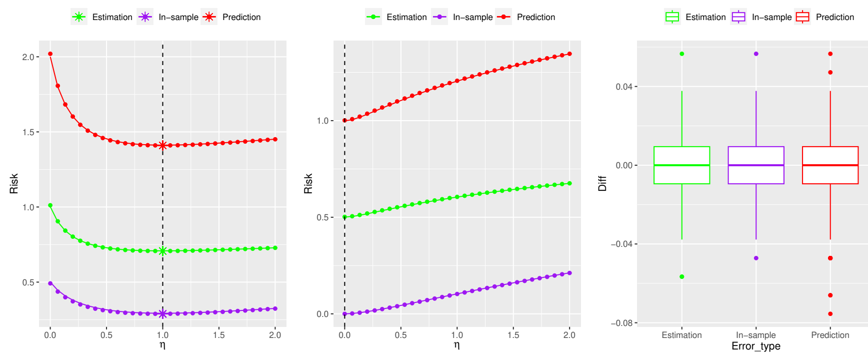

FIGURE 1. Validation of the phase transitions in Theorem 3.4. The theoretical risks $\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ are computed by solving (2.1); the empirical risks $R^{\#}_{(\Sigma,\mu_0)}(\eta)$ are computed via the Monte Carlo simulation over 200 repetitions. *Left panel*: Noisy case with minimal empirical risks attained at $\eta_* = \mathsf{SNR}^{-1}_{\mu_0} = 1$ (marked with $*$). *Middle panel*: Noiseless case with all risks minimized at the interpolation regime $\eta_* = \mathsf{SNR}^{-1}_{\mu_0} = 0$. *Right panel*: Differences of the global minimizer of the risk curves and the oracle η_* are concentrated around 0 over 500 different μ_0's.

the global minimum of both the theoretical and empirical risk curves are attained roughly at $\eta_* = \mathsf{SNR}^{-1}_{\mu_0}$.

In order to demonstrate the validity of the aforementioned phenomenon for 'most' μ_0's as claimed in Theorem 3.4, we uniformly generate 500 different μ_0 over $\partial B_n(1)$. Next, we discretize $\eta \in [0, 1.5]$ into 160 grid points. We then select the optimal value $\eta^{\#}$ by minimizing the empirical prediction, estimation, and in-sample risks. The difference between the chosen empirical optimal $\eta^{\#}$ and the theoretically optimal tuning is depicted in the right panel of Figure 1 through a boxplot of $\eta^{\#} - \eta_*$. It is easily seen that the differences for all three risks are highly concentrated around 0.

5.3. **Optimality of (generalized) cross-validation schemes.** Next, we investigate the efficacy of two cross validation schemes in Section 4, namely $\widehat{\eta}^{\mathsf{GCV}}$ in (4.3) and $\widehat{\eta}^{\mathsf{CV}}$ in (4.5). We keep the sample size fixed at $m = 500$, and allow the signal dimension n to vary so that the aspect ratio $\phi = m/n$ ranges from $[0.5, 1.5]$. To facilitate the tuning process, we employ 31 equidistant η's within the range of $[0, 1.5]$. Moreover, the k-fold cross validation scheme $\widehat{\eta}^{\mathsf{CV}}$ is carried out with the default choice $k = 5$.

To empirically verify Theorem 4.2 and 4.3, we report in the left panel of Figure 2 the empirical risks $R^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}), R^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{CV}})$ for all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$. All the empirical risk curves are found to concentrate around their theoretical optimal counterparts $\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta_*)$. We note again that as $\widehat{\eta}^{\mathsf{GCV}}$ and $\widehat{\eta}^{\mathsf{CV}}$ are designed to tune the prediction risk, it is not surprising that $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}), R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{CV}})$ concentrate

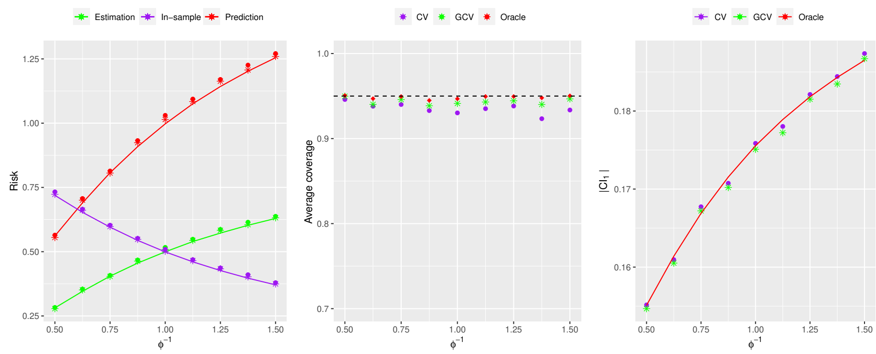

FIGURE 2. Validation of $\mathsf{GCV}, \mathsf{CV}$ in Theorems 4.2-4.4. The empirical risks are computed via the Monte Carlo average over 100 repetitions. *Left panel*: Comparison between empirical risks and theoretical risks for $* = \mathsf{GCV}$ and $\bullet = \mathsf{CV}$ with $k = 5$. *Middle panel*: Averaged coverage $\mathscr{C}^{\mathsf{dR}}(\widehat{\eta}^{\#})$ for $\# \in \{\mathsf{GCV}, \mathsf{CV}\}$ and the oracle $\mathscr{C}^{\mathsf{dR}}(\eta_*)$. *Right panel*: Length of the confidence intervals $\mathrm{CI}_1(\widehat{\eta}^{\#})$ for $\# \in \{\mathsf{GCV}, \mathsf{CV}\}$ and the oracle $\mathrm{CI}_1(\eta_*)$.

around $\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta_*)$. The major surprise appears to be that $\widehat{\eta}^{\mathsf{GCV}}$ and $\widehat{\eta}^{\mathsf{CV}}$ also provide optimal tuning for estimation and in-sample risks, both theoretically validated in our Theorems 4.2 and 4.3 and empirically confirmed here.

To empirically verify Theorem 4.4, we report in the middle and right panels of Figure 2 the averaged coverage and length for the 95%-debiased Ridge CI's with cross-validation, namely $\{\mathrm{CI}_j(\widehat{\eta}^{\#})\}$ for $\# \in \{\mathsf{GCV}, \mathsf{CV}\}$, and with oracle tuning $\eta_* = \mathsf{SNR}_{\mu_0}^{-1}$. For the middle panel, we observe that adaptive tuning via $\widehat{\eta}^{\mathsf{GCV}}$ and $\widehat{\eta}^{\mathsf{CV}}$ both provide approximate nominal coverage for a moderate sample size m and signal dimension n. For the right panel, as the lengths of $\{\mathrm{CI}_j(\widehat{\eta}^{\#})\}$ are solely determined by $\widehat{\gamma}_{\widehat{\eta}^{\#}}$, we report here only the length of $\mathrm{CI}_1(\widehat{\eta}^{\#})$. We observe that the CI length for both $\mathrm{CI}_1(\widehat{\eta}^{\mathsf{GCV}}), \mathrm{CI}_1(\widehat{\eta}^{\mathsf{CV}})$ are also in excellent agreement to the oracle length across different aspect ratios.

6. PROOF OUTLINES

6.1. **Technical tools.**

The main technical tool we use for the proof of Theorem 2.3 is the following version of convex Gaussian min-max theorem, taken from [MM21, Corollary G.1].

Theorem 6.1 (Convex Gaussian Min-Max Theorem)**.** *Suppose $D_u \in \mathbb{R}^{n_1+n_2}, D_v \in \mathbb{R}^{m_1+m_2}$ are compact sets, and $Q : D_u \times D_v \to \mathbb{R}$ is continuous. Let $G = (G_{ij})_{i\in[n_1], j\in[m_1]}$ with G_{ij}'s i.i.d. $\mathcal{N}(0,1)$, and $g \sim \mathcal{N}(0, I_{n_1})$, $h \sim \mathcal{N}(0, I_{m_1})$ be independent Gaussian vectors. For $u \in \mathbb{R}^{n_1+n_2}, v \in \mathbb{R}^{m_1+m_2}$, write $u_1 \equiv u_{[n_1]} \in \mathbb{R}^{n_1}, v_1 \equiv v_{[m_1]} \in \mathbb{R}^{m_1}$. Define*

$$\Phi^p(G) = \min_{u\in D_u} \max_{v\in D_v} \Big(u_1^\top G v_1 + Q(u, v)\Big),$$

$$\Phi^a(g,h) = \min_{u\in D_u}\max_{v\in D_v}\Big(\|v_1\|g^\top u_1 + \|u_1\|h^\top v_1 + Q(u,v)\Big).$$

Then the following hold.

(1) *For all $t\in\mathbb{R}$, $\mathbb{P}(\Phi^p(G)\le t)\le 2\,\mathbb{P}(\Phi^a(g,h)\le t)$.*

(2) *If $(u,v)\mapsto u_1^\top G v_1 + Q(u,v)$ satisfies the conditions of Sion's min-max theorem for the pair (D_u,D_v) a.s. (for instance, D_u,D_v are convex, and Q is convex-concave), then for any $t\in\mathbb{R}$, $\mathbb{P}(\Phi^p(G)\ge t)\le 2\,\mathbb{P}(\Phi^a(g,h)\ge t)$.*

Clearly, $\ge$ (resp. $\le$) in (1) (resp. (2)) can be replaced with $>$ (resp $<$). In the proofs below, we shall assume without loss of generality that G,g,h are independent Gaussian matrix/vectors defined on the same probability space.

As mentioned above, the CGMT above has been utilized for deriving precise risk/distributional asymptotics for a number of canonical statistical estimators across various important models; we only refer the readers to [TOH15, TAH18, SAH19, LGC+21, CMW22, DKT22, Han22, LS22, WWM22, ZZY22, MRSY23] for some selected references.

6.2. **Reparametrization and further notation.** Consider the reparametrization

$$w = \Sigma^{1/2}(\mu-\mu_0),\quad \widehat{w}_{\eta;Z}\equiv \Sigma^{1/2}(\widehat{\mu}_{\eta;Z}-\mu_0).$$

Then with

$$F(w)\equiv F_{(\Sigma,\mu_0)}(w) = \frac{1}{2}\|\mu_0+\Sigma^{-1/2}w\|^2, \tag{6.1}$$

we have the following reparametrized version of $\widehat{\mu}_{\eta;Z}$:

$$\widehat{w}_{\eta;Z} = \begin{cases}\arg\min_{w\in\mathbb{R}^n}\{F(w): Zw=\xi\}, & \eta=0;\\ \arg\min_{w\in\mathbb{R}^n}\{F(w)+\frac{1}{\eta}\cdot\frac{1}{2n}\|Zw-\xi\|^2\}, & \eta>0.\end{cases}$$

Next we give some further notation for cost functions. Let for $\eta\ge 0$,

$$\begin{aligned} h_{\eta;Z}(w,v) &\equiv \frac{1}{\sqrt{n}}\langle v, Zw-\xi\rangle + F(w) - \frac{\eta\|v\|^2}{2},\\ \ell_\eta(w,v) &\equiv \frac{1}{\sqrt{n}}\Big(-\|v\|\langle g,w\rangle + \|w\|\langle h,v\rangle - \langle v,\xi\rangle\Big) + F(w) - \frac{\eta\|v\|^2}{2}, \end{aligned}\tag{6.2}$$

and for $L_v\in[0,\infty]$,

$$H_{\eta;Z}(w;L_v)\equiv \max_{v\in B_n(L_v)} h_{\eta;Z}(w,v)\equiv \max_{v\in B_n(L_v)}\Big\{\frac{\langle v, Zw-\xi\rangle}{\sqrt{n}} + F(w) - \frac{\eta\|v\|^2}{2}\Big\}, \tag{6.3}$$

$$L_\eta(w;L_v)\equiv \max_{v\in B_n(L_v)}\ell_\eta(w,v) = \max_{\beta\in[0,L_v]}\Big\{\frac{\beta}{\sqrt{n}}\Big(\big\|\|w\|h-\xi\big\| - \langle g,w\rangle\Big) + F(w) - \frac{\eta\beta^2}{2}\Big\}.$$

We shall simply write $H_{\eta;Z}(\cdot) = H_{\eta;Z}(\cdot;\infty)$ and $L_\eta(\cdot)=L_\eta(\cdot;\infty)$. When $Z=G$, we sometimes write $h_{\eta;G}=h_\eta$ and $H_{\eta;G}=H_\eta$ for simplicity of notation.

Let the empirical noise σ_m^2 and its modified version be

$$\sigma_m^2\equiv\frac{\|\xi\|^2}{\|h\|^2},\quad \sigma_\pm^2(L_w)\equiv\Big(\sigma_m^2\pm 2L_w\frac{|\langle h,\xi\rangle|}{\|h\|^2}\Big)_+. \tag{6.4}$$

Finally we define $\mathsf{D}_{\eta,\pm}$ and its deterministic version $\overline{\mathsf{D}}_\eta$ as follows:

$$\mathsf{D}_{\eta,\pm}(\beta,\gamma) \equiv \frac{\beta}{2}\Big(\gamma(\phi e_h^2 - e_g^2) + \frac{\sigma_\pm^2}{\gamma}\Big) - \frac{\eta\beta^2}{2} + \mathsf{e}_F\Big(\frac{\gamma}{\sqrt{n}}g;\frac{\gamma}{\beta}\Big),$$

$$\overline{\mathsf{D}}_\eta(\beta,\gamma) = \frac{\beta}{2}\Big(\gamma(\phi-1) + \frac{\sigma_\xi^2}{\gamma}\Big) - \frac{\eta\beta^2}{2} + \mathbb{E}\,\mathsf{e}_F\Big(\frac{\gamma}{\sqrt{n}}g;\frac{\gamma}{\beta}\Big). \tag{6.5}$$

Here recall e_F is the Moreau envelope of F in (6.1). Note that $\mathsf{D}_{\eta,\pm}$ depends on the choice of L_w, but for notational convenience we drop this dependence here.

6.3. **Proof outline for Theorem 2.3 for $\eta = 0$.** We shall outline below the main steps for the proof of Theorem 2.3 for $\eta = 0$ in the regime $\phi^{-1} > 1$ under a stronger condition $\|\Sigma^{-1}\|_{\mathrm{op}} \lesssim 1$. The high level strategy of the proof shares conceptual similarities to [MM21, CMW22], but the details differ significantly.

(***Step 1**: Localization of the primal optimization*). In this step, we show that for $L_w, L_v > 0$ such that $L_w \wedge L_v \gtrsim 1$, with high probability (w.h.p.),

$$\min_{w\in B_n(L_w)} H_0(w;L_v) = \min_{w\in\mathbb{R}^n} H_0(w). \tag{6.6}$$

A formal statement of the above localization can be found in Proposition 9.1. The key point here is that despite $\min_w H_0(w)$ optimizes a deterministic function with a random constraint, it can be efficiently rewritten (in a probabilistic sense) in a minimax form indexed by *compact sets* that facilitate the application of the convex Gaussian min-max Theorem 6.1.

(***Step 2**: Characterization of the Gordon cost optimum*). In this step, we show that a suitably localized version of $\min_w L_0(w)$ concentrates around some *deterministic* quantity involving the function $\overline{\mathsf{D}}_0$ in (6.5). In particular, we show in Theorem 9.2 that for $L_w, L_v \asymp 1$ chosen large enough, w.h.p.,

$$\min_{w\in B_n(L_w)} L_0(w;L_v) \approx \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_0(\beta,\gamma). \tag{6.7}$$

The proof of (6.7) is fairly involved, as the minimax problem $\min_w L_0(w) = \min_w \max_v \ell_0(w,v)$ (and its suitably localized versions) cannot be computed exactly. We get around this technical issue by the following bracketing strategy:

- (***Step 2.1***). We show in Proposition 9.3 that for the prescribed choice of L_w, L_v, w.h.p., both

  $$\max_{\beta>0}\min_{\gamma>0} \mathsf{D}_{0,-}(\beta,\gamma) \le \min_{w\in B_n(L_w)} L_0(w;L_v) \le \max_{\beta>0}\min_{\gamma>0} \mathsf{D}_{0,+}(\beta,\gamma),$$

  and the localization

  $$\max_{\beta>0}\min_{\gamma>0} \mathsf{D}_{0,\pm}(\beta,\gamma) = \max_{1/C\le\beta\le C}\ \min_{1/C\le\gamma\le C} \mathsf{D}_{0,\pm}(\beta,\gamma)$$

  hold for some large $C > 0$.
- (***Step 2.2***). We show in Proposition 9.4 that for localized minimax problems, we may replace $\mathsf{D}_{0,\pm}$ by $\overline{\mathsf{D}}_{0,\pm}$: w.h.p.,

  $$\max_{1/C\le\beta\le C}\ \min_{1/C\le\gamma\le C} \mathsf{D}_{0,\pm}(\beta,\gamma) \approx \max_{1/C\le\beta\le C}\ \min_{1/C\le\gamma\le C} \overline{\mathsf{D}}_0(\beta,\gamma).$$

- (***Step 2.3***). We show in Proposition 9.5 that (de)localization holds for the (deterministic) max-min optimization problem with $\overline{\mathsf{D}}_0$:

$$\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_0(\beta,\gamma)=\max_{1/C\leq\beta\leq C}\ \min_{1/C\leq\gamma\leq C}\overline{\mathsf{D}}_0(\beta,\gamma).$$

Combining the above Steps 2.1-2.3 yields (6.7). An important step to prove the (de)localization claims above is to derive apriori estimates for the solutions of the fixed point equation (2.1) and its sample version, to be defined in (8.12). These estimates will be detailed in Section 8.

(***Step 3****: Locating the global minimizer of the Gordon objective*). In this step, we show that a suitably localized version of the Gordon objective $w\mapsto L_0(w)$ attains its global minimum approximately at $w_{0,*}\equiv\Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{0,*};\tau_{0,*})-\mu_0)$ in the following sense. For any $\varepsilon>0$ and any $g:\mathbb{R}^n\to\mathbb{R}$ that is 1-Lipschitz with respect to $\|\cdot\|_{\Sigma^{-1}}$, let $D_{0;\varepsilon}(\mathsf{g})\equiv\{w\in\mathbb{R}^n:|\mathsf{g}(w)-\mathbb{E}\,\mathsf{g}(w_{0,*})|\geq\varepsilon\}$ be the 'exceptional set'. We show in Theorem 9.6 that again for $L_w,L_v\asymp 1$ chosen large enough, w.h.p.,

$$\min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_n(L_w)}L_0(w;L_v)\geq\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_0(\beta,\gamma)+\Omega_\varepsilon(1).\tag{6.8}$$

The main challenge in proving (6.8) is partly attributed to the possible violation of strong convexity of the map $w\mapsto L_0(w;L_v)$, due to the necessity of working with non-Gaussian ξ's. We will get around this technical issue in similar spirit to Step 2 by another bracketing strategy. In particular:

- (***Step 3.1***). In Lemma 9.7, we will use surrogate, strongly convex functions $L_{0,\pm}(\cdot;L_v)$, formally defined in (9.16), to provide a sufficiently tight bracket for $L_0(\cdot;L_v)$ over large enough compact sets.
- (***Step 3.2***). In Proposition 9.8, we show that the minimizers of $w\mapsto L_{0,\pm}(\cdot;L_v)$ can be computed exactly and are close enough to $w_{0,*}$.
- (***Step 3.3***). In Proposition 9.9, combined with the tight bracketing and certain apriori estimates, we then conclude that all minimizers of $w\mapsto L_0(\cdot;L_v)$ must be close to $w_{0,*}$.

With all the above steps, finally we prove (6.8) by (i) using the proximity of L_0 and its surrogate $L_{0,\pm}$ and (ii) exploiting the strong convexity of $L_{0,\pm}$.

(***Step 4****: Putting pieces together and establishing uniform guarantees*). In this final step, we shall use the convex Gaussian min-max theorem to translate the estimates (6.7) in Step 2 and (6.8) in Step 3 to their counterparts with primal cost function H_0. For the global cost optimum, with the help of the localization in (6.6), by choosing $L_w,L_v\asymp 1$, we have w.h.p.,

$$\min_{w\in\mathbb{R}^n}H_0(w)\stackrel{(6.6)}{=}\min_{w\in B_n(L_w)}H_0(w;L_v)\stackrel{\mathbb{P}}{\approx}\min_{w\in B_n(L_w)}L_0(w;L_v)\stackrel{(6.7)}{\approx}\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_0(\beta,\gamma).$$

For the cost over the exceptional set, we have w.h.p.,

$$\begin{aligned}\min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_n(L_w)}H_0(w)&\geq\min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_n(L_w)}H_0(w;L_v)\\&\stackrel{\mathbb{P}}{\geq}\min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_n(L_w)}L_0(w;L_v)\stackrel{(6.8)}{\geq}\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_0(\beta,\gamma)+\Omega_\varepsilon(1).\end{aligned}$$

Combining the above two displays, we then conclude that w.h.p., $\widehat{w}_0 \notin D_{0;\varepsilon}(\mathsf{g}) \cap B_n(L_w)$. Finally using apriori estimate on $\|\widehat{w}_0\|$ we may conclude that w.h.p., $\widehat{w}_0 \notin D_{0;\varepsilon}(\mathsf{g})$, i.e., $|\mathsf{g}(\widehat{w}_0) - \mathbb{E}\,\mathsf{g}(w_{0,*})| \leq \varepsilon$.

The uniform guarantee in η is then proved by (i) extending the above arguments to include any positive $\eta > 0$, and (ii) establishing (high probability) Lipschitz continuity (w.r.t. $\|\cdot\|_{\Sigma^{-1}}$) of the maps $\eta \mapsto \widehat{w}_\eta$ and $\eta \mapsto w_{\eta,*}$.

Details of the above outline are implemented in Section 9.

6.4. **Proof outline for Theorem 2.4 for $\eta = 0$.** The main tool we will use to prove the universality Theorem 2.4 is the following set of comparison inequalities developed in [HS22]: Suppose Z matches the first two moments of G, and possesses enough high moments. Then for any measurable sets $\mathcal{S}_w \subset [-L_n/\sqrt{n}, L_n/\sqrt{n}]^n, \mathcal{S}_v \subset [-L_n/\sqrt{n}, L_n/\sqrt{n}]^m$, and any smooth test function $\mathsf{T} : \mathbb{R} \to \mathbb{R}$ (standardized with derivatives of order 1 in $\|\cdot\|_\infty$),

$$\Big|\mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_w}\max_{v\in\mathcal{S}_v} h_{\eta;Z}(w,v)\Big) - \mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_w}\max_{v\in\mathcal{S}_v} h_{\eta;G}(w,v)\Big)\Big| \leq \mathsf{r}_n(L_n),$$

$$\Big|\mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_w} H_{\eta;Z}(w)\Big) - \mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_w} H_{\eta;G}(w)\Big)\Big| \leq \mathsf{r}_n(L_n). \tag{6.9}$$

Here $\mathsf{r}_n(L_n) \to 0$ for $L_n = n^{\vartheta}$ with sufficiently small $\vartheta > 0$. The readers are referred to Theorems 10.1 and 10.2 for a precise statement of (6.9).

An important technical subtlety here is that while the first inequality in (6.9) holds down to $\eta = 0$, the second inequality does not. This is so because $\min_w H_{0;Z}(w)$, which minimizes a deterministic function under a random constraint due to the unbounded constraint in the maximization of v, is qualitatively different from $\min_w H_{\eta;Z}(w)$ for any $\eta > 0$.

Now we shall sketch how the comparison inequalities (6.9) lead to universality.

(***Step 1****: Universality of the global cost optimum*). In this step, we shall use the first inequality in (6.9) to establish the universality of the global Gordon cost:

$$\min_{w\in\mathbb{R}^n} H_{0;Z}(w) = \min_{w\in\mathbb{R}^n}\max_{v\in\mathbb{R}^m} h_{0;Z}(w,v) \overset{\mathbb{P}}{\approx} \min_{w\in\mathbb{R}^n}\max_{v\in\mathbb{R}^m} h_{0;G}(w,v). \tag{6.10}$$

See Theorem 10.4 for a formal statement of (6.10).

The crux to establish (6.10) via the first inequality of (6.9) is to show that, the ranges of the minimum and the maximum of $\min_w \max_v h_{0;Z}(w,v)$ can be localized into an L_∞ ball of order close to $\mathcal{O}(1/\sqrt{n})$. This amounts to showing that the stationary points $(\widehat{w}_{0;Z}, \widehat{v}_{0;Z})$, where $\widehat{w}_{\eta;Z} = \Sigma^{1/2}(\widehat{\mu}_{\eta;Z} - \mu_0)$ and $\widehat{v}_{\eta;Z} = -n^{-1/2}(XX^\top/n + \eta I_m)^{-1}Y$ (cf. Eqn. (10.3)), are delocalized. We prove such delocalization properties in Proposition 10.3 for 'most' $\mu_0 \in B_n(1)$.

(***Step 2****: Universality of the cost over exceptional sets*). In this step, we shall use the second inequality in (6.9) to establish the universality of the Gordon cost over exceptional sets $D_{0;\varepsilon}(\mathsf{g})$. In particular, we show in Theorem 10.5 that with $L_n = Cn^{\vartheta}$ for sufficiently small $\vartheta > 0$ and a large enough $C_0 > 0$, w.h.p.,

$$\min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{0;Z}(w) \geq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_0(\beta,\gamma) + \Omega_\varepsilon(1). \tag{6.11}$$

Here $B_{(2,\infty)}(C_0, L_n/\sqrt{n}) = B_n(C_0) \cap L_\infty(L_n/\sqrt{n})$. As mentioned above, a technical difficulty to apply the second inequality of (6.9) rests in its singular behavior near the interpolation regime $\eta = 0$. Also, we note that for a general exceptional set $D_{0;\varepsilon}(\mathsf{g})$, the maximum over v in $\min_{w\in D_{0;\varepsilon}(\mathsf{g})} H_{0;Z}(w) = \min_{w\in D_{0;\varepsilon}(\mathsf{g})} \max_v h_{0;Z}(w,v)$ need not be delocalized, so the first inequality of (6.9) cannot be applied. This singularity issue will be resolved in two steps:

- (***Step 2.1***). First, we use the second inequality of (6.9) to show that, (6.11) is valid for a version with small enough $\eta > 0$:

$$\mathbb{P}\bigg(\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) \geq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + \Omega_\varepsilon(1)\bigg) \geq 1 - c_\eta \cdot \mathfrak{o}(1).$$

  See (10.15) for a precise statement. As expected, c_η blows up as $\eta \downarrow 0$.
- (***Step 2.2***). Next, by using the 'stability' of the set $D_{\eta;\varepsilon}(\mathsf{g})$ (cf. Lemma 10.6) and $\max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma)$ (cf. Eqn. (9.13)) with respect to η, for a small enough $\eta > 0$, we have the following series of inequalities:

$$\begin{aligned}
&\min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{0;Z}(w)\\
&\geq \min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) \quad \text{(by definition of } H_{\eta;Z})\\
&\geq \min_{w\in D_{\eta;\varepsilon_\eta}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) \quad (\varepsilon_\eta \approx \varepsilon \text{ by Lemma 10.6})\\
&\overset{\mathbb{P}}{\geq} \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + \Omega_\varepsilon(1) \quad \text{(by Step 2.1 above)}\\
&\geq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_0(\beta,\gamma) - \mathcal{O}(\eta) + \Omega_\varepsilon(1) \quad \text{(by Eqn. (9.13))}.
\end{aligned}$$

  Now for a given $\varepsilon > 0$, we may choose $\eta > 0$ small enough so that the term $-\mathcal{O}(\eta)$ is absorbed into $\Omega_\varepsilon(1)$, and therefore concluding (6.11).

A complete proof of the above outline is detailed in Section 10.

7. Proof preliminaries

7.1. **Some properties of e_F and prox_F.** We write $g_n \equiv g/\sqrt{n}$ in this subsection. First we give an explicit expression for $\mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau)$ and $\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau)$.

Lemma 7.1. *For any $(\gamma,\tau) \in (0,\infty)^2$,*

$$\begin{aligned}
\mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau) &= \tau^2\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\mu_0\|^2 + \gamma^2\cdot n^{-1}\,\mathrm{tr}\,(\Sigma^2(\Sigma+\tau I)^{-2}),\\
\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau) &= \gamma^2\cdot n^{-1}\,\mathrm{tr}\,(\Sigma(\Sigma+\tau I)^{-1}).
\end{aligned}$$

Proof. Using the closed-form of $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}$, we may compute

$$\Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) - \mu_0) = (\Sigma+\tau I)^{-1}\Sigma^{1/2}(-\tau\mu_0 + \gamma\Sigma^{1/2}g_n). \tag{7.1}$$

The claims follow from direct calculations. $\square$

Next we give explicit expression for $\mathsf{prox}_F(\gamma g_n;\tau)$ and $\mathsf{e}_F(\gamma g_n;\tau)$.

Lemma 7.2. *It holds that*

$$\mathsf{prox}_F(\gamma g_n;\tau) = \Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) - \mu_0),$$

$$\mathsf{e}_F(\gamma g_n;\tau) = \frac{1}{2\tau}\|\Sigma^{1/2}\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) - y^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma)\|^2 + \frac{1}{2}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)\|^2.$$

Furthermore,

$$\begin{aligned}\mathbb{E}\,\mathsf{e}_F(\gamma g_n;\tau) = \frac{1}{2\tau}\big(\,\mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau) - 2\,\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau) + \gamma^2\big)\\ + \frac{1}{2}\Big(\|(\Sigma+\tau I)^{-1}\Sigma\mu_0\|^2 + \gamma^2\cdot\frac{1}{n}\,\mathrm{tr}(\Sigma(\Sigma+\tau I)^{-2})\Big).\end{aligned}$$

Proof. The two identities in the first display follows from the definition of F. For the second display, note that $\mathbb{E}\,\mathsf{e}_F(\gamma g_n;\tau)$ is equal to

$$\frac{1}{2\tau}\big(\,\mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau) - 2\,\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau) + \gamma^2\big) + \frac{1}{2}\,\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)\|^2.$$

Using $\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)\|^2 = \|(\Sigma+\tau I)^{-1}\Sigma\mu_0\|^2 + \gamma^2\cdot n^{-1}\,\mathrm{tr}\,(\Sigma(\Sigma+\tau I)^{-2})$ to conclude. □

The derivative formula below for e_F will be useful.

Lemma 7.3. *It holds that*

$$\nabla_x\mathsf{e}_F(x;\tau) = \frac{1}{\tau}(x - \mathsf{prox}_F(x;\tau)),\quad \partial_\tau\mathsf{e}_F(x;\tau) = -\frac{1}{2\tau^2}\|x - \mathsf{prox}_F(x;\tau)\|^2.$$

Proof. See e.g., [TAH18, Lemmas B.5 and D.1]. □

Finally we provide a concentration inequality for $\mathsf{e}_F(\gamma g_n;\tau)$.

Proposition 7.4. *There exists some universal constant $C > 0$ such that*

$$\mathbb{P}\left(\big|\mathsf{e}_F(\gamma g_n;\tau) - \mathbb{E}\,\mathsf{e}_F(\gamma g_n;\tau)\big| \geq C\Big\{v\,\mathbb{E}^{1/2}\,\mathsf{e}_F(\gamma g_n;\tau)\sqrt{\frac{t}{n}} + v^2\cdot\frac{t}{n}\Big\}\right) \leq Ce^{-t/C}$$

holds for any $t \geq 0$. Here $v^2 \equiv v^2(\gamma,\tau) \equiv \gamma^2(\tau\|(\Sigma+\tau I)^{-1}\|_{\mathrm{op}}^2 + \|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\|_{\mathrm{op}}^2)$.

Proof. Using that $\nabla_g\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) = \frac{\gamma}{\sqrt{n}}(\Sigma+\tau I)^{-1}\Sigma^{1/2}$ and $\nabla_g y^{\mathsf{seq}}(\gamma) = \frac{\gamma}{\sqrt{n}}I$,

$$\begin{aligned}\nabla_g\mathsf{e}_F(\gamma g_n;\tau) = \frac{1}{\tau}\cdot\frac{\gamma}{\sqrt{n}}((\Sigma+\tau I)^{-1}\Sigma - I)(\Sigma^{1/2}\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) - y^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma))\\ + \frac{\gamma}{\sqrt{n}}(\Sigma+\tau I)^{-1}\Sigma^{1/2}\nabla_g\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau).\end{aligned}$$

This means

$$\begin{aligned}&\|\nabla_g\mathsf{e}_F(\gamma g_n;\tau)\|^2\\ &\leq 2\gamma^2\cdot n^{-1}\Big\{\|(\Sigma+\tau I)^{-1}\|_{\mathrm{op}}^2\|\Sigma^{1/2}\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) - y^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma)\|^2\\ &\qquad + \|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\|_{\mathrm{op}}^2\|\nabla_g\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)\|^2\Big\}\\ &\leq 4\gamma^2\cdot n^{-1}\Big(\tau\|(\Sigma+\tau I)^{-1}\|_{\mathrm{op}}^2 + \|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\|_{\mathrm{op}}^2\Big)\cdot\mathsf{e}_F(\gamma g_n;\tau). \qquad (7.2)\end{aligned}$$

From here we may conclude by setting $H(g) \equiv \mathsf{e}_F(\gamma g_n;\tau)$ and $\Gamma^2 \equiv 4\gamma^2 n^{-1}(\tau\|(\Sigma+\tau I)^{-1}\|_{\mathrm{op}}^2 + \|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\|_{\mathrm{op}}^2)$ in Proposition B.1. □

7.2. **Some high probability events.** Let

$$e_h^2 = \|h\|^2/m, \quad e_g^2 \equiv \|g\|^2/n. \tag{7.3}$$

For $M, \delta > 0$, consider the event

$$\begin{aligned}
\mathscr{E}_0(M) &\equiv \{(\|G\|_{\mathrm{op}}/\sqrt{n}) \vee [\|(GG^\top/n)^{-1}\|_{\mathrm{op}}\mathbf{1}_{\eta=0}] \le M\},\\
\mathscr{E}_{1,0}(\delta) &\equiv \{|e_g^2 - 1| \vee |e_h^2 - 1| \vee |n^{-1/2}\langle\Sigma^{1/2}g, \mu_0\rangle| \vee |n^{-1}\langle h, \xi\rangle| \le \delta\},\\
\mathscr{E}_{1,\xi}(\delta) &\equiv \{|(\|\xi\|^2/m) - \sigma_\xi^2| \le \delta\},\\
\mathscr{E}_1(\delta) &\equiv \mathscr{E}_{1,0}(\delta) \cap \mathscr{E}_{1,\xi}(\delta).
\end{aligned}$$

Here in the definition of $\mathscr{E}_0(M)$, we interpret $\infty \cdot 0 = 0$. Typically we think of $M \asymp 1$ and $\delta \asymp 1/\sqrt{n}$.

Lemma 7.5. *Fix $\delta \in (0, 1/2)$ and $L_w > 0$. Then $\mathscr{E}_1(\delta) \subset \mathscr{E}_2(4(\sigma_\xi^2 + 1 + \phi^{-1}L_w)\delta, L_w)$, where $\mathscr{E}_2(\delta, L_w) \equiv \{|\sigma_\pm^2(L_w) - \sigma_\xi^2| \le \delta\}$.*

Proof. Using the definition of $\sigma_\pm^2(L_w)$ in (6.4), on $\mathscr{E}_1(\delta)$, we have

$$\big|\sigma_\pm^2(L_w) - \sigma_\xi^2\big| \le \frac{\|\xi\|^2}{\|h\|^2}|e_h^2 - 1| + \left|\frac{\|\xi\|^2}{m} - \sigma_\xi^2\right| + \frac{2L_w|\langle h, \xi\rangle|}{\|h\|^2} \le 4(\sigma_\xi^2 + 1 + \phi^{-1}L_w)\delta.$$

The claim follows. □

Lemma 7.6. *Suppose $1/K \le \phi^{-1} - \mathbf{1}_{\eta=0} \le K$. Then there exists some $C = C(K) > 0$ such that $\mathbb{P}(\mathscr{E}_0(C)) \ge 1 - Ce^{-n/C}$.*

Proof. The claim for $\|G\|_{\mathrm{op}}/\sqrt{n}$ follows from standard concentration estimates. The claim for $\|(GG^\top/n)^{-1}\|_{\mathrm{op}}$ follows from, e.g., [RV09, Theorem 1.1]. □

Lemma 7.7. *Suppose $1/K \le \phi^{-1} \le K$, and $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \le K$ for some $K > 0$, and Assumption B hold with $\sigma_\xi^2 > 0$. There exists some constant $C = C(K, \sigma_\xi) > 0$ such that for all $t \ge 0$, with $\delta(t, n) \equiv C(\sqrt{t/n} + t/n)$, for $\xi \in \mathscr{E}_{1,\xi}(\delta(t, n))$, we have $\mathbb{P}^\xi(\mathscr{E}_1(\delta(t, n))) \ge 1 - e^{-t}$.*

Proof. The claim follows by standard concentration inequalities. □

8. Properties of the fixed point equations

8.1. **The fixed point equation (2.1).**

Proposition 8.1. *The following hold.*

1. *The fixed point equation (2.1) admits a unique solution $(\gamma_{\eta,*}, \tau_{\eta,*}) \in (0, \infty)^2$, for all $(m, n) \in \mathbb{N}^2$ when $\eta > 0$ and $m < n$ when $\eta = 0$.*
2. *The following apriori bounds hold:*

$$\frac{1 - \phi + \sqrt{(1 - \phi)^2 + 4\mathcal{H}_\Sigma\eta}}{2\mathcal{H}_\Sigma} \le \tau_{\eta,*} \le \inf_{k \in [0:\min\{m-1, n\}]}\left\{\frac{\sum_{j>k}\lambda_j}{m - k} + \frac{n}{m - k}\cdot\eta\right\},$$

$$\frac{\sigma_\xi^2}{\phi} \le \gamma_{\eta,*}^2 \le \frac{\sigma_\xi^2 + \|\Sigma\|_{\mathrm{op}}\|\mu_0\|^2}{\phi}\left(1 + \frac{\|\Sigma\|_{\mathrm{op}}}{\tau_{\eta,*}}\right).$$

(3) If $1/K \leq \phi^{-1} \leq K$ and $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \leq K$ for some $K > 1$, then there exists some $C = C(K) > 1$ such that uniformly in $\eta \in \Xi_K$,

$$1/C \leq \tau_{\eta,*} \leq C, \quad 1/C \leq (-1)^{q+1}\partial_\eta^q \tau_{\eta,*} \leq C, \quad q \in \{1,2\}.$$

If furthermore $1/K \leq \sigma_\xi^2 \leq K$ and $\|\mu_0\| \leq K$, then uniformly in $\eta \in \Xi_K$,

$$1/C \leq \gamma_{\eta,*} \leq C, \quad |\partial_\eta \gamma_{\eta,*}| \leq C.$$

Proof. We shall write $(\gamma_{\eta,*}, \tau_{\eta,*}) = (\gamma_*, \tau_*)$ for notational simplicity. All the constants in $\lesssim, \gtrsim, \asymp$ below may depend on K.

(1). First we prove the existence and uniqueness of τ_*. We rewrite the second equation of (2.1) as

$$\phi = \frac{1}{n}\operatorname{tr}\big((\Sigma + \tau_* I)^{-1}\Sigma\big) + \frac{\eta}{\tau_*} = \frac{1}{n}\sum_{j=1}^n \frac{\lambda_j}{\lambda_j + \tau_*} + \frac{\eta}{\tau_*} \equiv \mathsf{f}(\tau_*). \tag{8.1}$$

Clearly $\mathsf{f}(\tau)$ is smooth, non-increasing, $\mathsf{f}(0) = 1 > \phi$ for $\eta = 0$ and $\mathsf{f}(0) = \infty$ for $\eta > 0$, and $\mathsf{f}(\infty) = 0$, so $\tau \mapsto \mathsf{f}(\tau) - \phi$ must admit a unique zero $\tau_* \in (0, \infty)$.

Next we prove the existence and uniqueness of γ_*. Using Lemma 7.1, the equation $\phi\gamma_*^2 = \sigma_\xi^2 + \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_*; \tau_*)$ reads

$$\phi = \frac{1}{\gamma_*^2}\big(\sigma_\xi^2 + \tau_*^2\|(\Sigma + \tau_* I)^{-1}\Sigma^{1/2}\mu_0\|^2\big) + \frac{1}{n}\operatorname{tr}\big((\Sigma + \tau_* I)^{-2}\Sigma^2\big). \tag{8.2}$$

As $n^{-1}\operatorname{tr}\big((\Sigma + \tau_* I)^{-2}\Sigma^2\big) < n^{-1}\operatorname{tr}\big((\Sigma + \tau_* I)^{-1}\Sigma\big) \leq \phi$ by (8.1) and the fact $\tau_* > 0$, the above equation admits a unique solution $\gamma_* \in (0, \infty)$, analytically given by

$$\gamma_*^2 = \frac{\sigma_\xi^2 + \tau_*^2\|(\Sigma + \tau_* I)^{-1}\Sigma^{1/2}\mu_0\|^2}{\phi - \frac{1}{n}\operatorname{tr}\big((\Sigma + \tau_* I)^{-2}\Sigma^2\big)} = \frac{\sigma_\xi^2 + \tau_*^2\|(\Sigma + \tau_* I)^{-1}\Sigma^{1/2}\mu_0\|^2}{\frac{\eta}{\tau_*} + \frac{\tau_*}{n}\operatorname{tr}\big((\Sigma + \tau_* I)^{-2}\Sigma\big)}. \tag{8.3}$$

(2). For the upper bound for τ_*, using the equation (8.1), we have

$$m = n\phi \leq k + \frac{1}{\tau_*}\sum_{j>k}\lambda_j + \frac{n\eta}{\tau_*}, \quad \forall k \in [0:n],\ k \leq m-1.$$

Solving for τ_* yields the desired upper bound. For the lower bound for τ_*, note that (8.1) leads to

$$\phi = 1 - \tau_* \cdot \frac{1}{n}\sum_{j=1}^n \frac{1}{\lambda_j + \tau_*} + \frac{\eta}{\tau_*} \geq 1 - \tau_*\mathcal{H}_\Sigma + \frac{\eta}{\tau_*},$$

or equivalently $\mathcal{H}_\Sigma\tau_*^2 + (\phi - 1)\tau_* - \eta \geq 0$. Solving this quadratic inequality yields the lower bound for τ_*.

On the other hand, the lower bound $\gamma_*^2 \geq \sigma_\xi^2/\phi$ is trivial by (8.3). For the upper bound for γ_*, using that

$$\begin{aligned}\phi - \frac{1}{n}\operatorname{tr}\big((\Sigma + \tau_* I)^{-2}\Sigma^2\big) &\geq \phi - \frac{1}{n}\operatorname{tr}\big((\Sigma + \tau_* I)^{-1}\Sigma\big) \cdot \max_{j\in[n]}\frac{\lambda_j}{\lambda_j + \tau_*}\\ &\geq \phi \cdot \frac{\tau_*}{\|\Sigma\|_{\mathrm{op}} + \tau_*},\end{aligned} \tag{8.4}$$

and the first identity in (8.3), we have

$$\gamma_*^2 \le \phi^{-1}(\sigma_\xi^2 + \|\Sigma\|_{\mathrm{op}}\|\mu_0\|^2)(1 + \|\Sigma\|_{\mathrm{op}}/\tau_*).$$

Collecting the bounds proves the claim.

(3). The claim on γ_*, τ_* is a simple consequence of (2). We shall prove the other claim on their derivatives. Viewing $\tau_* = \tau_*(\eta)$ and taking derivative with respect to η on both sides of (8.1) yield that, with $T_{-p,q}(\eta) \equiv n^{-1}\operatorname{tr}\big((\Sigma + \tau_*(\eta)I)^{-p}\Sigma^q\big)$ for $p, q \in \mathbb{N}$,

$$0 = -T_{-2,1}(\eta)\cdot \tau_*'(\eta) + \frac{1}{\tau_*(\eta)} - \frac{\eta}{\tau_*^2(\eta)}\cdot \tau_*'(\eta).$$

Solving for $\tau_*'(\eta)$ yields that

$$\tau_*'(\eta) = \frac{\tau_*(\eta)}{\eta + \tau_*^2(\eta)\cdot T_{-2,1}(\eta)} \equiv \frac{\tau_*(\eta)}{G_0(\eta)}. \tag{8.5}$$

Further taking derivative with respect to η on both sides of the above display (8.5), we have

$$\begin{aligned}
\tau_*''(\eta) &= \frac{1}{G_0^2(\eta)}\big(\tau_*'(\eta)G_0(\eta) - \tau_*(\eta)G_0'(\eta)\big)\\
&= \frac{1}{G_0^2(\eta)}\Big\{\tau_*(\eta) - \tau_*(\eta)\Big(1 + 2\tau_*(\eta)\tau_*'(\eta)T_{-2,1}(\eta) - 2\tau_*^2(\eta)\tau_*'(\eta)T_{-3,1}(\eta)\Big)\Big\}\\
&= \frac{2\tau_*^2(\eta)\tau_*'(\eta)}{G_0^2(\eta)}\Big(\tau_*(\eta)T_{-3,1}(\eta) - T_{-2,1}(\eta)\Big) = -\frac{2\tau_*^2(\eta)\tau_*'(\eta)}{G_0^2(\eta)}T_{-3,2}(\eta).
\end{aligned} \tag{8.6}$$

Using the apriori estimate for $\tau_*(\eta)$ proved in (2), it follows that for $q \in \{1, 2\}$,

$$1 \lesssim \inf_{\eta\in\Xi_K}(-1)^{q+1}\tau_*^{(q)}(\eta) \le \sup_{\eta\in\Xi_K}(-1)^{q+1}\tau_*^{(q)}(\eta) \lesssim 1. \tag{8.7}$$

For $\gamma_*'(\eta)$, let us define

$$\begin{aligned}
G_1(\eta) &\equiv \sigma_\xi^2 + \tau_*^2(\eta)\|(\Sigma + \tau_*(\eta)I)^{-1}\Sigma^{1/2}\mu_0\|^2,\\
G_2(\eta) &\equiv \phi - n^{-1}\operatorname{tr}\big((\Sigma + \tau_*(\eta)I)^{-2}\Sigma^2\big).
\end{aligned}$$

Then

$$\gamma_*'(\eta) = \frac{G_1'(\eta)G_2(\eta) - G_1(\eta)G_2'(\eta)}{2\gamma_*(\eta)G_2^2(\eta)}. \tag{8.8}$$

We shall now prove bounds for G_1, G_1', G_2, G_2'. First, using (8.4), we have

$$\sigma_\xi^2 \le G_1(\eta) \le \sigma_\xi^2 + \frac{\tau_*(\eta)}{2}\|\mu_0\|^2, \quad \phi\cdot\frac{\tau_*(\eta)}{\|\Sigma\|_{\mathrm{op}} + \tau_*(\eta)} \le G_2(\eta) \le \phi.$$

In particular, uniformly in $\eta \in \Xi_K$,

$$G_1(\eta), G_2(\eta) \asymp 1. \tag{8.9}$$

The derivatives G_1', G_2' are

$$G_1'(\eta) = 2\tau_*(\eta)\tau_*'(\eta)\|(\Sigma + \tau_*(\eta)I)^{-1}\Sigma^{1/2}\mu_0\|^2$$

$$- 2\tau_*^2(\eta)\|(\Sigma + \tau_*(\eta)I)^{-3/2}\Sigma^{1/2}\mu_0\|^2 \cdot \tau_*'(\eta),$$

$$G_2'(\eta) = 2 \cdot n^{-1} \operatorname{tr}\big((\Sigma + \tau_*(\eta)I)^{-3}\Sigma^2\big) \cdot \tau_*'(\eta).$$

Using the apriori estimates on $\tau_*(\eta)$ and (8.7), it now follows that

$$\sup_{\eta\in\Xi_K} \{|G_1'(\eta)| \vee |G_2'(\eta)|\} \lesssim 1. \tag{8.10}$$

Combining (8.8)-(8.10) and using apriori estimates on $\gamma_*(\eta)$, we arrive at

$$\sup_{\eta\in\Xi_K} |\gamma_*'(\eta)| \lesssim 1. \tag{8.11}$$

The claim follows by collecting (8.7) and (8.11). □

8.2. **Sample version of (2.1).** Let the sample version of (2.1) be defined by

$$\begin{cases} \phi e_h^2 \gamma^2 = \sigma_\pm^2(L_w) + \mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau), \\ (\phi e_h^2 - \frac{\eta}{\tau}) \cdot \gamma^2 = \mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau). \end{cases} \tag{8.12}$$

Here recall that e_h^2 is defined in (7.3), and $\sigma_\pm^2(L_w)$ is defined in (6.4).

Proposition 8.2. *$1/K \le \phi^{-1}, \sigma_\xi^2 \le K$ and $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ for some $K > 0$. There exist some $C, C_0 > 1$ depending on K, such that with $\delta \in (0, 1/C^{100})$, $1 \le M \le \sqrt{n}/C$ and $L_w \le C$, on the event $\mathscr{E}_1(\delta) \cap \mathscr{E}_{\Delta,\Xi}(M)$, where*

$$\mathscr{E}_{\Delta,\Xi}(M) \equiv \Big\{ \max_{\ell=1,2} \sup_{\tau\ge0} |\Delta_\ell(\tau)| \vee \max_{\ell=1,2} \sup_{\tau\ge0} n^{-1/2} |\Xi_\ell(\tau) - \mathbb{E}\,\Xi_\ell(\tau)| \le M \Big\}$$

with Δ_ℓ, Ξ_ℓ defined in Lemmas 8.3 and 8.4 ahead, the following hold.

1. *All solutions $(\gamma_{n,\eta,\pm}, \tau_{n,\eta,\pm})$ to the system of equations in (8.12) satisfy*
$$1/C_0 \le \tau_{n,\eta,\pm} \le C_0, \quad 1/C_0 \le \gamma_{n,\eta,\pm} \le C_0$$
uniformly in $\eta \in \Xi_K$.
2. *Moreover,*
$$\sup_{\eta\in\Xi_K} \{|\tau_{n,\eta,\pm} - \tau_{\eta,*}| \vee |\gamma_{n,\eta,\pm} - \gamma_{\eta,*}|\} \le C_0 \cdot (M/\sqrt{n} + \delta).$$

We need two concentration lemmas before the proof of Proposition 8.2.

Lemma 8.3. *Let $\Delta_\ell(\tau) \equiv -\ell \cdot \tau\langle(\Sigma + \tau I)^{-\ell}\Sigma^{\ell-1/2}\mu_0, g\rangle$ for $\ell = 1, 2$. Suppose that $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ for some $K > 0$. Then there exists some constant $C = C(K) > 1$ such that for $t \ge C\log(en)$,*

$$\mathbb{P}\Big(\max_{\ell=1,2} \sup_{\tau\ge0} |\Delta_\ell(\tau)| \ge C\sqrt{t} \Big) \le e^{-t}.$$

Lemma 8.4. *Let $\Xi_\ell(\tau) \equiv \|(\Sigma + \tau I)^{-\ell/2}\Sigma^{\ell/2} g\|^2$ for $\ell = 1, 2$. Suppose that $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ for some $K > 0$. Then there exists some constant $C = C(K) > 1$ such that for $t \ge C\log(en)$,*

$$\mathbb{P}\Big(\max_{\ell=1,2} \sup_{\tau\ge0} \big|\Xi_\ell(\tau) - \mathbb{E}\,\Xi_\ell(\tau)\big| \ge C(\sqrt{nt} + t) \Big) \le e^{-t}.$$

The proofs of these lemmas are deferred to the next subsection.

Proof of Proposition 8.2. All the constants in $\lesssim$, $\gtrsim$, $\asymp$ and O below may possibly depend on K. We often suppress the dependence of $\sigma_\pm^2(L_w)$ on L_w for simplicity.
(1). We shall write $(\gamma_{n,\eta,\pm}, \tau_{n,\eta,\pm})$ as (γ_n, τ_n) and $(\gamma_{\eta,*}, \tau_{\eta,*}) = (\gamma_*, \tau_*)$ for notational simplicity. Using (7.1), any solution (γ_n, τ_n) to the equations in (8.12) satisfies

$$\begin{cases} \phi e_h^2 - \frac{\eta}{\tau_n} + \frac{\Delta_1(\tau_n)}{\sqrt{n}\gamma_n} = \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-1}\Sigma\big) + \frac{1}{n}(\mathrm{id}-\mathbb{E})\Xi_1(\tau_n), \\ \phi e_h^2 + \frac{\Delta_2(\tau_n)}{\sqrt{n}\gamma_n} = \frac{1}{\gamma_n^2}\big(\sigma_\pm^2 + \tau_n^2\|(\Sigma+\tau_n I)^{-1}\Sigma^{1/2}\mu_0\|^2\big) \\ \qquad\qquad + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-2}\Sigma^2\big) + \frac{1}{n}(\mathrm{id}-\mathbb{E})\Xi_2(\tau_n). \end{cases} \tag{8.13}$$

On the event $\mathscr{E}_1(\delta)\cap\mathscr{E}_{\Delta,\Xi}(M)$ with $\delta\in(0,1/C^{100})$, $1\le M\le \sqrt{n}/C$ and $L_w\le C$, using Lemma 7.5, the second equation in (8.13) becomes

$$\begin{aligned} &\phi + O\big(M(1\vee\gamma_n^{-1})/\sqrt{n}+\delta\big) \\ &= \frac{1}{\gamma_n^2}\Big(\sigma_\pm^2 + \tau_n^2\|(\Sigma+\tau_n I)^{-1}\Sigma^{1/2}\mu_0\|^2\Big) + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-2}\Sigma^2\big) \gtrsim \frac{1}{\gamma_n^2}. \end{aligned}$$

Rearranging terms we obtain the inequality

$$\frac{1}{\gamma_n^2} \lesssim 1 + \frac{M}{\sqrt{n}} + \frac{M}{\sqrt{n}\gamma_n} \;\Rightarrow\; \gamma_n \gtrsim \frac{1}{1+M/\sqrt{n}} \gtrsim 1.$$

So with $\varepsilon_n\equiv\varepsilon_n(M,\delta)\equiv M/\sqrt{n}+\delta$, the equations in (8.13) reduce to

$$\begin{cases} \phi - \frac{\eta}{\tau_n} + O(\varepsilon_n) = \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-1}\Sigma\big), \\ \phi + O(\varepsilon_n) = \frac{1}{\gamma_n^2}\big(\sigma_\xi^2 + \tau_n^2\|(\Sigma+\tau_n I)^{-1}\Sigma^{1/2}\mu_0\|^2\big) + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-2}\Sigma^2\big). \end{cases} \tag{8.14}$$

The above equations match (2.1) up to the small perturbation $O(\varepsilon_n) = O(\varepsilon_n(M,\delta))$ that can be assimilated into the leading term ϕ with small enough $c_0>0$ such that $M\le c_0\sqrt{n}$. From here the existence (but not uniqueness) and apriori bounds for γ_n,τ_n can be established similarly to the proof of Proposition 8.1.
(2). Now we shall prove the claimed error bounds. By using (8.1) and the first equation of (8.14), we have

$$\frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-1}\Sigma\big) + \frac{\eta}{\tau_n} = \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_* I)^{-1}\Sigma\big) + \frac{\eta}{\tau_*} + O(\varepsilon_n).$$

Let $\mathsf{f}(\tau)\equiv\frac{1}{n}\operatorname{tr}\big((\Sigma+\tau I)^{-1}\Sigma\big)+\frac{\eta}{\tau}$. Then it is easy to calculate $\mathsf{f}'(\tau) = -\frac{1}{n}\operatorname{tr}\big((\Sigma+\tau I)^{-2}\Sigma\big) - \frac{\eta}{\tau^2}\le 0$, and for any $C_0>1$,

$$\inf_{1/C_0\le\tau\le C_0}|\mathsf{f}'(\tau)| \ge \inf_{1/C_0\le\tau\le C_0}\frac{1}{n}\operatorname{tr}\big((\Sigma+\tau I)^{-2}\Sigma\big) \ge (\|\Sigma\|_{\mathrm{op}}+C_0)^{-2}\mathcal{H}_\Sigma^{-1}.$$

Now using the apriori estimates on τ_*,τ_n, we may conclude

$$\sup_{\eta\in\Xi_K}|\tau_n-\tau_*| \lesssim \varepsilon_n. \tag{8.15}$$

On the other hand, using (8.2) and the second equation of (8.14), we have

$$\begin{aligned} &\frac{1}{\gamma_n^2}\Big(\sigma_\xi^2 + \tau_n^2\|(\Sigma+\tau_n I)^{-1}\Sigma^{1/2}\mu_0\|^2\Big) + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-2}\Sigma^2\big) \\ &= \frac{1}{\gamma_*^2}\Big(\sigma_\xi^2 + \tau_*^2\|(\Sigma+\tau_* I)^{-1}\Sigma^{1/2}\mu_0\|^2\Big) + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_* I)^{-2}\Sigma^2\big) + O(\varepsilon_n). \end{aligned} \tag{8.16}$$

Using the error bound in (8.15) and apriori estimates for τ_n, τ_*, and the fact that $\mathcal{H}_\Sigma \lesssim 1$, by an easy derivative estimate we have

- $\left|\frac{1}{n}\operatorname{tr}\left((\Sigma+\tau_n I)^{-2}\Sigma^2\right) - \frac{1}{n}\operatorname{tr}\left((\Sigma+\tau_* I)^{-2}\Sigma^2\right)\right| \lesssim \varepsilon_n$, and
- $\left|\tau_n^2\|(\Sigma+\tau_n I)^{-1}\Sigma^{1/2}\mu_0\|^2 - \tau_*^2\|(\Sigma+\tau_* I)^{-1}\Sigma^{1/2}\mu_0\|^2\right| \lesssim \varepsilon_n$.

Now plugging these estimates into (8.16), with $\mathscr{C}_0 \equiv \sigma_\xi^2 + \tau_*^2\|(\Sigma+\tau_* I)^{-1}\Sigma^{1/2}\mu_0\|^2$ satisfying $\mathscr{C}_0 \asymp 1$, we arrive at

$$\frac{\mathscr{C}_0 + O(\varepsilon_n)}{\gamma_n^2} = \frac{\mathscr{C}_0}{\gamma_*^2} + O(\varepsilon_n).$$

Using apriori estimates on γ_n, γ_*, we may then invert the above estimate into

$$\sup_{\eta\in\Xi_K}|\gamma_n - \gamma_*| \lesssim \varepsilon_n. \tag{8.17}$$

The claimed error bounds follow by combining (8.15) and (8.17). □

8.3. **Proofs of Lemmas 8.3 and 8.4.**

Proof of Lemma 8.3. We only handle the case $\ell = 1$. The case $\ell = 2$ is similar. Note that the assumption on μ_0 invariant over orthogonal transforms, so for notational simplicity we assume without loss of generality that Σ is diagonal. As $\sup_{\tau\geq Kn}|\Delta_1(\tau)| \leq \left|\sum_{j=1}^n \lambda_j^{1/2}\mu_{0,j}g_j\right| + Ce_g\cdot n^{-1/2}$, a standard concentration for the first term shows for $t\geq 1$, with probability $1-e^{-t}$,

$$\sup_{\tau\geq Kn}|\Delta_1(\tau)| \leq C_0\sqrt{t}. \tag{8.18}$$

On the other hand, for $\varepsilon>0$ to be chosen later, by taking an ε-net $\mathcal{S}_\varepsilon$ of $[0,Kn]$, a union bound shows that with probability at least $1-(Kn/\varepsilon+1)e^{-t}$,

$$\begin{aligned}\sup_{\tau\in[0,Kn]}|\Delta_1(\tau)| &\leq \max_{\tau\in\mathcal{S}_\varepsilon}|\Delta_1(\tau)| + \sup_{\tau,\tau'\in[0,Kn]:|\tau-\tau'|\leq\varepsilon}|\Delta_1(\tau)-\Delta_1(\tau')|\\ &\leq C_1\cdot\Big(\sqrt{t}+\sqrt{n}(\sqrt{\log n}+\sqrt{t})\varepsilon\Big).\end{aligned}$$

Here in the last inequality we used the simple estimate $\sup_{\tau\in[0,Kn]}|\partial_\tau\Delta_1(\tau)| \leq C\sqrt{n}\|\mu_0\|\|g\|_\infty$. Finally by choosing $\varepsilon \equiv \sqrt{t}/\{\sqrt{n}(\sqrt{\log n}+\sqrt{t})\}$, we conclude that for $t\geq C_2\log(en)$, with probability $1-e^{-t}$,

$$\sup_{\tau\in[0,Kn]}|\Delta_1(\tau)| \leq C_2\sqrt{t}. \tag{8.19}$$

The claim follows by combining (8.18) and (8.19). □

Proof of Lemma 8.4. We focus on the case $\ell = 1$ and will follow a similar idea used in the proof of Lemma 8.3 above. Similarly we assume Σ is diagonal without loss of generality. All the constants in $\lesssim, \gtrsim, \asymp$ below may depend on K.

First note by a standard concentration, for any $t\geq 1$, with probability at least $1-e^{-t}$, $\sup_{\tau>Kn}|\Xi_1(\tau)| \lesssim e_g^2 \lesssim 1+t/n$. Similarly we have $\sup_{\tau>Kn}\mathbb{E}|\Xi_1(\tau)| \lesssim 1$. This means for any $t\geq 1$, with probability at least $1-e^{-t}$,

$$\sup_{\tau>Kn}\Big(|\Xi_1(\tau)| \vee \mathbb{E}|\Xi_1(\tau)|\Big) \lesssim 1+t/n. \tag{8.20}$$

Next we handle the suprema over $[0, Kn]$ by discretization over an ε-net $\mathcal{S}_\varepsilon$. To this end, we shall establish a pointwise concentration. Note that $\|\nabla\Xi_1(\tau)\|^2 = 4\|(\Sigma + \tau I)^{-1}\Sigma g\|^2 \le 4\Xi_1(\tau)$. An application of Proposition B.1 then yields that, for each $\tau \ge 0$ and $t \ge 1$, with probability at least $1 - e^{-t}$,

$$|\Xi_1(\tau) - \mathbb{E}\,\Xi_1(\tau)| \le C\big(\mathbb{E}^{1/2}\,\Xi_1(\tau)\cdot\sqrt{t} + t\big) \lesssim (\sqrt{nt} + t).$$

On the other hand, as $\sup_{\tau\in[0,Kn]}|\partial_\tau\Xi_1(\tau)| \lesssim n\|g\|_\infty^2$ and $\sup_{\tau\in[0,Kn]}|\partial_\tau\,\mathbb{E}\,\Xi_1(\tau)| \lesssim n\log n$, we deduce that with probability at least $1 - (Kn/\varepsilon + 1)e^{-t}$,

$$\begin{aligned}\sup_{\tau\in[0,Kn]}|\Xi_1(\tau) - \mathbb{E}\,\Xi_1(\tau)| &\le \max_{\tau\in\mathcal{S}_\varepsilon}|\Xi_1(\tau) - \mathbb{E}\,\Xi_1(\tau)| + \sup_{\tau,\tau'\in[0,Kn]:|\tau-\tau'|\le\varepsilon}|\Xi_1(\tau) - \Xi_1(\tau')|\\ &\quad + \sup_{\tau,\tau'\in[0,Kn]:|\tau-\tau'|\le\varepsilon}|\mathbb{E}\,\Xi_1(\tau) - \mathbb{E}\,\Xi_1(\tau')|\\ &\lesssim \sqrt{nt} + t + n(\log n + t)\varepsilon.\end{aligned}$$

From here the claim follows by the same arguments used in the proof of Lemma 8.3 above. □

9. Gaussian designs: Proof of Theorem 2.3

We assume without loss of generality that $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, so $\mathsf{V} = I$ unless otherwise specified. Recall $\mathcal{H}_\Sigma = \mathrm{tr}(\Sigma^{-1})/n$.

9.1. Localization of the primal problem.

Proposition 9.1. *Suppose $1/K \le \phi^{-1} - \mathbf{1}_{\eta=0}, \sigma_\xi^2 \le K$, and $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \le K$ for some $K > 0$. Fix $M > 1, \delta \in (0, 1/2)$ and $\eta \ge 0$. On the event $\mathscr{E}_0(M) \cap \mathscr{E}_1(\delta)$, there exists some $C = C(K) > 0$ such that for any deterministic choice of (L_w, L_v) with*

$$L_w \wedge L_v \ge C\{1 + (\|\Sigma^{-1}\|_{\mathrm{op}} M\mathbf{1}^{-1}_{\phi^{-1}\ge 1+1/K} \wedge \eta^{-1})\cdot M^2\},$$

we have $\min_{w\in B_n(L_w)} H_\eta(w; L_v) = \min_{w\in\mathbb{R}^n} H_\eta(w)$.

Proof. Using the first-order optimality condition for the minimax problem

$$\min_{w\in\mathbb{R}^n} H_\eta(w) = \min_{w\in\mathbb{R}^n}\max_{v\in\mathbb{R}^m}\Big\{\frac{1}{\sqrt{n}}\langle v, Gw - \xi\rangle + F(w) - \frac{\eta\|v\|^2}{2}\Big\}, \tag{9.1}$$

any saddle point (w_*, v_*) of (9.1) must satisfy $\nabla F(w_*) = -\frac{1}{\sqrt{n}}G^\top v_*$ and $\frac{1}{\sqrt{n}}(Gw_* - \xi) = \eta v_*$, or equivalently,

$$\begin{cases} w_* = -\Sigma^{1/2}\mu_0 + \frac{1}{n}\Sigma G^\top(\phi\check{\Sigma} + \eta I)^{-1}(G\Sigma^{1/2}\mu_0 + \xi),\\ v_* = -\frac{1}{\sqrt{n}}(\phi\check{\Sigma} + \eta I)^{-1}(G\Sigma^{1/2}\mu_0 + \xi).\end{cases}$$

Here recall $\check{\Sigma} = m^{-1}G\Sigma G^\top$. On the event $\mathscr{E}_0(M)$,

$$\|(\phi\check{\Sigma} + \eta I)^{-1}\|_{\mathrm{op}} \lesssim_K \|\Sigma^{-1}\|_{\mathrm{op}} M\mathbf{1}^{-1}_{\phi^{-1}\ge 1+1/K} \wedge \eta^{-1}.$$

So on $\mathscr{E}_0(M) \cap \mathscr{E}_1(\delta)$,

$$\|w_*\| \vee \|v_*\| \lesssim_K 1 + (\|\Sigma^{-1}\|_{\mathrm{op}} M\mathbf{1}^{-1}_{\phi^{-1}\ge 1+1/K} \wedge \eta^{-1})M^2.$$

This means that on the event $\mathscr{E}_0(M) \cap \mathscr{E}_1(\delta)$, for any L_w, L_v chosen as in the statement of the lemma,

$$\min_{w\in\mathbb{R}^n} H_\eta(w) = \min_{w\in B_n(L_w)} \max_{v\in B_m(L_v)} \Big\{\frac{1}{\sqrt{n}}\langle v, Gw-\xi\rangle + F(w) - \frac{\eta\|v\|^2}{2}\Big\}.$$

The proof is complete by recalling the definition of $H_\eta(\cdot; L_v)$. □

9.2. **Characterization of the Gordon cost optimum.**

Theorem 9.2. *Suppose the following hold for some $K > 0$.*

- $1/K \le \phi^{-1} \le K$, $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$.
- *Assumption B with $\sigma_\xi^2 \in [1/K, K]$.*

There exist some $C, C' > 1$ depending on K such that for any deterministic choice of $L_w, L_v \in [C, C^2]$, it holds for any $C' \log(en) \le t \le n/C'$, $\eta \in \Xi_K$ and $\xi \in \mathscr{E}_{1,\xi}(\sqrt{t/n})$,

$$\mathbb{P}^\xi\Big(\Big|\min_{w\in B_n(L_w)} L_\eta(w; L_v) - \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma)\Big| \ge \sqrt{t/n}\Big) \le Ce^{-t/C}.$$

In the next subsection we will show that for large $L_v > 0$, the map $w \mapsto L_\eta(w; L_v)$ attains its global minimum in an ℓ_2 ball of constant order radius (under $\mathcal{H}_\Sigma \lesssim 1$) with high probability. This means that although the initial localization radius for the primal optimization may be highly suboptimal (which involves $\|\Sigma^{-1}\|_{\mathrm{op}}$), the Gordon objective can be further localized into an ℓ_2 ball with constant order radius.

To prove Theorem 9.2, we shall first relate $\min_{w\in B_n(L_w)} L_\eta(w; L_v)$ to $\max_{\beta>0}\min_{\gamma>0} \mathsf{D}_{\eta,\pm}(\beta,\gamma)$ and its localized versions.

Proposition 9.3. *Suppose $1/K \le \phi^{-1}, \sigma_\xi^2 \le K$, and $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ for some $K > 0$. There exists constant $C = C(K) > 1$ such that for any deterministic choice of $L_w, L_v \in [C, C^2]$, on the event $\mathscr{E}_1(\delta) \cap \mathscr{E}_{\Delta,\Xi}(M)$ (defined in Proposition 8.2) with $\delta \in (0, 1/C^{100})$ and $M \le \sqrt{n}/C$, we have for any $\eta \in \Xi_K$,*

$$\max_{\beta>0}\min_{\gamma>0} \mathsf{D}_{\eta,-}(\beta,\gamma) \le \min_{w\in B_n(L_w)} L_\eta(w; L_v) \le \max_{\beta>0}\min_{\gamma>0} \mathsf{D}_{\eta,+}(\beta,\gamma),$$

and the following localization holds:

$$\max_{\beta>0}\min_{\gamma>0} \mathsf{D}_{\eta,\pm}(\beta,\gamma) = \max_{1/C\le\beta\le C}\min_{1/C\le\gamma\le C} \mathsf{D}_{\eta,\pm}(\beta,\gamma).$$

Proof. We write $g_n \equiv g/\sqrt{n}$ in the proof.
(**Step 1**). Fix any $L_w, L_v > 0$. We may compute

$$\begin{aligned}
&\min_{w\in B_n(L_w)} L_\eta(w; L_v) \qquad (9.2)\\
&= \min_{w\in B_n(L_w)} \max_{\beta\in[0,L_v]} \Big\{\frac{\beta}{\sqrt{n}}\Big(\big\|\|w\|h-\xi\big\| - \langle g, w\rangle\Big) + F(w) - \frac{\eta\beta^2}{2}\Big\}\\
&= \max_{\beta\in[0,L_v]} \min_{\gamma>0} \Big\{\frac{\beta\gamma\|h\|^2}{2n} - \frac{\eta\beta^2}{2} + \min_{w\in B_n(L_w)} \Big(\frac{\beta}{2\gamma}\frac{\|\|w\|h-\xi\|^2}{\|h\|^2} - \langle w, \beta g_n\rangle + F(w)\Big)\Big\}.
\end{aligned}$$

Here in the last line we used Sion's min-max theorem to flip the order of minimum and maximum in $\min_{w\in B_n(L_w)}\max_{\beta\in[0,L_v]}$. The minimum over γ is achieved exactly at $\frac{\|\|w\|h-\xi\|/\|h\|}{\|h\|/\sqrt{n}}$, so when $\sigma_-^2\neq 0$, using the simple inequality

$$\|w\|^2+\sigma_-^2\le \|\|w\|h-\xi\|^2/\|h\|^2\le \|w\|^2+\sigma_+^2, \tag{9.3}$$

on the event $\mathscr{E}_1(\delta)$, we may further bound (9.2) as follows:

$$\begin{aligned}\pm \min_{w\in B_n(L_w)} L_\eta(w;L_v) &\le \pm \max_{\beta\in[0,L_v]}\min_{\gamma>0}\\ &\bigg\{\frac{\beta\gamma\|h\|^2}{2n}-\frac{\eta\beta^2}{2}+\min_{w\in B_n(L_w)}\bigg(\frac{\beta}{2\gamma}(\|w\|^2+\sigma_\pm^2(L_w))-\langle w,\beta g_n\rangle+F(w)\bigg)\bigg\}. \end{aligned}\tag{9.4}$$

We note that $\sigma_\pm^2$ depends on L_w, but this notational dependence will be dropped from now on for convenience.

(**Step 2**). Consider the minimax optimization problem in (9.4):

$$\begin{aligned}&\max_{\beta>0}\min_{\gamma>0,w\in\mathbb{R}^n}\bigg\{\frac{\beta\gamma\|h\|^2}{2n}-\frac{\eta\beta^2}{2}+\bigg(\frac{\beta}{2\gamma}(\|w\|^2+\sigma_\pm^2)-\langle w,\beta g_n\rangle+F(w)\bigg)\bigg\}\\ &=\max_{\beta>0}\min_{\gamma>0}\bigg\{\frac{\beta}{2}\bigg(\gamma(\phi e_h^2-e_g^2)+\frac{\sigma_\pm^2}{\gamma}\bigg)-\frac{\eta\beta^2}{2}+\mathsf{e}_F(\gamma g_n;\gamma/\beta)\bigg\}.\end{aligned}\tag{9.5}$$

Any saddle point $(\beta_{n,\eta,\pm},\gamma_{n,\eta,\pm},w_{n,\eta,\pm})=(\beta_{n,\pm},\gamma_{n,\pm},w_{n,\pm})$ of the above program must satisfy the first-order optimality condition

$$\begin{cases}0=\frac{1}{2}(\gamma_{n,\pm}(\phi e_h^2-e_g^2)+\frac{\sigma_\pm^2}{\gamma_{n,\pm}})-\eta\beta_{n,\pm}+\partial_\beta\mathsf{e}_F(\gamma_{n,\pm}g_n;\gamma_{n,\pm}/\beta_{n,\pm}),\\ 0=\frac{\beta_{n,\pm}}{2}((\phi e_h^2-e_g^2)-\frac{\sigma_\pm^2}{\gamma_{n,\pm}^2})+\partial_\gamma\mathsf{e}_F(\gamma_{n,\pm}g_n;\gamma_{n,\pm}/\beta_{n,\pm}),\\ w_{n,\pm}=\mathsf{prox}_F\,(\gamma_{n,\pm}g_n;\gamma_{n,\pm}/\beta_{n,\pm}).\end{cases}\tag{9.6}$$

Using the derivative formula in Lemma 7.3 and the form of prox_F in Lemma 7.2, we may compute

$$\begin{cases}\partial_\beta\mathsf{e}_F(\gamma g_n;\gamma/\beta)=\frac{1}{2\gamma}\Big(\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\gamma/\beta)-2\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\gamma/\beta)+\gamma^2e_g^2\Big),\\ \partial_\gamma\mathsf{e}_F(\gamma g_n;\gamma/\beta)=\frac{\beta}{2\gamma^2}\Big(\gamma^2e_g^2-\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\gamma/\beta)\Big).\end{cases}\tag{9.7}$$

Plugging (9.7) into (9.6), the first-order optimality condition for $(\beta_{n,\pm},\gamma_{n,\pm})$ in the minimax program (9.5) is given by

$$\begin{cases}(\phi e_h^2-e_g^2)\gamma_{n,\pm}^2+\sigma_\pm^2=2\eta\cdot\gamma_{n,\pm}\beta_{n,\pm}-\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_{n,\pm};\gamma_{n,\pm}/\beta_{n,\pm})\\ \qquad\qquad\qquad\qquad +2\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma_{n,\pm};\gamma_{n,\pm}/\beta_{n,\pm})-e_g^2\gamma_{n,\pm}^2,\\ (\phi e_h^2-e_g^2)\gamma_{n,\pm}^2-\sigma_\pm^2=-e_g^2\gamma_{n,\pm}^2+\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_{n,\pm};\gamma_{n,\pm}/\beta_{n,\pm}).\end{cases}$$

Equivalently,

$$\begin{cases}\phi e_h^2\gamma_{n,\pm}^2=\sigma_\pm^2+\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_{n,\pm};\gamma_{n,\pm}/\beta_{n,\pm}),\\ (\phi e_h^2-\frac{\eta}{\gamma_{n,\pm}/\beta_{n,\pm}})\gamma_{n,\pm}^2=\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma_{n,\pm};\gamma_{n,\pm}/\beta_{n,\pm}).\end{cases}\tag{9.8}$$

Using the apriori estimates in Proposition 8.2, on the event $\mathscr{E}_1(\delta)\cap\mathscr{E}_{\Delta,\Xi}(M)$ we have $\gamma_{n,\pm}/\beta_{n,\pm}\asymp_K 1$ and $\gamma_{n,\pm}^2\asymp_K 1$. This implies on the same event,

$$\gamma_{n,\pm}\asymp_K 1,\quad \beta_{n,\pm}\asymp_K 1.\tag{9.9}$$

Using the last equation of (9.6), we have

$$\|w_{n,\pm}\| = \Big\|\Sigma^{1/2}\Big(\Sigma + \frac{\gamma_{n,\pm}}{\beta_{n,\pm}} I\Big)^{-1}\Big(-\frac{\gamma_{n,\pm}}{\beta_{n,\pm}}\mu_0 + \gamma_{n,\pm}\Sigma^{1/2} g_n\Big)\Big\| \lesssim_K 1. \tag{9.10}$$

In view of (9.9)-(9.10), by choosing $L_w, L_v \in [C, C^2]$ for large enough $C > 0$, the constraints in the optimization in (9.4) can be dropped for free. $\square$

Next we replace the random function $\mathsf{D}_{\eta,\pm}$ in the above proposition by its deterministic counterpart $\overline{\mathsf{D}}_\eta$ in their localized versions.

Proposition 9.4. *Suppose $1/K \le \phi^{-1}, \sigma_\xi^2 \le K$, and $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ for some $K > 0$. There exist some $C, C' > 1$ depending on K such that for $L_w \in [C, C^2]$, $\delta \in (0, 1/C^{100})$, $\xi \in \mathscr{E}_{1,\xi}(\delta)$ and $t \ge C' \log(en)$,*

$$\mathbb{P}^\xi\Big[\sup_{\eta\in\Xi_K}\Big|\max_{1/C\le\beta\le C}\min_{1/C\le\gamma\le C}\mathsf{D}_{\eta,\pm}(\beta,\gamma) - \max_{1/C\le\beta\le C}\min_{1/C\le\gamma\le C}\overline{\mathsf{D}}_\eta(\beta,\gamma)\Big|$$
$$\ge C\big(\sqrt{t/n} + t/n + \delta\big)\Big] \le Ce^{-t/C} + \mathbb{P}^\xi\big(\mathscr{E}_{1,0}(\delta)^c\big).$$

Proof. In the proof, we write $g_n \equiv g/\sqrt{n}$. All the constants in $\lesssim, \gtrsim, \asymp$ and O below may depend on K.

(**Step 1**). We first prove the following: On the event $\mathscr{E}_1(\delta)$, for any $C_0 > 1$,

$$\sup_{\gamma,\tau\in[1/C_0, C_0]^2} |\partial_\# \mathsf{e}_F(\gamma g_n; \tau)| \vee |\partial_\# \mathbb{E}\, \mathsf{e}_F(\gamma g_n; \tau)| \lesssim 1. \tag{9.11}$$

To this end, with $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}, y^{\mathsf{seq}}_{(\Sigma,\mu_0)}$ written as $\widehat{\mu}, y$, and using $\partial_\gamma\widehat{\mu} = (\Sigma + \tau I)^{-1}\Sigma^{1/2} g_n$, $\partial_\gamma y = g_n$, $\partial_\tau\widehat{\mu} = -(\Sigma + \tau I)^{-2}\Sigma^{1/2}(\Sigma^{1/2}\mu_0 + \gamma g_n)$,

$$\partial_\gamma \mathsf{e}_F(\gamma g_n; \tau) = \tau^{-1}\langle \Sigma^{1/2}\widehat{\mu} - y, \Sigma^{1/2}\partial_\gamma\widehat{\mu} - \partial_\gamma y\rangle + \langle\widehat{\mu}, \partial_\gamma\widehat{\mu}\rangle,$$
$$\partial_\tau \mathsf{e}_F(\gamma g_n; \tau) = -\frac{1}{2\tau^2}\|\Sigma^{1/2}\widehat{\mu} - y\|^2 + \frac{1}{\tau}\langle\Sigma^{1/2}\widehat{\mu} - y, \Sigma^{1/2}\partial_\tau\widehat{\mu}\rangle + \langle\widehat{\mu}, \partial_\tau\widehat{\mu}\rangle,$$

on the event $\mathscr{E}_1(\delta)$, we may estimate $|\partial_\gamma \mathsf{e}_F(\gamma g_n;\tau)| \vee |\partial_\tau \mathsf{e}_F(\gamma g_n;\tau)| \lesssim 1$. A similar estimate applies to the expectation versions, proving (9.11).

(**Step 2**). Next we show that for any $C_0 > 1$, there exists $C_1 > 0$ such that for $t \ge C_1 \log(en)$,

$$\mathbb{P}\Big(\sup_{(\gamma,\tau)\in[C_0^{-1}, C_0]^2}\big|(\mathrm{id} - \mathbb{E})\mathsf{e}_F(\gamma g_n;\tau)\big| \ge C_1\big(\sqrt{t/n} + t/n\big), \mathscr{E}_1(\delta)\Big) \le C_1 e^{-t/C_1}. \tag{9.12}$$

To prove the claim, we fix $\varepsilon > 0$ to be chosen later, and take an ε-net $\mathcal{S}(\varepsilon)$ for $[1/C_0, C_0]$. Then $|\mathcal{S}(\varepsilon)| \le C_0/\varepsilon + 1$. So on the event $\mathscr{E}_1(\delta)$, using the estimate in (9.11) and a union bound via the pointwise concentration inequality in Proposition 7.4, for $t \ge 1$, with probability at least $1 - C\varepsilon^{-2}e^{-t/C}$,

$$\sup_{(\gamma,\tau)\in[C_0^{-1}, C_0]^2}\big|(\mathrm{id} - \mathbb{E})\mathsf{e}_F(\gamma g_n;\tau)\big| \lesssim \sup_{\gamma,\tau\in\mathcal{S}(\varepsilon)}\big|(\mathrm{id} - \mathbb{E})\mathsf{e}_F(\gamma g_n;\tau)\big| + C\varepsilon \lesssim \sqrt{\frac{t}{n}} + \frac{t}{n} + \varepsilon.$$

Here in the last inequality we used Lemma 7.2 to estimate $\sup_{(\gamma,\tau)} v^2(\gamma,\tau) \vee \sup_{(\gamma,\tau)} v^2(\gamma,\tau)\,\mathbb{E}\,\mathsf{e}_F(\gamma g_n;\tau) \lesssim 1$, where $v^2(\gamma,\tau)$ is defined in Proposition 7.4. The claim (9.12) follows by choosing $\varepsilon \equiv \sqrt{t/n} + t/n$ and some calculations.

(**Step 3**). By (9.12), for $t \geq C\log(en)$, on the event $\mathscr{E}_1(\delta)$, it holds with probability at least $1 - C_2 e^{-t/C_2}$ that

$$
\begin{aligned}
&\max_{1/C\leq\beta\leq C}\min_{1/C\leq\gamma\leq C}\mathsf{D}_{\eta,\pm}(\beta,\gamma)\\
&= \max_{1/C\leq\beta\leq C}\min_{1/C\leq\gamma\leq C}\bigg\{\frac{\beta}{2}\Big(\gamma(\phi e_h^2 - e_g^2) + \frac{\sigma_\pm^2}{\gamma}\Big) - \frac{\eta\beta^2}{2} + \mathsf{e}_F(\gamma g_n;\gamma/\beta)\bigg\}\\
&= \max_{1/C\leq\beta\leq C}\min_{1/C\leq\gamma\leq C}\overline{\mathsf{D}}_\eta(\beta,\gamma) + \mathcal{O}(\sqrt{t/n} + t/n + \delta).
\end{aligned}
$$

The estimate in $\mathcal{O}$ is uniform in $\eta \in \Xi_K$, so the claim follows. □

Finally we delocalize the range constraints for β,γ in the deterministic minimax problem with $\overline{\mathsf{D}}_\eta$ in the above proposition.

Proposition 9.5. *Suppose $1/K \leq \phi^{-1}, \sigma_\xi^2 \leq K$, and $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \leq K$ for some $K > 0$. There exists some $C = C(K) > 1$ such that for any $\eta \in \Xi_K$,*

$$
\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) = \max_{1/C\leq\beta\leq C}\min_{1/C\leq\gamma\leq C}\overline{\mathsf{D}}_\eta(\beta,\gamma).
$$

Consequently,

$$
\Big|\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) - \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_0(\beta,\gamma)\Big| \leq C\eta. \tag{9.13}
$$

Proof. The proof is essentially a deterministic version of Step 2 in the proof of Proposition 9.3. We give some details below. We write $g_n \equiv g/\sqrt{n}$. First, using similar calculations as that of (9.7),

$$
\begin{cases}
\partial_\beta\,\mathbb{E}\,\mathsf{e}_F(\gamma g_n;\gamma/\beta) = \frac{1}{2\gamma}\Big(\mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\gamma/\beta) - 2\,\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\gamma/\beta) + \gamma^2\Big),\\
\partial_\gamma\,\mathbb{E}\,\mathsf{e}_F(\gamma g_n;\gamma/\beta) = \frac{\beta}{2\gamma^2}\Big(\gamma^2 - \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\gamma/\beta)\Big).
\end{cases}
$$

Then the first-order optimality condition for (β_*,γ_*) to be the saddle point of $\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)$, i.e., a deterministic version of (9.8), is given by

$$
\begin{cases}
\phi\gamma_*^2 = \sigma_\xi^2 + \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_*;\gamma_*/\beta_*),\\
(\phi - \frac{\eta}{\gamma_*/\beta_*})\gamma_*^2 = \mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma_*;\gamma_*/\beta_*).
\end{cases}
$$

Finally using the apriori estimates in Proposition 8.1, we obtain a deterministic analogue of (9.9) in that $\gamma_* \asymp_K 1$, $\quad \beta_* \asymp_K 1$. The claimed localization follows. The continuity follows by the definition of $\overline{\mathsf{D}}_\eta$ and the proven localization. □

Proof of Theorem 9.2. By Propositions 9.3, 9.4 and 9.5, there exist $C, C' > 0$ such that for any $\delta \in (0, 1/C^{100})$, $M \leq \sqrt{n}/C$, $t \geq C'\log(en)$, $\xi \in \mathscr{E}_{1,\xi}(\delta)$ and $\eta \in \Xi_K$,

$$
\begin{aligned}
&\mathbb{P}^\xi\Big[\Big|\min_{w\in B_n(L_w)} L_\eta(w;L_v) - \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)\Big| \geq C(\sqrt{t/n} + t/n + \delta)\Big]\\
&\leq Ce^{-t/C} + \mathbb{P}^\xi\,(\mathscr{E}_{1,0}(\delta)^c) + \mathbb{P}(\mathscr{E}_{\Delta,\Xi}(M)^c).
\end{aligned}
$$

The claim now follows from the concentration estimates in Lemmas 7.7, 8.3 and 8.4, by choosing $M \equiv \sqrt{n}/C$ and $\delta \equiv C(\sqrt{t/n} + t/n)$, which is valid in the regime $t \leq n/C_0$ for large C_0. □

9.3. **Locating the global minimizer of the Gordon objective.** With $(\gamma_{\eta,*}, \tau_{\eta,*})$ denoting the unique solution to the system of equations (2.1), let

$$w_{\eta,*} \equiv \mathsf{prox}_F\,(\gamma_{\eta,*}g/\sqrt{n}; \tau_{\eta,*}) = \Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*}; \tau_{\eta,*}) - \mu_0). \tag{9.14}$$

For any $\varepsilon > 0$, let the exceptional set be defined as

$$D_{\eta;\varepsilon}(\mathsf{g}) \equiv \{w \in \mathbb{R}^n : |\mathsf{g}(w) - \mathbb{E}\,\mathsf{g}(w_{\eta,*})| \geq \varepsilon\}. \tag{9.15}$$

Theorem 9.6. *Suppose the following hold for some $K > 0$.*

- $1/K \leq \phi^{-1} \leq K$, $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \leq K$.
- *Assumption B holds with $\sigma_\xi^2 \in [1/K, K]$.*

Fix any $\mathsf{g} : \mathbb{R}^n \to \mathbb{R}$ that is 1-Lipschitz with respect to $\|\cdot\|_{\Sigma^{-1}}$. There exist constants $C, C' > 10$ depending on K such that for $L_w, L_v \in [C, C^2]$, $C' \log(en) \leq t \leq n/C'$, $\xi \in \mathscr{E}_{1,\xi}(\sqrt{t/n})$ and $\eta \in \Xi_K$,

$$\mathbb{P}^\xi\bigg(\min_{w \in D_{\eta;C(t/n)^{1/4}}(\mathsf{g}) \cap B_n(L_w)} L_\eta(w; L_v) \leq \max_{\beta>0} \min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta, \gamma) + \sqrt{t/n}\bigg) \leq Ce^{-t/C}.$$

Roughly speaking, the above theorem will be proved by approximating L_η both from above and below by nicer strongly convex, surrogate functions whose minimizers can be directly located. Then we may relate the minimizer of L_η and those of the surrogate functions.

We first formally define these surrogate functions. For $L_w > 0, L_v > 0$, let

$$L_{\eta,\pm}(w; L_v) \equiv \max_{\beta \in [0, L_v]} \bigg\{\frac{\beta}{\sqrt{n}}\Big(\|h\|\sqrt{\|w\|^2 + \sigma_\pm^2(L_w)} - \langle g, w\rangle\Big) - \frac{\eta\beta^2}{2} + F(w)\bigg\}. \tag{9.16}$$

Again we omit notational dependence of $L_{\eta,\pm}$ on L_w for simplicity.

The following lemma provides uniform (bracketing) approximation of L_η via $L_{\eta,\pm}$ on compact sets.

Lemma 9.7. *Fix $L_v > 0$. The following hold when $\sigma_-^2(L_w) \neq 0$.*

(1) For any $w \in B_n(L_w)$, $L_{\eta,-}(w; L_v) \leq L_\eta(w; L_v) \leq L_{\eta,+}(w; L_v)$.
(2) For any $L_w > 0$,

$$\sup_{w \in \mathbb{R}^n} |L_{\eta,+}(w; L_v) - L_{\eta,-}(w; L_v)| \leq \frac{4e_h}{\sigma_m} \cdot L_v L_w \frac{|\langle h, \xi\rangle|}{\|h\|^2}.$$

Proof. The first claim (1) follows by the definition of $\sigma_\pm^2(L_w)$ in (6.4) and the simple inequality (9.3). For (2), note that

$$\begin{aligned} \big|L_{\eta,+}(w; L_v) - L_{\eta,-}(w; L_v)\big| &\leq L_v e_h \cdot \Big|\sqrt{\|w\|^2 + \sigma_+^2(L_w)} - \sqrt{\|w\|^2 + \sigma_-^2(L_w)}\Big| \\ &\leq L_v e_h \cdot \frac{|\sigma_+^2(L_w) - \sigma_-^2(L_w)|}{\sigma_+(L_w) + \sigma_-(L_w)} \leq \frac{4e_h}{\sigma_m} \cdot L_v L_w \frac{|\langle h, \xi\rangle|}{\|h\|^2}, \end{aligned}$$

as desired. □

Next, we will study the properties of the global minimizers for $L_{\eta,\pm}$.

Proposition 9.8. *Suppose $1/K \le \phi^{-1}, \sigma_\xi^2 \le K$, and $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ for some $K > 0$. There exists some constant $C = C(K) > 1$ such that for any deterministic choice of $L_w, L_v \in [C, C^2]$, on the event $\mathscr{E}_1(\delta) \cap \mathscr{E}_{\Delta,\Xi}(M)$ (defined in Proposition 8.2) with $\delta \in (0, 1/C^{100})$ and $M \le \sqrt{n}/C$, for any $\eta \in \Xi_K$, the maps $w \mapsto L_{\eta,\pm}(w; L_v)$ attain its global minimum at $w_{n,\eta,\pm}$ with $\|w_{n,\eta,\pm}\|_{\Sigma^{-1}} \le C$. Moreover, $\|w_{n,\eta,\pm} - w_{\eta,*}\|_{\Sigma^{-1}} \le C(M/\sqrt{n} + \delta)^{1/2}$.*

Proof. Note that the optimization problem

$$
\begin{aligned}
&\min_{w \in \mathbb{R}^n} L_{\eta,\pm}(w; L_v)\\
&= \min_{w \in \mathbb{R}^n} \max_{\beta \in [0, L_v]} \bigg\{ \frac{\beta}{\sqrt{n}} \Big(\|h\| \sqrt{\|w\|^2 + \sigma_\pm^2(L_w)} - \langle g, w \rangle \Big) - \frac{\eta \beta^2}{2} + F(w) \bigg\}\\
&\overset{(*)}{=} \max_{\beta \in [0, L_v]} \min_{w \in \mathbb{R}^n} \bigg\{ \frac{\beta}{\sqrt{n}} \Big(\|h\| \sqrt{\|w\|^2 + \sigma_\pm^2(L_w)} - \langle g, w \rangle \Big) - \frac{\eta \beta^2}{2} + F(w) \bigg\}\\
&= \max_{\beta \in [0, L_v]} \min_{\gamma > 0, w \in \mathbb{R}^n} \bigg\{ \frac{\beta \gamma \|h\|^2}{2n} - \frac{\eta \beta^2}{2} + \bigg(\frac{\beta}{2\gamma} (\|w\|^2 + \sigma_\pm^2(L_w)) - \Big\langle w, \frac{\beta}{\sqrt{n}} g \Big\rangle + F(w) \bigg) \bigg\}.
\end{aligned}
$$

Here in $(*)$ we used Sion's min-max theorem to exchange minimum and maximum, as the maximum is taken over a compact set. The difference of the above minimax problem compared to (9.5) rests in its range constraint on β. As proven in (9.9), all solutions $\beta_{n,\pm}$ to the unconstrained minimax problem (9.5) must satisfy $\beta_{n,\eta,\pm} \le C$ on the event $\mathscr{E}_1(\delta) \cap \mathscr{E}_{\Delta,\Xi}(M)$. So on this event, for the choice $L_w, L_v \in [C, C^2]$ for some large $C > 0$, $\min_{w \in \mathbb{R}^n} L_{\eta,\pm}(w; L_v)$ exactly corresponds to (9.5), whose minimizers $w_{n,\pm}$ admit the apriori estimate (9.10) (with minor modifications that change $\|\cdot\|$ to the stronger estimate in $\|\cdot\|_{\Sigma^{-1}}$).

Next, for the error bound, using the last equation in (9.6) and the definition of $w_{\eta,*}$ in (9.14), along with the estimates in Proposition 8.2, we have

$$
\begin{aligned}
\|w_{n,\eta,\pm} - w_{\eta,*}\|_{\Sigma^{-1}}^2 &= \big\| \widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{n,\eta,\pm}; \tau_{n,\eta,\pm}) - \widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*}; \tau_{\eta,*}) \big\|^2\\
&\lesssim_K |\gamma_{n,\eta,\pm} - \gamma_{\eta,*}| \vee |\tau_{n,\eta,\pm} - \tau_{\eta,*}| \le C(M/\sqrt{n} + \delta),
\end{aligned}
$$

as desired. ◻

Finally we shall relate back to the global minimizer of L_η. We note that the proposition below by itself is not formally used in the proof of Theorem 9.6, but will turn out to be useful in the proof of Theorem 2.3 ahead.

Proposition 9.9. *Suppose the conditions in Theorem 9.6 hold for some $K > 0$. There exist constants $C, C' > 1$ depending on K such that for $L_w, L_v \in [C, C^2]$, $C' \log(en) \le t \le n/C'$, $\xi \in \mathscr{E}_{1,\xi}(\sqrt{t/n})$ and $\eta \in \Xi_K$,*

$$
\begin{aligned}
\mathbb{P}^\xi \Big(&\text{The map } w \mapsto L_\eta(w; L_v) \text{ attains its global minimum at } w_{n,\eta} \text{ with } \|w_{n,\eta}\|_{\Sigma^{-1}} \le C,\\
&\text{and } \|w_{n,\eta} - w_{\eta,*}\|_{\Sigma^{-1}} \le C(t/n)^{1/4} \Big) \ge 1 - Ce^{-t/C}.
\end{aligned}
$$

Proof. Let us fix $\xi \in \mathscr{E}_{1,\xi}(\sqrt{t/n})$.

(**Step 1**). We first prove the apriori estimate for $\|w_{n,\eta}\|_{\Sigma^{-1}}$. To this end, for large enough $C_0, C_0' > 0$ depending on K, we choose $L_w \equiv C_0, \delta \equiv 1/C_0^{100}$ and $M \equiv \delta\sqrt{n}$ in Proposition 9.8, it follows that

$$\mathbb{P}^{\xi}\Big(E_1 \equiv \Big\{\|w_{n,\eta,\pm}\|_{\Sigma^{-1}} \vee \|w_{n,\eta,\pm}\| \leq C_0/2, \\ L_{\eta,\pm}(w_{n,\eta,\pm}; L_v) = (9.5)\Big\}\Big) \geq 1 - C_0 e^{-n/C_0}. \tag{9.17}$$

On the other hand, choosing $\delta \equiv \sqrt{t/n}$ with $C_0' \log(en) \leq t \leq n/C_0'$ leads to

$$\mathbb{P}^{\xi}\Big(E_2(t) \equiv \Big\{\|w_{n,\eta,\pm} - w_{\eta,*}\|_{\Sigma^{-1}} \leq C_0 (t/n)^{1/4}\Big\}\Big) \geq 1 - C_0 e^{-t/C_0}. \tag{9.18}$$

On E_1, we may characterize the value of $L_{\eta,\pm}(w_{n,\eta,\pm}; L_v)$ by applying Propositions 9.3-9.5: for $C_0' \log(en) \leq t \leq n/C_0'$,

$$\mathbb{P}^{\xi}\Big(E_3(t) \equiv \Big\{\Big|L_{\eta,\pm}(w_{n,\eta,\pm}; L_v) - \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma)\Big| \leq C_0\sqrt{t/n}\Big\}\Big) \\ \geq 1 - C_0 e^{-t/C_0}. \tag{9.19}$$

Note by the strong convexity of $L_{\eta,\pm}(\cdot; L_v)$ with respect to $\|\cdot\|_{\Sigma^{-1}}$, we have

$$\inf_{w\in\mathbb{R}^n: \|w - w_{n,\eta,\pm}\|_{\Sigma^{-1}} \geq \sqrt{6C_0}(t/n)^{1/4}} L_{\eta,\pm}(w; L_v) - L_{\eta,\pm}(w_{n,\eta,\pm}; L_v) \geq 3C_0\sqrt{t/n}.$$

This means on $E_3(t)$,

$$\inf_{w\in\mathbb{R}^n: \|w - w_{n,\eta,\pm}\|_{\Sigma^{-1}} \geq \sqrt{6C_0}(t/n)^{1/4}} L_{\eta,\pm}(w; L_v) \geq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + 2C_0\sqrt{t/n},$$
$$L_{\eta,\pm}(w_{n,\eta,\pm}; L_v) \leq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + C_0\sqrt{t/n}.$$

This in particular means on $E_1 \cap E_3(t)$,

$$w_{n,\eta,\pm} \in \Big\{w \in \mathbb{R}^n : L_{\eta,\pm}(w; L_v) \leq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + C_0\sqrt{t/n}\Big\} \\ \subset \{w \in \mathbb{R}^n : \|w\|_{\Sigma^{-1}} \leq \sqrt{6C_0}(t/n)^{1/4} + C_0/2\}.$$

Consequently, by enlarging $C_0 > 0$ if necessary, using Lemma 9.7-(1), on $E_1 \cap E_3(t)$

$$\Big\{w \in \mathbb{R}^n : L(w; L_v) \leq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + C_0\sqrt{t/n}\Big\} \\ \subset \{w \in \mathbb{R}^n : \|w\|_{\Sigma^{-1}} \leq 3C_0/5\} \subset B_n(3C_0/4) \subsetneq B_n(C_0) = B_n(L_w).$$

This implies, on $E_1 \cap E_3(t)$, we have $\|w_{n,\eta}\|_{\Sigma^{-1}} \vee \|w_{n,\eta}\| \leq 3C_0/4$, proving the apriori bound.

(**Step 2**). Next we establish the announced error bound. On the event $\mathscr{E}_{1,0}(\sqrt{t/n})$, by Lemma 9.7-(2),

$$\sup_{w\in B_n(C_0)} \Big|L_\eta(w; L_v) - L_{\eta,\pm}(w; L_v)\Big| \leq C_1\sqrt{t/n}. \tag{9.20}$$

Consequently, on $E_1 \cap E_3(t) \cap \mathscr{E}_{1,0}(\sqrt{t/n})$,

$$\Big|\min_{w\in\mathbb{R}^n} L_\eta(w; L_v) - \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma)\Big| \leq C_2\sqrt{t/n}. \tag{9.21}$$

On this event, combining (9.20)-(9.21) with (9.19), and using again the strong convexity of $L_{\eta,+}(\cdot; L_v)$ respect to $\|\cdot\|_{\Sigma^{-1}}$, we have for $C_3 = 2\sqrt{(C_0 + C_1 + C_2)}$,

$$\inf_{w\in B_n(C_0):\|w-w_{n,\eta,+}\|_{\Sigma^{-1}}\geq C_3(t/n)^{1/4}} L_\eta(w; L_v) - \min_{w\in\mathbb{R}^n} L_\eta(w; L_v)$$

$$\geq \inf_{w\in B_n(C_0):\|w-w_{n,\eta,+}\|_{\Sigma^{-1}}\geq C_3(t/n)^{1/4}} L_{\eta,+}(w; L_v) - \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) - (C_1 + C_2)\sqrt{t/n}$$

$$\geq \inf_{w\in B_n(C_0):\|w-w_{n,\eta,+}\|_{\Sigma^{-1}}\geq C_3(t/n)^{1/4}} L_{\eta,+}(w; L_v) - L_{\eta,+}(w_{n,+}; L_v) - (C_0 + C_1 + C_2)\sqrt{t/n}$$

$$\geq (C_3^2/2)\sqrt{t/n} - (C_0 + C_1 + C_2)\sqrt{t/n} = (C_0 + C_1 + C_2)\sqrt{t/n}.$$

This means that $\|w_{n,\eta} - w_{n,\eta,+}\|_{\Sigma^{-1}} \leq C_3(t/n)^{1/4}$ on $E_1 \cap E_3(t) \cap \mathscr{E}_{1,0}(\sqrt{t/n})$. The claim follows by intersecting the prescribed event with $E_2(t)$ in (9.18) that controls the $\mathbb{P}^\xi$-probability of $\|w_{n,\eta,+} - w_{\eta,*}\|_{\Sigma^{-1}} \leq C_0(t/n)^{1/4}$. □

Proof of Theorem 9.6. Fix $\xi \in \mathscr{E}_{1,\xi}(\sqrt{t/n})$, and $\varepsilon > 0$ to be chosen later on. First, as g is Lipschitz with respect to $\|\cdot\|_{\Sigma^{-1}}$, by the Gaussian concentration inequality, there exists $C_0 = C_0(K) > 0$ such that for $t \geq 1$, on an event $E_0(t)$ with $\mathbb{P}^\xi$-probability at least $1 - e^{-t}$,

$$|\mathsf{g}(w_{\eta,*}) - \mathbb{E}\,\mathsf{g}(w_{\eta,*})| \leq C_0\sqrt{t/n}.$$

Moreover, by Proposition 9.8 and Propositions 9.3-9.5, there exist some $C_1, C_1' > 0$ depending on K such that for $C_1' \log(en) \leq t \leq n/C_1'$, on an event $E_1(t)$ with $\mathbb{P}^\xi$-probability $1 - C_1 e^{-t/C_1}$, we have

(1) $\|w_{n,\eta,-}\|_{\Sigma^{-1}} \vee \|w_{n,\eta,-}\| \leq C_1$, $\|w_{n,\eta,-} - w_{\eta,*}\|_{\Sigma^{-1}} \leq C_1(t/n)^{1/4}$, and
(2) $|L_{\eta,-}(w_{n,\eta,-}; L_v) - \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma)| \leq C_1\sqrt{t/n}$.

Consequently, for $C_1' \log(en) \leq t \leq n/C_1'$, on the event $E_0(t) \cap E_1(t)$, uniformly in $w \in D_{\eta;\varepsilon}(\mathsf{g}) \cap B_n(L_w)$,

$$\begin{aligned}
\varepsilon &\leq |\mathsf{g}(w) - \mathbb{E}\,\mathsf{g}(w_{\eta,*})| \\
&\leq |\mathsf{g}(w) - \mathsf{g}(w_{\eta,*})| + |\mathsf{g}(w_{\eta,*}) - \mathbb{E}\,\mathsf{g}(w_{\eta,*})| \\
&\leq \|w - w_{\eta,n,-}\|_{\Sigma^{-1}} + \|w_{\eta,n,-} - w_{\eta,*}\|_{\Sigma^{-1}} + C_0\sqrt{t/n} \\
&\leq \|w - w_{\eta,n,-}\|_{\Sigma^{-1}} + (C_0 + C_1)(t/n)^{1/4}.
\end{aligned}$$

This implies that, for the prescribed range of t and on the event $E_0(t) \cap E_1(t)$,

$$\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_n(L_w)} \|w - w_{\eta,n,-}\|_{\Sigma^{-1}} \geq \big(\varepsilon - (C_0 + C_1)(t/n)^{1/4}\big)_+.$$

Using the strong convexity of $L_{\eta,-}(\cdot; L_v)$ with respect to $\|\cdot\|_{\Sigma^{-1}}$, we have for $C_1' \log(en) \leq t \leq n/C_1'$, on the event $E_0(t) \cap E_1(t)$,

$$\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_n(L_w)} L_\eta(w; L_v) \geq \min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_n(L_w)} L_{\eta,-}(w; L_v)$$

$$\geq L_{\eta,-}(w_{\eta,n,-}; L_v) + \frac{1}{2}\big(\varepsilon - (C_0 + C_1)(t/n)^{1/4}\big)_+^2$$

$$\geq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + \frac{1}{2}\big(\varepsilon - (C_0 + C_1)(t/n)^{1/4}\big)_+^2 - C_1\sqrt{t/n}.$$

Now we may choose $\varepsilon \equiv \varepsilon(t,n) \equiv (C_0 + C_1 + 2\sqrt{C_1})(t/n)^{1/4}$ to conclude by adjusting constants. $\square$

9.4. **Proof of Theorem 2.3 for $\widehat{\mu}_{\eta;G}$.** Fix $\xi \in \mathscr{E}_{1,\xi}(\sqrt{t/n})$. All the constants in $\lesssim, \gtrsim, \asymp$ below may depend on K.

(**Step 1**). In this step, we will obtain an upper bound $\min_{w\in\mathbb{R}^n} H_\eta(w)$. By Proposition 9.1 and the concentration estimate in Lemma 7.6, there exists some $C_0 = C_0(K) > 0$ such that on an event E_0 with $\mathbb{P}^\xi(E_0) \geq 1 - C_0 e^{-n/C_0}$,

$$\min_{w\in\mathbb{R}^n} H_\eta(w) = \min_{w\in\mathbb{R}^n} H_\eta(w; L_0) = \min_{w\in B_n(L_0)} H_\eta(w) = \min_{w\in B_n(L_0)} H_\eta(w; L_0). \tag{9.22}$$

where

$$L_0 \equiv C_0\Big\{1 + \Big(\|\Sigma^{-1}\|_{\mathrm{op}} \mathbf{1}^{-1}_{\phi^{-1}\geq 1+1/K} \wedge \eta^{-1}\Big)\Big\}. \tag{9.23}$$

Now we shall apply the convex(-side) Gaussian min-max theorem to obtain an upper bound for the right hand side of (9.22). Recall the definition of $h_\eta = h_{\eta;G}$ and ℓ_η in (6.2). Using Theorem 6.1-(2), for any $z \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_\eta(w) \geq z\Big) &\leq \mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} H_\eta(w; L_0) \geq z\Big) + \mathbb{P}^\xi(E_0^c)\\
&= \mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} \max_{v\in B_m(L_0)} h_\eta(w,v) \geq z\Big) + \mathbb{P}^\xi(E_0^c)\\
&\leq 2\,\mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} \max_{v\in B_m(L_0)} \ell_\eta(w,v) \geq z\Big) + \mathbb{P}^\xi(E_0^c)\\
&= 2\,\mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} L_\eta(w; L_0) \geq z\Big) + \mathbb{P}^\xi(E_0^c).
\end{aligned} \tag{9.24}$$

By Proposition 9.9, there exist some $C_1, C_1' > 0$ depending on K (which we assume without loss of generality $L_0 > C_1$ and C_1 exceeds the constants in Theorems 9.2 and 9.6), such that on an event E_1 with $\mathbb{P}^\xi$-probability at least $1 - C_1 e^{-n/C_1}$, the map $w \mapsto L_\eta(w; L_0)$ attains its global minimum in $B_n(C_1)$. We may now apply Theorem 9.2: with $z \equiv \bar{z}(t) = \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + \sqrt{t/n}$, for $C_1'\log(en) \leq t \leq n/C_1'$,

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} L_\eta(w; L_0) \geq \bar{z}(t)\Big)\\
&\leq \mathbb{P}^\xi\Big(\min_{w\in B_n(C_1)} L_\eta(w; L_0) \geq \bar{z}(t)\Big) + \mathbb{P}^\xi(E_1^c) \leq C_1 e^{-t/C_1} + \mathbb{P}^\xi(E_1^c).
\end{aligned} \tag{9.25}$$

Combining (9.24)-(9.25), by enlarging C_1 if necessary, for $C_1'\log(en) \leq t \leq n/C_1'$, and $\eta \in \Xi_K$,

$$\mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_\eta(w) \geq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + \sqrt{t/n}\Big) \leq C_1 e^{-t/C_1}. \tag{9.26}$$

An entirely similar argument leads to a lower bound (which will be used later on):

$$\mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_\eta(w) \leq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) - \sqrt{t/n}\Big) \leq C_1 e^{-t/C_1}. \tag{9.27}$$

(**Step 2**). In this step, we will obtain a lower bound on $\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})} H_\eta(w)$ for the exceptional set $D_\varepsilon(\mathsf{g})$ defined in (9.15), with a suitable choice of ε. Let us take $C_2, C_2' > 0$ to be the constants in Theorem 9.6, and let $\varepsilon(t,n) \equiv C_2(t/n)^{1/4}$ for

$C_2' \log(en) \le t \le n/C_2'$. To this end, using Theorem 6.1-(1) (that holds without convexity), for any $z \in \mathbb{R}$ and $L_v > 0$

$$\begin{aligned}
\mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)\cap D_{\eta;\varepsilon}(\mathsf{g})} H_\eta(w) \le z\Big) &\le \mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)\cap D_{\eta;\varepsilon}(\mathsf{g})} \max_{v\in B_m(L_v)} h_\eta(w,v) \le z\Big)\\
&\le 2\,\mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)\cap D_{\eta;\varepsilon}(\mathsf{g})} \max_{v\in B_m(L_v)} \ell_\eta(w,v) \le z\Big)\\
&= 2\,\mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)\cap D_{\eta;\varepsilon}(\mathsf{g})} L_\eta(w;L_v) \le z\Big).
\end{aligned}$$

By choosing $L_v \asymp 1$ of constant order but large enough, $\varepsilon \equiv \varepsilon(t,n)$ and $z \equiv \bar{z}(t) = \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + 2\sqrt{t/n}$, we have for $C_2'\log(en) \le t \le n/C_2'$,

$$\mathbb{P}^\xi\bigg(\min_{w\in B_n(L_0)\cap D_{\eta;\varepsilon(t,n)}(\mathsf{g})} H_\eta(w) \le \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + 2\sqrt{\frac{t}{n}}\bigg) \le 2C_2 e^{-t/C_2}. \tag{9.28}$$

(**Step 3**). Combining (9.28) and the localization in (9.22), there exist some $C_3, C_3' > 0$ depending on K such that for $C_3'\log(en) \le t \le n/C_3'$, on an event $E_3(t)$ with $\mathbb{P}^\xi(E_3(t)) \ge 1 - C_3 e^{-t/C_3}$,

$$\begin{aligned}
&\min_{w\in B_n(L_0)\cap D_{\eta;\varepsilon(t,n)}(\mathsf{g})} H_\eta(w) \ge \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + 2\sqrt{t/n}\\
&\qquad > \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + \sqrt{t/n} \ge \min_{w\in\mathbb{R}^n} H_\eta(w) = \min_{w\in B_n(L_0)} H_\eta(w).
\end{aligned}$$

So on $E_3(t)$, $\widehat{w}_\eta \notin D_{\eta;\varepsilon(t,n)}(\mathsf{g}) \cap B_n(L_0)$, i.e., for $C_3'\log(en) \le t \le n/C_3'$,

$$\mathbb{P}^\xi\Big(|\mathsf{g}(\widehat{w}_\eta) - \mathbb{E}\,\mathsf{g}(w_{\eta,*})| \ge C_3(t/n)^{1/4}\Big) \le C_3 e^{-t/C_3}.$$

Using a change of variable and suitably adjusting the constant C_3, for any 1-Lipschitz function $\mathsf{g}_0 : \mathbb{R}^n \to \mathbb{R}$, $\eta \in \Xi_K$ and $\varepsilon \in (0, 1/2]$,

$$\mathbb{P}^\xi\Big(\big|\mathsf{g}_0(\widehat{\mu}_\eta) - \mathbb{E}\,\mathsf{g}_0(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}))\big| \ge \varepsilon\Big) \le C_3 n e^{-n\varepsilon^4/C_3}.$$

(**Step 4**). In this step we shall establish uniform guarantees. We write $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) = \widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}$ in this part of the proof. First, in the case $\phi^{-1} \ge 1 + 1/K$, using $\widehat{\mu}_\eta = n^{-1}X^\top(XX^\top/n + \eta I)^{-1}Y$, for $\eta_1, \eta_2 \in [0,K]$,

$$\begin{aligned}
\|\widehat{\mu}_{\eta_1} - \widehat{\mu}_{\eta_2}\| &\lesssim n^{-1}\|G\|_{\mathrm{op}}(\|G\|_{\mathrm{op}} + \|\xi\|) \cdot \|(XX^\top/n + \eta_1 I)^{-1} - (XX^\top/n + \eta_2 I)^{-1}\|_{\mathrm{op}}\\
&\lesssim \|\Sigma^{-1}\|_{\mathrm{op}}^2 \cdot \Big(1 + \frac{\|G\|_{\mathrm{op}} + \|\xi\|}{\sqrt{n}}\Big)^2 \cdot \|(GG^\top/n)^{-1}\|_{\mathrm{op}}^2 \cdot |\eta_1 - \eta_2|.
\end{aligned} \tag{9.29}$$

Here the last inequality follows by the fact that any p.s.d. matrix A, $\|(A+\eta_1 I)^{-1} - (A + \eta_2 I)^{-1}\|_{\mathrm{op}} \le \lambda_{\min}^{-2}(A)|\eta_1 - \eta_2|$. As $\|\Sigma^{-1}\|_{\mathrm{op}} \lesssim n$ under $\mathcal{H}_\Sigma \le K$, there exists $C_4 = C_4(K) > 0$ such that on an event E_4 with $\mathbb{P}^\xi(E_4) \ge 1 - C_4 e^{-n/C_4}$,

$$\|\widehat{\mu}_{\eta_1} - \widehat{\mu}_{\eta_2}\| \le C_4 n^2 |\eta_1 - \eta_2|. \tag{9.30}$$

On the other hand, note that for $\eta_1, \eta_2 \in [0,K]$, using Proposition 8.1-(3),

$$\|\widehat{\mu}^{\mathsf{seq},*}_{\eta_1;(\Sigma,\mu_0)} - \widehat{\mu}^{\mathsf{seq},*}_{\eta_2;(\Sigma,\mu_0)}\| \lesssim (1 \vee e_g)\|\Sigma^{-1}\|_{\mathrm{op}}^2 |\eta_1 - \eta_2|. \tag{9.31}$$

So we have

$$\left|\mathbb{E}\,\mathsf{g}_0(\widehat{\mu}^{\mathsf{seq},*}_{\eta_1;(\Sigma,\mu_0)})-\mathbb{E}\,\mathsf{g}_0(\widehat{\mu}^{\mathsf{seq},*}_{\eta_2;(\Sigma,\mu_0)})\right|\leq C_4n^2|\eta_1-\eta_2|. \tag{9.32}$$

Now by taking an $\varepsilon/(2C_4n^2)$-net Λ_ε of $[0,K]$ and a union bound,

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\sup_{\eta\in[0,K]}\big|\mathsf{g}_0(\widehat{\mu}_\eta)-\mathbb{E}\,\mathsf{g}_0(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)})\big|\geq 2\varepsilon\Big)\\
&\leq \mathbb{P}^\xi\Big(\max_{\eta\in\Lambda_\varepsilon}\big|\mathsf{g}_0(\widehat{\mu}_\eta)-\mathbb{E}\,\mathsf{g}_0(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)})\big|\geq \varepsilon\Big)+\mathbb{P}(E_4^c)\\
&\leq (1+2C_4Kn^2/\varepsilon)\cdot C_3ne^{-n\varepsilon^4/C_3}+C_4e^{-n/C_4}\leq C\cdot\varepsilon^{-1}n^3e^{-n\varepsilon^4/C}. \qquad (9.33)
\end{aligned}$$

By adjusting constants, we may replace n^3/ε by n. We then conclude by further taking expectation with respect to ξ, and noting that $\mathbb{P}(\xi\in\mathscr{E}_{1,\xi}(\sqrt{t/n}))\geq 1-Ce^{-t/C}$.

Next, in the case $\phi^{-1}<1+1/K$, we work with $\eta\in[1/K,K]$ and use the standard form of $\widehat{\mu}_\eta$ with $\widehat{\mu}_\eta=n^{-1}(X^\top X/n+\eta I)^{-1}X^\top Y$. As $\eta\geq 1/K$, the spectrum of the middle inverse matrix is bounded by $1/\eta\leq K$, so we may replicate the above calculations in (9.30) and (9.32) to reach a similar estimate as in (9.33). □

9.5. **Proof of Theorem 2.3 for $\widehat{r}_{\eta;G}$.** Recall the cost function $h_\eta=h_{\eta;G},\ell_\eta$ defined in (6.2). It is easy to see that

$$\widehat{v}_\eta\equiv\arg\max_{v\in\mathbb{R}^m}\min_{w\in\mathbb{R}^n}h_\eta(w,v)=\frac{1}{\sqrt{n}\eta}(G\widehat{w}_\eta-\xi)=-\frac{\widehat{r}_\eta}{\eta}. \tag{9.34}$$

We shall define the 'population' version of $\widehat{v}_\eta$ as

$$v_{\eta,*}\equiv\frac{1}{\phi\tau_{\eta,*}}\Big(\sqrt{\phi\gamma^2_{\eta,*}-\sigma^2_\xi}\cdot\frac{h}{\sqrt{n}}-\frac{\xi}{\sqrt{n}}\Big) \tag{9.35}$$

in the Gordon problem.

Proposition 9.10. *Suppose the following hold for some $K>0$.*

- $1/K\leq\phi^{-1},\eta\leq K$, $\|\mu_0\|\vee\|\Sigma\|\vee\mathcal{H}_\Sigma\leq K$.
- *Assumption B holds with $\sigma^2_\xi\in[1/K,K]$.*

There exist constants $C,C'>0$ depending on K such that for $C'\log(en)\leq t\leq n/C'$ and $\xi\in\mathscr{E}_{1,\xi}(\sqrt{t/n})$,

$$\begin{aligned}
\mathbb{P}^\xi\Big(&\textit{The map } v\mapsto\ell_\eta(w_{\eta,*};v)\textit{ is }\eta\textit{-strongly concave with unique maximizer }v_{\eta,n}\\
&\quad\textit{satisfying }\|v_{\eta,n}\|\leq C\textit{ and }\|v_{\eta,n}-v_{\eta,*}\|\leq C\sqrt{t/n}.\\
&\quad\textit{Furthermore, }\Big|\max\nolimits_v\ell_\eta(w_{\eta,*},v)-\max\nolimits_{\beta>0}\min\nolimits_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)\Big|\leq C\sqrt{t/n}.\Big)\\
&\geq 1-Ce^{-t/C}.
\end{aligned}$$

We need the following before the proof of Proposition 9.10.

Lemma 9.11. *Suppose $1/K\leq\phi^{-1},\sigma^2_\xi\leq K$, and $\|\mu_0\|\vee\|\Sigma\|_{\mathrm{op}}\vee\mathcal{H}_\Sigma\leq K$ for some $K>0$. Recall $w_{\eta,*}$ defined in (9.14). Then there exist constants $C,C'>0$*

depending on K such that for $C'\log(en)\leq t\leq n/C'$, $\eta\in\Xi_K$ and $\xi\in\mathscr{E}_{1,\xi}(\sqrt{t/n})$,

$$\mathbb{P}^{\xi}\Big(\max\Big\{|(\mathrm{id}-\mathbb{E})\langle g/\sqrt{n},w_{\eta,*}\rangle|,\ |(\mathrm{id}-\mathbb{E})\|w_{\eta,*}\|^2|,\ |(\mathrm{id}-\mathbb{E})F(w_{\eta,*})|,$$
$$n^{-1}|(\mathrm{id}-\mathbb{E})\|\,\|w_{\eta,*}\|h-\xi\|^2|\Big\}\geq\sqrt{t/n}\Big)\leq Ce^{-t/C}.$$

Proof. All the constants in $\lesssim,\gtrsim,\asymp$ below may depend on K. Recall $w_{\eta,*}=(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^{1/2}(-\tau_{\eta,*}\mu_0+\gamma_{\eta,*}\Sigma^{1/2}g/\sqrt{n})$. Under the assumed conditions, $\gamma_{\eta,*},\tau_{\eta,*}\asymp 1$. We shall consider the four terms separately below.

For the first term, we have

$$n^{-1/2}|\langle g,w_{\eta,*}\rangle-\mathbb{E}\langle g,w_{\eta,*}\rangle|\leq\tau_{\eta,*}\cdot n^{-1/2}|\langle(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^{1/2}\mu_0,g\rangle|$$
$$+\gamma_{\eta,*}\cdot n^{-1}(\mathrm{id}-\mathbb{E})\|(\Sigma+\tau_{\eta,*}I)^{-1/2}\Sigma^{1/2}g\|^2\equiv A_{1,1}+A_{1,2}.$$

The concentration of the term $A_{1,1}$ can be handled using Gaussian tails and the fact that $\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^{1/2}\mu_0\|^2\lesssim 1$. For the term $A_{1,2}$, with $H_1(g)\equiv\|(\Sigma+\tau_{\eta,*}I)^{-1/2}\Sigma^{1/2}g\|^2$, it is easy to evaluate $\|\nabla H_1(g)\|^2=4\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma g\|^2\leq 4H_1(g)$ and $\mathbb{E}\,H_1(g)\leq n$, so Proposition B.1 applies to conclude the concentration of $A_{1,2}$.

For the second term, we may decompose

$$\Big|\|w_{\eta,*}\|^2-\mathbb{E}\,\|w_{\eta,*}\|^2\Big|\lesssim\tau_{\eta,*}\gamma_{\eta,*}\cdot n^{-1/2}|\langle(\Sigma+\tau_{\eta,*}I)^{-2}\Sigma^{3/2}\mu_0,g\rangle|$$
$$+\gamma_{\eta,*}^2\cdot n^{-1}(\mathrm{id}-\mathbb{E})\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma g\|^2.$$

From here we may handle the concentration of the above two terms in a completely similar fashion to $A_{1,1}$ and $A_{1,2}$ above.

For the third term, recall that $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)=(\Sigma+\tau I)^{-1}\Sigma^{1/2}(\Sigma^{1/2}\mu_0+\gamma g/\sqrt{n})$, so

$$\Big|F(w_{\eta,*})-\mathbb{E}\,F(w_{\eta,*})\Big|=\frac{1}{2}\Big|\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})\|^2-\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})\|^2\Big|$$
$$\lesssim\gamma_{\eta,*}\cdot n^{-1/2}|\langle(\Sigma+\tau_{\eta,*}I)^{-2}\Sigma^{3/2}\mu_0,g\rangle|+\gamma_{\eta,*}^2\cdot n^{-1}(\mathrm{id}-\mathbb{E})\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^{1/2}g\|^2.$$

The concentration properties of the two terms on the right hand side above can be handled similarly to the case for the second term.

For the last term, we have

$$n^{-1}\Big|\|\,\|w_{\eta,*}\|h-\xi\|^2-\mathbb{E}\|\,\|w_{\eta,*}\|h-\xi\|^2\Big|$$
$$\lesssim n^{-1}\Big|\|w_{\eta,*}\|^2\|h\|^2-\mathbb{E}\,\|w_{\eta,*}\|^2\|h\|^2\Big|+n^{-1}\|w_{\eta,*}\||\langle h,\xi\rangle|\equiv A_{4,1}+A_{4,2}.$$

On the other hand, on the event $\mathscr{E}_1(\sqrt{t/n})$,

$$A_{4,1}\lesssim(\|h\|^2/n)\Big|\|w_{\eta,*}\|^2-\mathbb{E}\,\|w_{\eta,*}\|^2\Big|+n^{-1}\,\mathbb{E}\,\|w_{\eta,*}\|^2\cdot|\|h\|^2-m|\lesssim\sqrt{t/n},$$

and $A_{4,2}\lesssim(1\vee e_g)\cdot n^{-1}|\langle h,\xi\rangle|\lesssim\sqrt{t/n}$. Combining the above estimates concludes the concentration claim for the last term. □

Proof of Proposition 9.10. Fix $\xi\in\mathscr{E}_{1,\xi}(\sqrt{t/n})$. All the constants in $\lesssim,\gtrsim,\asymp$ below may depend on K.

(**Step 1**). In this step, we establish both the uniqueness and the apriori estimates for $v_{\eta,n}$. Using Lemma 9.11, we may choose a sufficiently large $C,C'>0$ depending on K such that $C' \log(en) \le t \le n/C'$,

$$\mathbb{P}^{\xi}\Big(E_0(t) \equiv \Big\{\max\Big\{|(\mathrm{id}-\mathbb{E})\langle g/\sqrt{n}, w_{\eta,*}\rangle|,\ |(\mathrm{id}-\mathbb{E})\|w_{\eta,*}\|^2|,\ |(\mathrm{id}-\mathbb{E})F(w_{\eta,*})|,$$
$$n^{-1}|(\mathrm{id}-\mathbb{E})\|\,\|w_{\eta,*}\|h-\xi\|^2|\Big\} \le \sqrt{t/n}\Big\}\Big) \ge 1 - Ce^{-t/C}.$$

Therefore, on the event $E_0(t)$,

$$\langle g/\sqrt{n}, w_{\eta,*}\rangle \ge \mathbb{E}\langle g/\sqrt{n}, w_{\eta,*}\rangle - \sqrt{t/n} = \gamma_{\eta,*}\cdot n^{-1}\mathrm{tr}((\Sigma+\tau_{\eta,*}I)^{-1}\Sigma) - \sqrt{t/n}.$$

Note that $n^{-1}\mathrm{tr}((\Sigma+\tau_{\eta,*}I)^{-1}\Sigma) \gtrsim \mathcal{H}_\Sigma^{-1} \gtrsim 1$, by choosing sufficiently large C, we conclude $\langle g/\sqrt{n}, w_{\eta,*}\rangle > 0$ on the event $E_0(t)$. This implies that $v \mapsto \ell_\eta(w_{\eta,*}, v)$ is η-strongly concave with respect to $\|\cdot\|$, so $v_{\eta,n}$ exists uniquely on $E_0(t)$.

Next we derive apriori estimates. We claim that on $E_0(t)$, $v_{\eta,n} = \arg\max_{v\in\mathbb{R}^m} \ell_\eta(w_{\eta,*}, v)$ takes the following form:

$$v_{\eta,n} = \frac{1}{\sqrt{n}\eta}\bigg(1 - \frac{\langle g, w_{\eta,*}\rangle}{\|\,\|w_{\eta,*}\|h-\xi\|}\bigg)_+ \cdot (\|w_{\eta,*}\|h-\xi). \tag{9.36}$$

To see this, using the definition

$$\begin{aligned} v_{\eta,n} &= \arg\max_{v\in\mathbb{R}^m}\bigg\{\frac{1}{\sqrt{n}}\Big(-\|v\|\langle g, w_{\eta,*}\rangle + \|w_{\eta,*}\|\langle h, v\rangle - \langle v, \xi\rangle\Big) - \frac{\eta\|v\|^2}{2}\bigg\} \\ &= \arg\max_{\alpha\ge 0}\bigg\{\frac{\alpha}{\sqrt{n}}\Big(-\langle g, w_{\eta,*}\rangle + \|\,\|w_{\eta,*}\|h-\xi\|\Big) - \frac{\eta\alpha^2}{2}\bigg\}\cdot\frac{\|w_{\eta,*}\|h-\xi}{\|\,\|w_{\eta,*}\|h-\xi\|} \\ &= \frac{1}{\sqrt{n}\eta}\Big(-\langle g, w_{\eta,*}\rangle + \|\,\|w_{\eta,*}\|h-\xi\|\Big)_+ \cdot \frac{\|w_{\eta,*}\|h-\xi}{\|\,\|w_{\eta,*}\|h-\xi\|}. \end{aligned}$$

Some simple algebra leads to the expression in (9.36). The boundedness of $\|v_{\eta,n}\|$ then follows from the boundedness of $\|w_{\eta,*}\|$.

(**Step 2**). In this step, we establish the bound on $\|v_{\eta,n} - v_{\eta,*}\|$. The key observation is that we may rewrite $v_{\eta,*}$ defined via (9.35) into the following form

$$v_{\eta,*} = \frac{1}{\sqrt{n}\eta}\bigg(1 - \frac{\mathbb{E}\langle g, w_{\eta,*}\rangle}{\mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|\cdot h-\xi\|^2}\bigg)\cdot\big(\mathbb{E}^{1/2}\|w_{\eta,*}\|^2\cdot h - \xi\big). \tag{9.37}$$

This can be seen by observing

$$\begin{cases} \mathbb{E}\|w_{\eta,*}\|^2 = \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) = \phi\gamma_{\eta,*}^2 - \sigma_\xi^2, \\ \mathbb{E}\langle g, w_{\eta,*}\rangle = \frac{\sqrt{n}}{\gamma_{\eta,*}}\cdot\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) = \sqrt{n}\gamma_{\eta,*}\cdot\big(\phi - \frac{\eta}{\tau_{\eta,*}}\big), \\ \mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|\cdot h-\xi\|^2 = \sqrt{m}\big(\mathbb{E}\|w_{\eta,*}\|^2 + \sigma_\xi^2\big)^{1/2} = \sqrt{m\phi}\gamma_{\eta,*}, \end{cases} \tag{9.38}$$

and therefore $1 - \frac{\mathbb{E}\langle g, w_{\eta,*}\rangle}{\mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|\cdot h-\xi\|^2} = \frac{\eta}{\phi\tau_{\eta,*}}$. Now with (9.36)-(9.37), we may use Lemma 9.11 to estimate

$$\begin{aligned} \|v_{\eta,n} - v_{\eta,*}\| &\le \frac{1}{\sqrt{n}\eta}\Big|\|w_{\eta,*}\| - \mathbb{E}^{1/2}\|w_{\eta,*}\|^2\Big|\cdot\|h\| \\ &\quad + \frac{1}{\sqrt{n}\eta}\bigg|\frac{\langle g, w_{\eta,*}\rangle}{\|\,\|w_{\eta,*}\|h-\xi\|} - \frac{\mathbb{E}\langle g, w_{\eta,*}\rangle}{\mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|\cdot h-\xi\|^2}\bigg|\|\mathbb{E}^{1/2}\|w_{\eta,*}\|^2\cdot h - \xi\| \\ &\equiv V_1 + V_2. \end{aligned} \tag{9.39}$$

We first handle the term V_1. As $\mathbb{E}\|w_{\eta,*}\|^2 \geq \gamma_{\eta,*}^2 \mathrm{tr}(\Sigma^2(\Sigma+\tau_{\eta,*})^{-2})/n \gtrsim 1$, on the event $E_0(t)\cap \mathscr{E}_{1,0}(\sqrt{t/n})$,

$$V_1 \lesssim \frac{\|h\|}{\sqrt{n}}\cdot \frac{\big|\|w_{\eta,*}\|^2 - \mathbb{E}\,\|w_{\eta,*}\|^2\big|}{\mathbb{E}^{1/2}\,\|w_{\eta,*}\|^2} \lesssim \sqrt{t/n}. \tag{9.40}$$

Next we handle V_2. On the event $E_0(t)\cap \mathscr{E}_{1,0}(\sqrt{t/n})$,

$$\begin{aligned}
V_2 &\lesssim \|\,\|w_{\eta,*}\|h-\xi\|^{-1}\cdot \big|\langle g, w_{\eta,*}\rangle - \mathbb{E}\langle g, w_{\eta,*}\rangle\big|\\
&\qquad + \mathbb{E}\langle g, w_{\eta,*}\rangle \cdot \Big|\|\,\|w_{\eta,*}\|h-\xi\|^{-1} - \mathbb{E}^{-1/2}\|\,\|w_{\eta,*}\|\cdot h-\xi\|^2\Big|\\
&\lesssim n^{-1/2}\big|\langle g, w_{\eta,*}\rangle - \mathbb{E}\langle g, w_{\eta,*}\rangle\big| + n^{-1/2}\Big|\|\,\|w_{\eta,*}\|h-\xi\| - \mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|h-\xi\|^2\Big|\\
&\lesssim n^{-1/2}\big|\langle g, w_{\eta,*}\rangle - \mathbb{E}\langle g, w_{\eta,*}\rangle\big| + n^{-1}\Big|\|\,\|w_{\eta,*}\|h-\xi\|^2 - \mathbb{E}\|\,\|w_{\eta,*}\|h-\xi\|^2\Big|\\
&\lesssim \sqrt{t/n}.
\end{aligned} \tag{9.41}$$

The desired estimate for $\|v_{\eta,n}-v_{\eta,*}\|$ follows from (9.39)-(9.41).

(**Step 3**). In this step, we prove the claimed bound on $|\max_v \ell_\eta(w_{\eta,*},v) - \overline{\mathsf{D}}_\eta(\beta_{\eta,*},\gamma_{\eta,*})|$. First note that

$$\begin{aligned}
&\max_{v\in\mathbb{R}^m} \ell_\eta(w_{\eta,*},v)\\
&\equiv \max_{v\in\mathbb{R}^m}\bigg\{\frac{1}{\sqrt{n}}\Big(-\|v\|\langle g, w_{\eta,*}\rangle + \|w_{\eta,*}\|\langle h, v\rangle - \langle v,\xi\rangle\Big) + F(w_{\eta,*}) - \frac{\eta\|v\|^2}{2}\bigg\}\\
&= \frac{1}{2n\eta}(\|\,\|w_{\eta,*}\|h-\xi\| - \langle g, w_{\eta,*}\rangle)_+^2 + F(w_{\eta,*}).
\end{aligned} \tag{9.42}$$

On the other hand, with $\#^*_{\eta;(\Sigma,\mu_0)} \equiv \#_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})$, $\# \in \{\mathsf{err},\mathsf{dof}\}$,

$$\mathbb{E}\,\mathsf{e}_F\Big(\frac{\gamma_{\eta,*}}{\sqrt{n}}g;\frac{\gamma_{\eta,*}}{\beta_{\eta,*}}\Big) = \frac{\beta_{\eta,*}}{2\gamma_{\eta,*}}\big(\mathbb{E}\,\mathsf{err}^*_{\eta;(\Sigma,\mu_0)} - 2\,\mathbb{E}\,\mathsf{dof}^*_{\eta;(\Sigma,\mu_0)} + \gamma_{\eta,*}^2\big) + \mathbb{E}\,F(w_{\eta,*}),$$

so we may rewrite $\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) = \overline{\mathsf{D}}_\eta(\beta_{\eta,*},\gamma_{\eta,*})$ as follows:

$$\begin{aligned}
\overline{\mathsf{D}}_\eta(\beta_{\eta,*},\gamma_{\eta,*}) &= \frac{\beta_{\eta,*}}{2}\Big(\gamma_{\eta,*}(\phi-1) + \frac{\sigma_\xi^2}{\gamma_{\eta,*}}\Big) - \frac{\eta\beta_{\eta,*}^2}{2} + \mathbb{E}\,\mathsf{e}_F\Big(\frac{\gamma_{\eta,*}}{\sqrt{n}}g;\frac{\gamma_{\eta,*}}{\beta_{\eta,*}}\Big)\\
&= \frac{\beta_{\eta,*}}{2\gamma_{\eta,*}}\Big(\phi\gamma_{\eta,*}^2 + \sigma_\xi^2 + \mathbb{E}\,\mathsf{err}^*_{\eta;(\Sigma,\mu_0)} - 2\,\mathbb{E}\,\mathsf{dof}^*_{\eta;(\Sigma,\mu_0)}\Big) - \frac{\eta\beta_{\eta,*}^2}{2} + \mathbb{E}\,F(w_{\eta,*})\\
&= \frac{\beta_{\eta,*}}{\gamma_{\eta,*}}\Big(\phi\gamma_{\eta,*}^2 - \mathbb{E}\,\mathsf{dof}^*_{\eta;(\Sigma,\mu_0)}\Big) - \frac{\eta\beta_{\eta,*}^2}{2} + \mathbb{E}\,F(w_{\eta,*}).
\end{aligned}$$

Further using the second and third equations in (9.38), it now follows that

$$\begin{aligned}
\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) &= \frac{\beta_{\eta,*}}{\sqrt{n}}\Big(\mathbb{E}^{1/2}\,\|\,\|w_{\eta,*}\|h-\xi\|^2 - \mathbb{E}\langle g, w_{\eta,*}\rangle\Big) - \frac{\eta\beta_{\eta,*}^2}{2} + \mathbb{E}\,F(w_{\eta,*})\\
&= \frac{1}{2n\eta}\Big(\mathbb{E}^{1/2}\,\|\,\|w_{\eta,*}\|h-\xi\|^2 - \mathbb{E}\langle g, w_{\eta,*}\rangle\Big)^2 + \mathbb{E}\,F(w_{\eta,*}).
\end{aligned} \tag{9.43}$$

Now combining (9.42) and (9.43), on the event $E_0(t)\cap\mathscr{E}_{1,0}(\sqrt{t/n})$, we may estimate

$$\begin{aligned}&\Big|\max_{v\in\mathbb{R}^m}\ell_\eta(w_{\eta,*},v)-\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)\Big|\\&\lesssim n^{-1/2}\Big|\langle g,w_{\eta,*}\rangle-\mathbb{E}\langle g,w_{\eta,*}\rangle\Big|+n^{-1/2}\Big|\|\,\|w_{\eta,*}\|h-\xi\|-\mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|h-\xi\|^2\Big|\\&\qquad+|F(w_{\eta,*})-\mathbb{E}\,F(w_{\eta,*})|\lesssim\sqrt{t/n},\end{aligned}$$

completing the proof. □

Proof of Theorem 2.3 for $\widehat{r}_\eta$. Fix $\xi\in\mathscr{E}_{1,\xi}(\sqrt{t/n})$. All the constants in $\lesssim,\gtrsim,\asymp$ below may depend on K. We sometimes write $\overline{\mathscr{D}}_\eta\equiv\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)$.

As $\widehat{r}_\eta=-\eta\widehat{v}_\eta$, we only need to study $\widehat{v}_\eta$. Fix $\varepsilon>0$, and any $\mathsf{h}:\mathbb{R}^m\to\mathbb{R}$, let

$$D_{\eta;\varepsilon}(\mathsf{h})\equiv\{v\in\mathbb{R}^m:|\mathsf{h}(v)-\mathbb{E}^\xi\,\mathsf{h}(v_{\eta,*})|\geq\varepsilon\}.$$

(**Step 1**). In this step we establish the Gordon cost cap: there exist constants $C_1,C_1'>0$ depending on K such that for $C_1'\log(en)\leq t\leq n/C_1'$,

$$\mathbb{P}^\xi\Big(E_1(t)^c\equiv\Big\{\max_{v\in D_{\eta;C_1(t/n)^{1/4}}(\mathsf{h})}\ell_\eta(w_{\eta,*},v)\geq\overline{\mathscr{D}}_\eta-C_1^{-1}\sqrt{t/n}\Big\}\Big)\leq C_1e^{-t/C_1}.\tag{9.44}$$

To this end, first note that by the Lipschitz property of h, the Gaussian concentration and Proposition 9.10, there exist some $C_0,C_0'>0$ depending on K such that for $C_0'\log(en)\leq t\leq n/C_0'$, on an event $E_{1,0}(t)$ with probability at least $1-C_0e^{-t/C_0}$, we have uniformly in $v\in\mathsf{D}_{\eta;\varepsilon}(\mathsf{h})$,

$$\begin{aligned}\varepsilon\leq&|\mathsf{h}(v)-\mathbb{E}^\xi\,\mathsf{h}(v_{\eta,*})|\leq|\mathsf{h}(v)-\mathsf{h}(v_{\eta,*})|+|\mathsf{h}(v_{\eta,*})-\mathbb{E}^\xi\,\mathsf{h}(v_{\eta,*})|\\\leq&\|v-v_{\eta,n}\|+\|v_{\eta,*}-v_{\eta,n}\|+C\sqrt{t/n}\leq\|v-v_{\eta,n}\|+C_0\sqrt{t/n},\end{aligned}$$

and all the properties in Proposition 9.10 hold. In other word, on $E_{1,0}(t)$ with the prescribed range of t,

$$\inf_{w\in D_{\eta;\varepsilon}(\mathsf{h})}\|v-v_{\eta,n}\|\geq(\varepsilon-C_0\sqrt{t/n})_+.$$

Using the η-strong concavity of $v\mapsto\ell_\eta(w_{\eta,*},v)$ on $E_{1,0}(t)$, we have

$$\begin{aligned}\max_{v\in D_{\eta;\varepsilon}(\mathsf{h})}\ell_\eta(w_{\eta,*},v)&\leq\ell_\eta(w_{\eta,*},v_{\eta,n})-\frac{\eta}{2}\inf_{w\in\mathsf{D}_{\eta;\varepsilon}(\mathsf{h})}\|v-v_{\eta,n}\|^2\\&\leq\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)-\frac{\eta}{2}(\varepsilon-C_0\sqrt{t/n})_+^2+C_1\sqrt{t/n}.\end{aligned}$$

By choosing $\varepsilon\equiv\varepsilon_{\eta;v}(t,n)\equiv C_0\sqrt{t/n}+2\sqrt{C_1/\eta}\cdot(t/n)^{1/4}$, we have on $E_{1,0}(t)$,

$$\max_{v\in D_{\eta;\varepsilon_{\eta;v}(t,n)}(\mathsf{h})}\ell_\eta(w_{\eta,*},v)\leq\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)-C_1\sqrt{t/n}.\tag{9.45}$$

Adjusting constants proves the claim in (9.44).

(**Step 2**). In this step, we provide an upper bound for the original cost over exceptional set. More concretely, we will prove that there exist constants $C_2,C_2'>0$ depending on K such that for any $L_v>0$, and $C_2'\log(en)\leq t\leq n/C_2'$,

$$\mathbb{P}^\xi\Big(E_2(t)^c\equiv\Big\{\max_{v\in D_{\eta;C_2(t/n)^{1/4}}(\mathsf{h})\cap B_m(L_v)}\min_{w\in\mathbb{R}^n}h_\eta(w,v)$$

$$\geq \overline{\mathscr{D}}_\eta - C_2^{-1}\sqrt{t/n}\Big\}\Big) \leq C_2 e^{-t/C_2}. \tag{9.46}$$

To see this, first note by Proposition 9.9, there exists some $C_2 = C_2(K) > 0$ such that on an event $E_{2,0}$ with $\mathbb{P}^\xi(E_{2,0}) \geq 1 - C_2 e^{-n/C_2}$, $\|w_{\eta,*}\| \leq C_2$. So with $\bar{z}_{\eta;v}(t,n) \equiv \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) - C_1^{-1}\sqrt{t/n}$, for any $L_v > 0$, an application of Theorem 6.1-(1) yields that for $C_1'\log(en) \leq t \leq n/C_1'$,

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C_1(t/n)^{1/4}}(\mathsf{h})\cap B_m(L_v)}\min_{w\in\mathbb{R}^n} h_\eta(w,v) \geq \bar{z}_{\eta;v}(t,n)\Big)\\
&\leq \mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C_1(t/n)^{1/4}}(\mathsf{h})\cap B_m(L_v)}\min_{w\in B_n(C_2)} h_\eta(w,v) \geq \bar{z}_{\eta;v}(t,n)\Big)\\
&\leq 2\,\mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C_1(t/n)^{1/4}}(\mathsf{h})\cap B_m(L_v)}\min_{w\in B_n(C_2)} \ell_\eta(w,v) \geq \bar{z}_{\eta;v}(t,n)\Big)\\
&\leq 2\,\mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C_1(t/n)^{1/4}}(\mathsf{h})\cap B_m(L_v)} \ell_\eta(w_{\eta,*},v) \geq \bar{z}_{\eta;v}(t,n)\Big) + 2\,\mathbb{P}^\xi(E_{2,0}^c)\\
&\leq 2\,\mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C_1(t/n)^{1/4}}(\mathsf{h})} \ell_\eta(w_{\eta,*},v) \geq \bar{z}_{\eta;v}(t,n)\Big) + 2\,\mathbb{P}^\xi(E_{2,0}^c) \leq Ce^{-t/C},
\end{aligned}$$

proving the claim (9.44) by possibly adjusting constants.

(**Step 3**). In this step, we recall a lower bound for the original cost optimum, essentially established in the Step 1 in the proof of Theorem 2.3. In particular, using (9.22), (9.23) and (9.27), there exist $C_3, C_3', C_3'' > 0$ depending on K, such that for $C_3'\log(en) \leq t \leq n/C_3'$,

$$\mathbb{P}^\xi\Big(E_{3,0}(t)^c \equiv \Big\{\max_{v\in B_m(C_3'')}\min_{w\in\mathbb{R}^n} h_\eta(w,v) \leq \overline{\mathscr{D}}_\eta - C_3^{-1}\sqrt{t/n}\Big\}\Big) \leq C_3 e^{-t/C_3}, \tag{9.47}$$

and

$$\mathbb{P}^\xi\Big(E_{3,1}^c \equiv \Big\{\max_{v\in B_m(C_3'')}\min_{w\in\mathbb{R}^n} h_\eta(w,v) = \max_{v\in\mathbb{R}^m}\min_{w\in\mathbb{R}^n} h_\eta(w,v)\Big\}\Big) \leq C_3 e^{-n/C_3}. \tag{9.48}$$

(**Step 4**). By choosing without loss of generality $C_3 > C_2$, on the event $E_2(t) \cap E_{3,0}(t) \cap E_{3,1}$, (9.46)-(9.48) yield that for any $C'\log(en) \leq t \leq n/C'$,

$$\begin{aligned}
\max_{v\in \mathsf{D}_{\eta;C_2(t/n)^{1/4}}(\mathsf{h})\cap B_m(C_3'')}\min_{w\in\mathbb{R}^n} h_\eta(w,v) &\leq \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) - C_2^{-1}\sqrt{t/n}\\
&< \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) - C_3^{-1}\sqrt{t/n}\\
&\leq \max_{v\in B_m(C_3'')}\min_{w\in\mathbb{R}^n} h_\eta(w,v) = \max_{v\in\mathbb{R}^m}\min_{w\in\mathbb{R}^n} h_\eta(w,v).
\end{aligned}$$

This means on the event $E_2(t) \cap E_{3,0}(t) \cap E_{3,1}$, $\widehat{v}_\eta \notin \mathsf{D}_{\eta;C_2(t/n)^{1/4}}(\mathsf{h})$, i.e., there exist some $C_4, C_4' > 0$ depending on K such that for $C_4'\log(en) \leq t \leq n/C_4'$ and $1/K \leq \eta \leq K$,

$$\mathbb{P}^\xi\Big(|\mathsf{h}(\widehat{v}_\eta) - \mathbb{E}^\xi\,\mathsf{h}(v_{\eta,*})| \geq C_4(t/n)^{1/4}\Big) \leq C_4 e^{-t/C_4}. \tag{9.49}$$

(**Step 5**). In this final step, we shall prove uniform version of the estimate (9.49). For $\eta_1, \eta_2 \in [1/K, K]$, using the definition of $\widehat{v}_\eta$ in (9.34),

$$|\mathsf{h}(\widehat{v}_{\eta_1}) - \mathsf{h}(\widehat{v}_{\eta_2})| \leq \|\widehat{v}_{\eta_1} - \widehat{v}_{\eta_2}\|$$

$$\leq n^{-1/2}\big\|\eta_1^{-1}G\widehat{w}_{\eta_1}-\eta_2^{-1}G\widehat{w}_{\eta_2}\big\|+(\|\xi\|/\sqrt{n})\cdot|\eta_1^{-1}-\eta_2^{-1}|$$
$$\leq \frac{\|G\widehat{w}_{\eta_1}\|+\|\xi\|}{\sqrt{n}}\cdot|\eta_1^{-1}-\eta_2^{-1}|+\frac{1}{\sqrt{n}\eta_2}\big\|G(\widehat{w}_{\eta_1}-\widehat{w}_{\eta_2})\big\|$$
$$\lesssim\Big(1+\|\widehat{\mu}_{\eta_1}\|\frac{\|G\|_{\mathrm{op}}}{\sqrt{n}}\Big)\cdot|\eta_1-\eta_2|+\frac{\|G\|_{\mathrm{op}}}{\sqrt{n}}\cdot\|\widehat{\mu}_{\eta_1}-\widehat{\mu}_{\eta_2}\|.$$

Using that $\|\widehat{\mu}_\eta\|=\|n^{-1}(X^\top X/n+\eta I)^{-1}X^\top Y\|\leq\|X^\top Y\|/(n\eta)\lesssim(1+\|G\|_{\mathrm{op}}/\sqrt{n})^2$, we have

$$|\mathsf{h}(\widehat{v}_{\eta_1})-\mathsf{h}(\widehat{v}_{\eta_2})|\lesssim(1+\|G\|_{\mathrm{op}}/\sqrt{n})^3\cdot(|\eta_1-\eta_2|\vee\|\widehat{\mu}_{\eta_1}-\widehat{\mu}_{\eta_2}\|). \tag{9.50}$$

In view of (9.30), there exists some $C_5>0$ depending on K, such that on an event $E_{5,1}$ with $\mathbb{P}^\xi(E_{5,1})\geq 1-C_5e^{-n/C_5}$,

$$|\mathsf{h}(\widehat{v}_{\eta_1})-\mathsf{h}(\widehat{v}_{\eta_2})|\leq C_5n^2|\eta_1-\eta_2|. \tag{9.51}$$

On the other hand, using the definition of $v_{\eta,*}$ in (9.35), Proposition 8.1-(3) and the fact that $\phi\gamma_{\eta,*}^2-\sigma_\xi^2=\mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})\geq\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma^2)\gtrsim 1$, we have

$$\Big|\mathbb{E}^\xi\,\mathsf{h}(v_{\eta_1,*})-\mathbb{E}^\xi\,\mathsf{h}(v_{\eta_2,*})\Big|\leq\mathbb{E}^{1/2,\xi}\|v_{\eta_1,*}-v_{\eta_2,*}\|^2$$
$$\lesssim\Big|\tau_{\eta_1,*}^{-1}\sqrt{\phi\gamma_{\eta_1,*}^2-\sigma_\xi^2}-\tau_{\eta_2,*}^{-1}\sqrt{\phi\gamma_{\eta_2,*}^2-\sigma_\xi^2}\Big|+\Big|\tau_{\eta_1,*}^{-1}-\tau_{\eta_2,*}^{-1}\Big|$$
$$\lesssim\Big|\gamma_{\eta_1,*}^2-\gamma_{\eta_2,*}^2\Big|+\Big|\tau_{\eta_1,*}^{-1}-\tau_{\eta_2,*}^{-1}\Big|\leq C_5|\eta_1-\eta_2|. \tag{9.52}$$

Now we may mimic the proof in (9.33) to conclude that, by possibly enlarging $C_5>0$, for any $\varepsilon\in(0,1/2]$ and $\xi\in\mathscr{E}_{1,\xi}(\varepsilon^2/C_5)$,

$$\mathbb{P}^\xi\Big(\sup_{\eta\in[1/K,K]}|\mathsf{h}(\widehat{v}_\eta)-\mathbb{E}^\xi\,\mathsf{h}(v_{\eta,*})|\geq\varepsilon\Big)\leq C_5ne^{-n\varepsilon^4/C_5},$$

as desired. □

10. Universality: Proof of Theorem 2.4

10.1. **Comparison inequalities.** For $\mathsf{f}:\mathbb{R}^n\to\mathbb{R}$, let

$$\mathcal{H}_{\mathsf{f}}(w,A)\equiv\frac{1}{2n}\|Aw-\xi\|^2+\mathsf{f}(w).$$

The following theorem is proved in [HS22, Theorem 2.3].

Theorem 10.1. *Suppose $1/K\leq\phi^{-1}\leq K$ for some $K>1$. Let $A_0,B_0\in\mathbb{R}^{m\times n}$ be two random matrices with independent components, such that $\mathbb{E}\,A_{0;ij}=\mathbb{E}\,B_{0;ij}=0$ and $\mathbb{E}\,A_{0;ij}^2=\mathbb{E}\,B_{0;ij}^2$ for all $i\in[m]$, $j\in[n]$. Further assume that*

$$M\equiv\max_{i\in[m],j\in[n]}\big(\mathbb{E}|A_{0;ij}|^6+\mathbb{E}|B_{0;ij}|^6\big)<\infty.$$

Let $A\equiv A_0/\sqrt{n}$ and $B\equiv B_0/\sqrt{n}$. Then there exists some $C_0=C_0(K,M)>0$ such that the following hold: For any $\mathcal{S}_n\subset[-L_n,L_n]^n$ with $L_n\geq 1$, and any $\mathsf{T}\in C^3(\mathbb{R})$, we have

$$\Big|\mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_n}\mathcal{H}_{\mathsf{f}}(w,A)\Big)-\mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_n}\mathcal{H}_{\mathsf{f}}(w,B)\Big)\Big|\leq C_0\cdot K_{\mathsf{T}}\cdot\mathsf{r}_{\mathsf{f}}(L_n).$$

Here $K_{\mathsf{T}} \equiv 1 + \max_{\ell\in[0:3]} \|\mathsf{T}^{(\ell)}\|_\infty$, *and* $\mathsf{r}_{\mathsf{f}}(L_n)$ *is defined by*

$$\mathsf{r}_{\mathsf{f}}(L_n) \equiv \inf_{\delta\in(0,n^{-5/2})} \Big\{ \mathscr{N}_{\mathsf{f}}(L_n,\delta) + \Big(1 + \frac{1}{m}\sum_{i=1}^m \mathbb{E}|\xi_i|^3\Big)^{1/3} \cdot \frac{L_n^2 \log_+^{2/3}(L_n/\delta)}{n^{1/6}} \Big\},$$

where $\mathscr{N}_{\mathsf{f}}(L_n,\delta) \equiv \sup |\mathsf{f}(w) - \mathsf{f}(w')|$ *with the supremum taken over all* $w, w' \in [-L_n, L_n]^n$ *such that* $\|w - w'\|_\infty \le \delta$. *Consequently, for any* $z \in \mathbb{R}, \varepsilon > 0$,

$$\mathbb{P}\Big(\min_{w\in\mathcal{S}_n} \mathcal{H}_{\mathsf{f}}(w,A) > z + 3\varepsilon\Big) \le \mathbb{P}\Big(\min_{w\in\mathcal{S}_n} \mathcal{H}_{\mathsf{f}}(w,B) > z + \varepsilon\Big) + C_1(1\vee\varepsilon^{-3})\mathsf{r}_{\mathsf{f}}(L_n).$$

Here $C_1 > 0$ *is an absolute multiple of* C_0.

Let for $u \in \mathbb{R}^m, w \in \mathbb{R}^n, A \in \mathbb{R}^{m\times n}$ and a measurable function $Q : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$

$$X(u,w;A) \equiv u^\top A w + Q(u,w). \tag{10.1}$$

The following theorem is proved in [HS22, Theorem 2.5].

Theorem 10.2. *Let* $A, B \in \mathbb{R}^{m\times n}$ *be two random matrices with independent entries and matching first two moments, i.e.,* $\mathbb{E}\,A_{ij}^\ell = \mathbb{E}\,B_{ij}^\ell$ *for all* $i \in [m], j \in [n], \ell = 1, 2$. *There exists a universal constant* $C_0 > 0$ *such that the following hold. For any measurable subsets* $\mathcal{S}_u \subset [-L_u, L_u]^m$, $\mathcal{S}_w \subset [-L_w, L_w]^n$ *with* $L_u, L_w \ge 1$, *and any* $\mathsf{T} \in C^3(\mathbb{R})$, *we have*

$$\begin{aligned}&\Big|\mathbb{E}\,\mathsf{T}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X(u,w;A)\Big) - \mathbb{E}\,\mathsf{T}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X(u,w;B)\Big)\Big|\\&\le C_0 \cdot K_{\mathsf{T}} \cdot \inf_{\delta\in(0,1)} \Big\{ M_1 L\delta + \mathscr{N}_Q(L,\delta) + \log_+^{2/3}(L/\delta)\cdot(m+n)^{2/3} M_3^{1/3} L^2 \Big\}.\end{aligned}$$

Here $K_{\mathsf{T}} \equiv 1 + \max_{\ell\in[0:3]} \|\mathsf{T}^{(\ell)}\|_\infty$, $L \equiv L_u + L_w$, $M_\ell \equiv \sum_{i\in[m],j\in[n]} \big(\mathbb{E}|A_{ij}|^\ell + \mathbb{E}|B_{ij}|^\ell\big)$, *and* $\mathscr{N}_Q(L,\delta) \equiv \sup|Q(u,w) - Q(u',w')|$ *with the supremum taken over all* $u, u' \in [-L, L]^m, w, w' \in [-L, L]^n$ *such that* $\|u - u'\|_\infty \vee \|w - w'\|_\infty \le \delta$. *The conclusion continues to hold when max-min is flipped to min-max.*

10.2. **Delocalization.** Recall that $\widehat{\mu}_\eta$ defined in (1.3) can be rewritten as

$$\widehat{\mu}_\eta = \arg\min_{\mu\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} \Big\{ \frac{1}{2}\|\mu\|^2 + \frac{1}{\sqrt{n}}\langle v, X\mu - Y\rangle - \frac{\eta}{2}\|v\|^2 \Big\}.$$

For any $\eta > 0$, we have the following closed form for $\widehat{\mu}_\eta$:

$$\widehat{\mu}_\eta = n^{-1}(X^\top X/n + \eta I_n)^{-1} X^\top Y, \quad \widehat{v}_\eta = -(\sqrt{n}\eta)^{-1}(Y - X\widehat{\mu}_\eta). \tag{10.2}$$

The above formula does not include the interpolating case $\eta = 0$ when $n > m$. To give an alternative expression, note that the first-order condition for the above minimax optimization is $\widehat{\mu}_\eta = X^\top \widehat{v}_\eta/\sqrt{n}$, $Y - X\widehat{\mu}_\eta = -\sqrt{n}\eta\widehat{v}_\eta$, or equivalently,

$$\widehat{\mu}_\eta = n^{-1} X^\top (XX^\top/n + \eta I_m)^{-1} Y, \quad \widehat{v}_\eta = -n^{-1/2}(XX^\top/n + \eta I_m)^{-1} Y. \tag{10.3}$$

The following proposition proves delocalization for $\widehat{w}_\eta \equiv \Sigma^{1/2}(\widehat{\mu}_\eta - \mu_0)$ and $\widehat{v}_\eta$.

Proposition 10.3. *Suppose Assumption A holds and the following hold for some* $K > 0$.

- $1/K \le \phi^{-1} \le K$, $\|\Sigma^{-1}\|_{\mathrm{op}} \vee \|\Sigma\|_{\mathrm{op}} \le K$.

- *Assumption B holds with* $\sigma_\xi^2 \in [1/K, K]$.

Fix $\vartheta \in (0, 1/2]$. *Then there exist some constant* $C = C(K, \vartheta) > 0$*, two measurable sets* $\mathcal{U}_\vartheta \subset B_n(1), \mathcal{E}_\vartheta \subset \mathbb{R}^m$ *with* $\min\{\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)), \mathbb{P}(\xi \in \mathcal{E}_\vartheta)\} \geq 1 - Ce^{-n^{2\vartheta}/C}$*, such that*

$$\sup_{\mu_0 \in \mathcal{U}_\vartheta, \xi \in \mathcal{E}_\vartheta} \mathbb{P}^\xi\Big(\sup_{\eta \in \Xi_K}\Big\{\|\widehat{w}_\eta\|_\infty \vee \|\widehat{v}_\eta\|_\infty\Big\} \geq Cn^{-1/2+\vartheta}\Big) \leq Cn^{-100}.$$

The sets $\mathcal{U}_\vartheta, \mathcal{E}_\vartheta$ *can be taken as*

$$\mathcal{U}_\vartheta \equiv \Big\{\mu_0 \in B_n(1) : \sup_{\eta \in \Xi_K}\Big\|\Sigma^{1/2}\big(\mathbb{E}\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0\big)\Big\|_\infty \leq C_0 n^{-1/2+\vartheta}\Big\},$$

$$\mathcal{E}_\vartheta \equiv \Big\{\xi \in \mathbb{R}^m : \|\xi\|_\infty \leq C_0 n^\vartheta, \Big|\|\xi\|^2/m - \sigma_\xi^2\Big| \leq C_0 n^{-1/2+\vartheta}\Big\}$$

for some large enough $C_0 = C_0(K) > 0$.

Remark 6. Delocalization in the same sense of the above proposition holds for $\|\mathsf{P}\widehat{\mu}_\eta + \mathsf{q}\|_\infty$ with any deterministic matrix $\mathsf{P} \in \mathbb{R}^{n\times n}$ and vector $\mathsf{q} \in \mathbb{R}^n$ satisfying $\|\mathsf{P}\|_{\mathrm{op}} \vee \|\mathsf{q}\| \leq 1$, with a (slightly) different construction of $\mathcal{U}_\vartheta$.

Proof of Proposition 10.3. All the constants in $\lesssim, \gtrsim, \asymp$ below may depend on K.
(1). Let us consider delocalization for $\widehat{w}_\eta$. Using (10.3), for any $s \in [n]$,

$$\begin{aligned}\langle e_s, \widehat{w}_\eta\rangle &= n^{-1}\langle \Sigma^{1/2}e_s, X^\top(\phi\check{\Sigma} + \eta I_m)^{-1}X\mu_0\rangle - \langle\Sigma^{1/2}e_s, \mu_0\rangle\\ &\quad + n^{-1}\langle\Sigma^{1/2}e_s, X^\top(\phi\check{\Sigma} + \eta I_m)^{-1}\xi\rangle \equiv A_{1;s} + A_{2;s}.\end{aligned}\tag{10.4}$$

We first handle $A_{1;s}$. Let ρ be the asymptotic eigenvalue density of $\check{\Sigma} = XX^\top/m$ and fix $c > 0$. By [KY17, Theorem 3.16-(i), Remark 3.17 and Lemma 4.4-(i)], for any small $\vartheta > 0$ and large $D > 0$,

$$\begin{aligned}\mathbb{P}^\xi\Big(\Big|&m^{-1}\langle\Sigma^{1/2}e_s, X^\top(\check{\Sigma} - zI_m)^{-1}X\mu_0\rangle\\ &- \langle\Sigma^{1/2}e_s, \mathfrak{m}(z)\Sigma(I_n + \mathfrak{m}(z)\Sigma)^{-1}\mu_0\rangle\Big| \geq n^{-1/2+\vartheta}\sqrt{\Im\mathfrak{m}(z)/\Im z}\Big) \leq Cn^{-D}\end{aligned}$$

holds for all $z \in [-1/c, 1/c] \times (0, 1/c]$. With $\kappa \equiv \kappa(z) \equiv \mathrm{dist}(\Re z, \mathrm{supp}\,\rho) \geq n^{-2/3+c}$, by further using the simple relation $\Im m(z)/\Im z = \int\frac{\rho(\mathrm{d}x)}{(\Re z - x)^2 + \Im^2 z} \leq \kappa^{-2}$, the error bound $n^{-1/2+\vartheta}\sqrt{\Im\mathfrak{m}(z)/\Im z}$ in the above display can be replaced by $\kappa^{-1}n^{-1/2+\vartheta}$.

When $\phi^{-1} \geq 1 + 1/K$, according to [BS10, Theorem 6.3-(2)], $\mathrm{supp}\,\rho \in (C_0^{-1}, C_0)$ for some constant $C_0 > 1$. Therefore, for $z \equiv z(b) \equiv -\eta/\phi + \sqrt{-1}b$ with a small enough $b > 0$ to be chosen later, it is easy to see that $\kappa \geq \kappa_0 \equiv (\eta/\phi) \vee C_0^{-1}\mathbf{1}_{\phi^{-1}\geq 1+1/K}$. Therefore, on an event $E_{1,0;s}(b)$ with $\mathbb{P}^\xi(E_{1,0;s}(b)) \geq 1 - Cn^{-D}$,

$$\begin{aligned}\Big|&m^{-1}\langle\Sigma^{1/2}e_s, X^\top(\check{\Sigma} - z(0)I_m)^{-1}X\mu_0\rangle\\ &- \langle\Sigma^{1/2}e_s, \mathfrak{m}(z(0))\Sigma(I_n + \mathfrak{m}(z(0))\Sigma)^{-1}\mu_0\rangle\Big| \leq (I) + (II) + \kappa_0^{-1}n^{-1/2+\vartheta},\end{aligned}\tag{10.5}$$

where

- $(I) = |m^{-1}\langle\Sigma^{1/2}e_s, X^\top(\check{\Sigma} - z(b)I_m)^{-1}X\mu_0\rangle - m^{-1}\langle\Sigma^{1/2}e_s, X^\top(\check{\Sigma} - z(0)I_m)^{-1}X\mu_0\rangle|$,
- $(II) = |\langle\Sigma^{1/2}e_s, \mathfrak{m}(z(b))\Sigma(I_n + \mathfrak{m}(z(b))\Sigma)^{-1}\mu_0\rangle - \langle\Sigma^{1/2}e_s, \mathfrak{m}(z(0))\Sigma(I_n + \mathfrak{m}(z(0))\Sigma)^{-1}\mu_0\rangle|$.

By a derivative calculation, it is easy to derive

$$(I) \lesssim (\|Z\|_{\mathrm{op}}/\sqrt{n})^2 \cdot (\|(ZZ^\top/n)^{-1}\|_{\mathrm{op}} \mathbf{1}_{\phi^{-1}\geq 1+1/K} \wedge \eta^{-1})^2 \cdot b.$$

Now by using the concentration result in [RV09, Theorem 1.1], on an event $E_{1,1;s}$ with $\mathbb{P}^\xi(E_{1,1;s}) \geq 1 - e^{-n/C}$, we have $(I) \leq Cb$.

For (II), using the boundedness of $\mathfrak{m}(z(b))$ around 0 for $\phi^{-1} \geq 1 + 1/K$, we may estimate

$$\begin{aligned}
(II) &\lesssim (\mathbf{1}_{\phi^{-1}\geq 1+1/K} \wedge \eta^{-1}) \cdot |\mathfrak{m}(z(b)) - \mathfrak{m}(z(0))| \\
&\leq (\mathbf{1}_{\phi^{-1}\geq 1+1/K} \wedge \eta^{-1}) \cdot \int_{C_0^{-1}\mathbf{1}_{\phi^{-1}\geq 1+1/K}}^{\infty} \frac{b}{|x - z(b)||x - z(0)|}\,\rho(\mathrm{d}x) \\
&\leq \Big\{C_0^2 \mathbf{1}^{-1}_{\phi^{-1}\geq 1+1/K} \wedge \eta^{-3}\Big\} \cdot b.
\end{aligned}$$

Combining the above estimates, for b chosen small enough, say, $b = n^{-100}$, on the event $E_{1,0;s}(n^{-100}) \cap E_{1,1;s}$,

$$|A_{1;s} - \langle \Sigma^{1/2} e_s, \mathfrak{m}(-\eta/\phi)\Sigma(I_n + \mathfrak{m}(-\eta/\phi)\Sigma)^{-1}\mu_0 - \mu_0\rangle| \lesssim n^{-1/2+\vartheta}.$$

Using $\tau_{\eta,*}^{-1} = \mathfrak{m}(-\eta/\phi)$ and the definition of $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})$, recall $w_{\eta,*} = \Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0)$ defined in (9.14), we then have

$$\sup_{\mu_0\in B_n(1)} \mathbb{P}^\xi\Big(\max_{s\in[n]}|A_{1;s} - \langle e_s, \mathbb{E}\, w_{\eta,*}\rangle| \geq Cn^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{10.6}$$

The term $A_{2;s}$ can be handled similarly, now reading off the $(1,2)$ element in [KY17, Eqn. (3.10)], which shows that for any $\xi \in \mathbb{R}^m$,

$$\mathbb{P}^\xi\Big(\max_{s\in[n]}|A_{2;s}| \geq C(\|\xi\|/\sqrt{m}) \cdot n^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{10.7}$$

Combining (10.4), (10.6) and (10.7), we have

$$\sup_{\mu_0\in B_n(1),\xi\in\mathcal{E}_\vartheta} \mathbb{P}^\xi\Big(\|\widehat{w}_\eta\|_\infty \geq \|\mathbb{E}\, w_{\eta,*}\|_\infty + Cn^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{10.8}$$

Now we will construct $\mathcal{U}_\vartheta \subset B_n(1)$ with the desired volume estimate, and $\sup_{\mu_0\in\mathcal{U}_\vartheta}\sup_{\eta\in\Xi_K}\|\mathbb{E}\, w_{\eta,*}\|_\infty \leq Cn^{-1/2+\vartheta}$. To this end, we place a uniform prior on $\mu_0 \sim U_0 g_0/\|g_0\|$, where $U_0 \sim \mathrm{Unif}[0,1]$ and $g_0 \sim \mathcal{N}(0, I_n)$ are independent of all other random variables. Then $\sup_{\eta\in\Xi_K}\|\mathbb{E}\, w_{\eta,*}\|_\infty \leq \sup_{\eta\in\Xi_K}\tau_{\eta,*}\|(\Sigma + \tau_{\eta,*}I_n)^{-1}\Sigma^{1/2}g_0\|_\infty/\|g_0\|$. Using Proposition 8.1-(3) and a standard Gaussian tail bound, $\mathbb{P}_{\mu_0}(\mathcal{U}_\vartheta \equiv \{\sup_{\eta\in\Xi_K}\|\mathbb{E}\, w_{\eta,*}\|_\infty \geq C_1 n^{-1/2+\vartheta}\}) \leq Ce^{-n^{2\vartheta}/C}$. Moreover, $\mathbb{P}(\xi \notin \mathcal{E}_\vartheta) \leq e^{-n^{2\vartheta}/C}$. The pointwise-in-$\eta$ delocalization claim on $\widehat{w}_\eta$ follows. As $\eta \mapsto \|\widehat{w}_\eta\|_\infty$ is C-Lipschitz with exponentially high probability, the uniform version follows by a standard discretization and union bound argument.

(2). Let us consider delocalization for $\widehat{v}_\eta$. Using again (10.3), for any $t \in [m]$,

$$\begin{aligned}
-\langle e_t, \widehat{v}_\eta\rangle &= n^{-1/2}\langle e_t, (\phi\check{\Sigma} + \eta I_m)^{-1}X\mu_0\rangle + n^{-1/2}\langle e_t, (\phi\check{\Sigma} + \eta I_m)^{-1}\xi\rangle \\
&\equiv B_{1;t} + B_{2;t}.
\end{aligned}$$

The term $B_{1;t}$ can be handled, by reading off the $(2,1)$ element in [KY17, Eqn. (3.10)], which shows that

$$\sup_{\mu_0\in B_n(1)} \mathbb{P}^\xi\Big(\max_{t\in[m]}|B_{1;t}| \geq Cn^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{10.9}$$

The term $B_{2;t}$ relies on the local law described by the $(2,2)$ element in [KY17, Eqn. (3.10)]: for any $\xi \in \mathbb{R}^m$,

$$\mathbb{P}^\xi\Big(\max_{t\in[m]}|B_{2;t} - \phi^{-1}\mathfrak{m}(-\eta/\phi)\xi_t| \geq C(\|\xi\|/\sqrt{m})\cdot n^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{10.10}$$

Consequently, combining (10.9)-(10.10), we have

$$\sup_{\mu_0\in B_n(1),\xi\in\mathcal{E}_\vartheta} \mathbb{P}^\xi\Big(\|\widehat{v}_\eta\|_\infty \geq Cn^{-1/2+\vartheta}\Big) \leq Cn^{-D}.$$

The claim follows. □

10.3. **Universality of the global cost optimum.**

Theorem 10.4. *Suppose Assumption A holds and the following hold for some $K > 0$.*

- $1/K \leq \phi^{-1} \leq K$, $\|\Sigma\|_{\mathrm{op}} \vee \|\Sigma^{-1}\|_{\mathrm{op}} \leq K$.
- *Assumption B holds with $\sigma_\xi^2 \in [1/K, K]$.*

Fix $\vartheta \in (0, 1/18)$. There exists some $C = C(K,\vartheta) > 0$ such that for $\rho_0 \leq 1/C$, $\eta \in \Xi_K$ and $\xi \in \mathcal{E}_\vartheta$,

$$\sup_{\mu_0\in\mathcal{U}_\vartheta} \mathbb{P}^\xi\Big(\Big|\min_{w\in\mathbb{R}^n} H_{\eta;Z}(w) - \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)\Big| \geq \rho_0\Big) \leq C\rho_0^{-3}\cdot n^{-1/6+3\vartheta}.$$

Here $\mathcal{U}_\vartheta$ is specified as in Proposition 10.3.

Proof. Fix $\vartheta > 0$, $\mu_0 \in \mathcal{U}_\vartheta$ and $\xi \in \mathcal{E}_\vartheta$ as specified in Proposition 10.3. Let $L_n \equiv C_0 n^\vartheta$. By the same proposition, with $\mathbb{P}^\xi$-probability at least $1 - C_0 n^{-100}$,

$$\begin{aligned}\min_{w\in\mathbb{R}^n} H_{\eta;Z}(w) &= \min_{\|w\|_\infty\leq L_n/\sqrt{n}}\ \max_{\|v\|_\infty\leq L_n/\sqrt{n}} \Big\{\frac{1}{\sqrt{n}}\langle v, Zw\rangle - \frac{1}{\sqrt{n}}\langle v,\xi\rangle - \frac{\eta}{2}\|v\|^2 + F(w)\Big\}\\ &= \min_{\|\widetilde{w}\|_\infty\leq L_n}\ \max_{\|\widetilde{v}\|_\infty\leq L_n}\Big\{\frac{1}{n^{3/2}}\langle\widetilde{v}, Z\widetilde{w}\rangle - \frac{1}{n}\langle\widetilde{v},\xi\rangle - \frac{\eta}{2n}\|\widetilde{v}\|^2 + F(\widetilde{w}/\sqrt{n})\Big\},\end{aligned} \tag{10.11}$$

and

$$\begin{aligned}&\min_{w\in\mathbb{R}^n} H_{\eta;G}(w)\\ &= \min_{\|\widetilde{w}\|_\infty\leq L_n}\ \max_{\|\widetilde{v}\|_\infty\leq L_n}\Big\{\frac{1}{n^{3/2}}\langle\widetilde{v}, G\widetilde{w}\rangle - \frac{1}{n}\langle\widetilde{v},\xi\rangle - \frac{\eta}{2n}\|\widetilde{v}\|^2 + F(\widetilde{w}/\sqrt{n})\Big\}.\end{aligned} \tag{10.12}$$

By writing $Q(\widetilde{v},\widetilde{w}) \equiv -\frac{1}{n}\langle\widetilde{v},\xi\rangle - \frac{\eta}{2n}\|\widetilde{v}\|^2 + F(\widetilde{w}/\sqrt{n})$, we have

$$\mathscr{N}_Q(L,\delta) \equiv \sup_{\substack{\|\widetilde{v}\|_\infty\vee\|\widetilde{v}'\|_\infty\leq L, \|\widetilde{v}-\widetilde{v}'\|_\infty\leq\delta,\\ \|\widetilde{w}\|_\infty\vee\|\widetilde{w}'\|_\infty\leq L, \|\widetilde{w}-\widetilde{w}'\|_\infty\leq\delta}} \big|Q(\widetilde{v},\widetilde{w}) - Q(\widetilde{v}',\widetilde{w}')\big| \lesssim_K (1\vee L)\delta\cdot\Big(1 + \frac{\|\xi\|_1}{n}\Big).$$

Now with $X_Q(\widetilde{v},\widetilde{w};Z)\equiv n^{-3/2}\langle \widetilde{v}, Z\widetilde{w}\rangle + Q(\widetilde{v},\widetilde{w})$, for $\xi\in\mathcal{E}_\vartheta$, by applying Theorem 10.2, we have for any $\mathsf{T}\in C^3(\mathbb{R})$,

$$\begin{aligned}
&\Big|\mathbb{E}^\xi\,\mathsf{T}\Big(\min_{\|\widetilde{w}\|_\infty\le L_n}\max_{\|\widetilde{v}\|_\infty\le L_n} X_Q(\widetilde{v},\widetilde{w};Z)\Big)-\mathbb{E}^\xi\,\mathsf{T}\Big(\min_{\|\widetilde{w}\|_\infty\le L_n}\max_{\|\widetilde{v}\|_\infty\le L_n} X_Q(\widetilde{v},\widetilde{w};G)\Big)\Big|\\
&\lesssim_K K_{\mathsf{T}}\cdot\inf_{\delta\in(0,1)}\Big\{\sqrt{n}L_n\delta+L_n\delta+\log_+^{2/3}(L_n/\delta)\cdot n^{-1/6}L_n^2\Big\}\\
&\le C_1\cdot K_{\mathsf{T}}\cdot n^{-1/6+3\vartheta}. \qquad (10.13)
\end{aligned}$$

Replicating the last paragraph of proof of [HS22, Theorem 2.3] (right above Section 4.3 therein), for any $z>0,\rho_0>0$,

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\min_{\|\widetilde{w}\|_\infty\le L_n}\max_{\|\widetilde{v}\|_\infty\le L_n} X_Q(\widetilde{v},\widetilde{w};Z)>z+3\rho_0\Big)\\
&\le \mathbb{P}^\xi\Big(\min_{\|\widetilde{w}\|_\infty\le L_n}\max_{\|\widetilde{v}\|_\infty\le L_n} X_Q(\widetilde{v},\widetilde{w};G)>z+\rho_0\Big)+C\rho_0^{-3}n^{-1/6+3\vartheta}.
\end{aligned}$$

Combined with (10.11)-(10.12), we have

$$\mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_{\eta;Z}(w)>z+3\rho_0\Big)\le \mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_{\eta;G}(w)>z+\rho_0\Big)+C_2\rho_0^{-3}n^{-1/6+3\vartheta}.$$

In view of (9.26) (in Step 1 of the final proof of Theorem 2.3), for $\rho_0\in(C_3n^{-1/2+\vartheta},1/C_3)$, we take $z\equiv z_\eta\equiv\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)$ and $t\equiv\rho_0^2n/C_3$ therein, so that for $\xi\in\mathcal{E}_\vartheta\subset\mathscr{E}_{1,\xi}(\rho_0/C_3^{1/2})$,

$$\mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_{\eta;G}(w)>z_\eta+\rho_0\Big)\le C_3e^{-\rho_0^2n/C_3}.$$

Combining the estimates, for $\xi\in\mathcal{E}_\vartheta$, $\rho_0\in(C_3n^{-1/2+\vartheta},1/C_3)$,

$$\mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_{\eta;Z}(w)>z_\eta+3\rho_0\Big)\le C_4\{e^{-\rho_0^2n/C_4}+\rho_0^{-3}n^{-1/6+3\vartheta}\}.$$

The first term above can be assimilated into the second one, and $\rho_0\ge C_3n^{-1/2+\vartheta}$ can be dropped. The lower bound follow similarly by utilizing (9.27). □

10.4. Universality of the cost over exceptional sets.

Theorem 10.5. *Suppose Assumption A holds and the following hold for some $K>0$.*

- $1/K\le\phi^{-1}\le K$, $\|\Sigma\|_{\mathrm{op}}\vee\|\Sigma^{-1}\|_{\mathrm{op}}\le K$.
- *Assumption B with variance $\sigma_\xi^2\in[1/K,K]$.*

Fix $\vartheta\in(0,1/18)$. Then there exists some $C=C(K,\vartheta)>0$ such that for $\mathsf{g}:\mathbb{R}^n\to\mathbb{R}$ being 1-Lipschitz with respect to $\|\cdot\|_{\Sigma^{-1}}$, $\rho_0\le 1/C$, $\eta\in\Xi_K$ and $\xi\in\mathcal{E}_\vartheta$,

$$\sup_{\mu_0\in\mathcal{U}_\vartheta}\mathbb{P}^\xi\Big(\min_{w\in D_{\eta;C\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w)\le\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)+\rho_0\Big)\le C\rho_0^{-6}\cdot n^{-1/6+3\vartheta}.$$

Here $B_{(2,\infty)}(C,L_n/\sqrt{n})\equiv B_n(C)\cap L_\infty(L_n/\sqrt{n})$ with $L_n\equiv Cn^\vartheta$, and $\mathcal{U}_\vartheta$ is specified as in Proposition 10.3.

Proof. Fix $\varepsilon, \vartheta > 0$, $\mu_0 \in \mathcal{U}_\vartheta$ and $\xi \in \mathcal{E}_\vartheta$ as specified in Proposition 10.3. We define a renormalized version of $D_{\varepsilon;\eta}(\mathsf{g})$ as

$$\widetilde{D}_{\varepsilon;\eta}(\mathsf{g}) \equiv \{\widetilde{w} \in \mathbb{R}^n : |\mathsf{g}(\widetilde{w}/\sqrt{n}) - \mathbb{E}\,\mathsf{g}(\widetilde{w}_{\eta,*}/\sqrt{n})| \geq \varepsilon\},$$

where $\widetilde{w}_{\eta,*} = \sqrt{n} w_{\eta,*}$.

(**Step 1**). Let $L_n \equiv C_0 n^\vartheta$. For any $z \in \mathbb{R}$ and $\rho_0 > 0$, with $Z_n \equiv Z/\sqrt{n}$,

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\min_{w \in D_{\eta;\varepsilon}(\mathsf{g}) \cap B_{(2,\infty)}(C_0, \frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) \leq z + \rho_0\Big) \qquad (10.14)\\
&= \mathbb{P}^\xi\Big(\min_{w \in D_{\eta;\varepsilon}(\mathsf{g}) \cap B_{(2,\infty)}(C_0, \frac{L_n}{\sqrt{n}})} \Big\{F(w) + \frac{1}{2n\eta}\|Zw - \xi\|^2\Big\} \leq z + \rho_0\Big)\\
&= \mathbb{P}^\xi\Big(\min_{\widetilde{w} \in \widetilde{D}_{\eta;\varepsilon}(\mathsf{g}) \cap B_{(2,\infty)}(\sqrt{n}C_0, L_n)} \Big\{\eta F(\widetilde{w}/\sqrt{n}) + \frac{1}{2n}\|Z_n\widetilde{w} - \xi\|^2\Big\} \leq \eta(z + \rho_0)\Big).
\end{aligned}$$

Now we may apply Theorem 10.1. To do so, let us write $\mathsf{f}(\widetilde{w}) \equiv \eta F(\widetilde{w}/\sqrt{n})$ to match the notation. Then a simple calculation leads to

$$\mathscr{N}_{\mathsf{f}}(L, \delta) \equiv \sup_{\|\widetilde{w}\|_\infty \vee \|\widetilde{w}'\|_\infty \leq L, \|\widetilde{w} - \widetilde{w}'\|_\infty \leq \delta} |\mathsf{f}(\widetilde{w}) - \mathsf{f}(\widetilde{w}')| \lesssim_K (1 \vee L)\delta,$$

Consequently, an application of Theorem 10.1 leads to

$$\begin{aligned}
&\text{RHS of (10.14)} - C_1(1 \vee (\eta\rho_0)^{-3}) L_n^2 n^{-1/6} \log^{2/3}(L_n n)\\
&\leq \mathbb{P}^\xi\Big(\min_{\widetilde{w} \in \widetilde{D}_{\eta;\varepsilon}(\mathsf{g}) \cap B_{(2,\infty)}(\sqrt{n}C_0, L_n)} \Big\{\eta F(\widetilde{w}/\sqrt{n}) + \frac{1}{2n}\|G_n\widetilde{w} - \xi\|^2\Big\} \leq \eta(z + 3\rho_0)\Big)\\
&\leq \mathbb{P}^\xi\Big(\min_{w \in D_{\eta;\varepsilon}(\mathsf{g}) \cap B_{(2,\infty)}(C_0, \frac{L_n}{\sqrt{n}})} H_{\eta;G}(w) \leq z + 3\rho_0\Big)\\
&\leq \mathbb{P}^\xi\Big(\min_{w \in D_{\eta;\varepsilon}(\mathsf{g}) \cap B_n(C_0)} H_{\eta;G}(w) \leq z + 3\rho_0\Big).
\end{aligned}$$

Here in the last inequality we simply drop the L_∞ constraint. Now for $C_2 n^{-1/2+\vartheta} \leq \rho_0 \leq 1/C_2$, by choosing $z \equiv z_\eta \equiv \max_{\beta > 0} \min_{\gamma > 0} \overline{\mathsf{D}}_\eta(\beta, \gamma)$ and $t \equiv 2\rho_0^2 n/C_3$ in Theorem 9.6, where C_3 is the constant therein, we have

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\min_{w \in D_{\eta;C_4\rho_0^{1/2}}(\mathsf{g}) \cap B_{(2,\infty)}(C, \frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) \leq \max_{\beta > 0} \min_{\gamma > 0} \overline{\mathsf{D}}_\eta(\beta, \gamma) + \rho_0\Big)\\
&\leq C\Big\{e^{-\rho_0^2 n/C_3} + (\eta\rho_0)^{-3} \cdot n^{-1/6 + 3\vartheta}\Big\} \leq C_4 \cdot (\eta\rho_0)^{-3} \cdot n^{-1/6 + 3\vartheta}. \qquad (10.15)
\end{aligned}$$

The constraints $\rho_0 \geq C_2 n^{-1/2+\vartheta}$ can be removed by enlarging C_4 if necessary.

(**Step 2**). In this step we shall trade the dependence of the above bound with respect to $\eta > 0$ with a possible worsened dependence on ρ_0, primarily in the regime $\phi^{-1} \geq 1 + 1/K$. Fix $\eta_0 \in \Xi_K$. Let $\eta > 0$ be chosen later and $\eta_1 \equiv \eta_0 + \eta$. Without loss of generality we assume $\eta_0, \eta_1 \in \Xi_K$, so by (9.13) in Proposition 9.5, $|z_{\eta_1} - z_{\eta_0}| \leq C_5\eta$. By enlarging C_5 if necessary we assume that C_5 exceeds the constant in Lemma 10.6. Using Lemma 10.6, for $\varepsilon = 2C_4\rho_0^{1/2}$, with the choice

$\eta = C_4\rho_0/C_5 \le C_4\rho_0^{1/2}/C_5$ (we assume without loss of generality $\rho_0 \le 1$),

$$
\begin{aligned}
&\mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta_0;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta_0;Z}(w) \le z_{\eta_0} + \rho_0\Big)\\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta_0;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta_1;Z}(w) \le z_{\eta_0} + \rho_0\Big) \quad (\text{since } H_{\eta_1;Z} \le H_{\eta_0;Z})\\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta_1;(\varepsilon-C_5\eta)_+}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta_1;Z}(w) \le z_{\eta_0} + \rho_0\Big) \quad (\text{by Lemma 10.6})\\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta_1;(\varepsilon-C_5\eta)_+}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta_1;Z}(w) \le z_{\eta_1} + C_5\eta + \rho_0\Big)\\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta;C_4\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta_1;Z}(w) \le \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_{\eta_1}(\beta,\gamma) + C\rho_0\Big)\\
&\le C\cdot(\eta_0+\rho_0)^{-3}\rho_0^{-3}\cdot n^{-1/6+3\vartheta} \le C\cdot\rho_0^{-6}n^{-1/6+3\vartheta}.
\end{aligned}
$$

The proof is complete by adjusting constants. □

Lemma 10.6. *Suppose $\|\mu_0\| \vee \|\Sigma\|_{\mathrm{op}} \vee \|\Sigma^{-1}\|_{\mathrm{op}} \le K$. Let $\mathsf{g}: \mathbb{R}^n \to \mathbb{R}$ be 1-Lipschitz with respect to $\|\cdot\|_{\Sigma^{-1}}$. Then there exists some constant $C = C(K) > 0$ such that for any $\varepsilon > 0, \eta_0, \eta_1 \in \Xi_K$ with $\eta_1 \ge \eta_0$, we have $D_{\eta_0;\varepsilon}(\mathsf{g}) \subset D_{\eta_1;(\varepsilon - C(\eta_1-\eta_0))_+}(\mathsf{g})$.*

Proof. Using the definition of $w_{\eta,*}$ in (9.14), we have

$$
\begin{aligned}
&|\mathbb{E}\,\mathsf{g}(w_{\eta_1,*}) - \mathbb{E}\,\mathsf{g}(w_{\eta_0,*})| \le \mathbb{E}\|w_{\eta_1,*} - w_{\eta_0,*}\|_{\Sigma^{-1}}\\
&= \mathbb{E}\Big\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta_1,*};\tau_{\eta_1,*}) - \widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta_0,*};\tau_{\eta_0,*})\Big\| \le C\cdot(\eta_1-\eta_0).
\end{aligned}
$$

Here the last inequality follows from the calculations in (9.31). So for any $w \in D_{\eta_0;\varepsilon}(\mathsf{g})$, we have

$$\varepsilon \le |\mathsf{g}(w) - \mathbb{E}\,\mathsf{g}(w_{\eta_0,*})| \le |\mathsf{g}(w) - \mathbb{E}\,\mathsf{g}(w_{\eta_1,*})| + C(\eta_1-\eta_0).$$

This means $w \in D_{\eta_1;(\varepsilon - C(\eta_1-\eta_0))_+}(\mathsf{g})$, as desired. □

10.5. **Proof of the universality Theorem 2.4 for $\widehat{\mu}_{\eta;Z}$.** Fix $\vartheta > 0$, $\mu_0 \in \mathcal{U}_\vartheta$ and $\xi \in \mathcal{E}_\vartheta$. Let $L_n \equiv C_0 n^{\vartheta}$, and $E_0 \equiv \{\widehat{w}_{n;Z} \in B_{(2,\infty)}(C_0, L_n/\sqrt{n}) = B_n(C_0) \cap L_\infty(L_n/\sqrt{n})\}$. We assume that C_0 exceeds the constants in Proposition 10.3 and Theorem 10.5. By Proposition 10.3 and a simple ℓ_2 estimate, $\mathbb{P}^{\xi}(E_0^c) \le C_0 n^{-100}$. We further let $z_\eta \equiv \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)$ for $\eta \ge 0$.

Let $\mathsf{g}: \mathbb{R}^n \to \mathbb{R}$ be 1-Lipschitz with respect to $\|\cdot\|_{\Sigma^{-1}}$. Then for $\rho_0 \le 1/C_0$ and $\eta \in \Xi_K$, we have

$$
\begin{aligned}
&\mathbb{P}^{\xi}\big(\widehat{w}_{\eta;Z} \in D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g})\big)\\
&\le \mathbb{P}^{\xi}\big(\widehat{w}_{\eta;Z} \in D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g}) \cap B_{(2,\infty)}(C_0, L_n/\sqrt{n})\big) + \mathbb{P}^{\xi}(E_0^c)\\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) \ge z_\eta + \rho_0\Big)\\
&\quad + \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) \le z_\eta + 2\rho_0\Big) + C_0 n^{-100}.
\end{aligned}
$$

Here in the last inequality we used the simple fact that

$$\Big\{\min_{w\in B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) < z_\eta + \rho_0\Big\} \cap \Big\{\min_{w\in D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) > z_\eta + 2\rho_0\Big\}$$
$$\subset \{\widehat{w}_{\eta;Z} \notin D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g}) \cap B_{(2,\infty)}(C_0, L_n/\sqrt{n})\}.$$

Invoking Theorems 10.4 and 10.5, by enlarging C_0 if necessary, we have for $\rho_0 \le 1/C_0$ and $\eta \in \Xi_K$,

$$\mathbb{P}^\xi\Big(|\mathsf{g}(\widehat{w}_{\eta;Z}) - \mathbb{E}\,\mathsf{g}(w_{\eta,*})| \ge \rho_0^{1/2}\Big) \le C_0 \cdot \rho_0^{-6} n^{-1/6+3\vartheta},$$

or equivalently, for $\mathsf{g}_0 : \mathbb{R}^n \to \mathbb{R}$ being 1-Lipschitz with respect to $\|\cdot\|$,

$$\mathbb{P}^\xi\Big(\big|\mathsf{g}_0(\widehat{\mu}_{\eta;Z}) - \mathbb{E}\,\mathsf{g}_0(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}))\big| \ge \rho_0\Big) \le C_0 \cdot \rho_0^{-12} n^{-1/6+3\vartheta}.$$

Now we may follow Step 4 in the proof of Theorem 2.3 to strengthen the above statement to a uniform one in η; we only sketch the differences below. Using (9.29) with G therein replaced by Z, and the assumption $\|\Sigma^{-1}\|_{\mathrm{op}} \le K$, we arrive at a modified form of (9.30): on an event E_1 with $\mathbb{P}^\xi(E_1) \ge 1 - C_1 e^{-n/C_1}$, for any $\eta_1, \eta_2 \in \Xi_K$,

$$\|\widehat{\mu}_{\eta_1;Z} - \widehat{\mu}_{\eta_2;Z}\| \le C_1|\eta_1 - \eta_2|. \tag{10.16}$$

Using (9.31) with $\|\Sigma^{-1}\|_{\mathrm{op}} \le K$, we arrive at a modified form of (9.32): for any $\eta_1, \eta_2 \in \Xi_K$,

$$\Big|\mathbb{E}\,\mathsf{g}_0(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta_1,*};\tau_{\eta_1,*})) - \mathbb{E}\,\mathsf{g}_0(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta_2,*};\tau_{\eta_2,*}))\Big| \le C_1|\eta_1 - \eta_2|.$$

Now using a standard discretization and a union bound, we have

$$\mathbb{P}^\xi\Big(\sup_{\eta\in\Xi_K}\big|\mathsf{g}_0(\widehat{\mu}_{\eta;Z}) - \mathbb{E}\,\mathsf{g}_0(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}))\big| \ge \rho_0\Big) \le C_2 \cdot \rho_0^{-13} n^{-1/6+3\vartheta}.$$

The proof is complete by taking expectation with respect to ξ and note that $\mathbb{P}(\xi \in \mathcal{E}_\vartheta) \ge 1 - Ce^{-n^{2\vartheta}/C}$ as in Proposition 10.3. □

10.6. Proof of the universality Theorem 2.4 for $\widehat{r}_{\eta;Z}$.

Proposition 10.7. *Suppose Assumption A holds and the following hold for some* $K > 0$.

- $1/K \le \phi^{-1}, \eta \le K$, $\|\Sigma\|_{\mathrm{op}} \vee \|\Sigma^{-1}\|_{\mathrm{op}} \le K$.
- *Assumption B holds with* $\sigma_\xi^2 \in [1/K, K]$.

Fix $\vartheta \in (0, 1/18)$. *Then there exists some* $C = C(K, \vartheta) > 0$ *such that for any* 1*-Lipschitz function* $\mathsf{h} : \mathbb{R}^m \to \mathbb{R}$, $\rho_0 \le 1/C$, $\eta \in \Xi_K$ *and* $\xi \in \mathcal{E}_\vartheta$,

$$\sup_{\mu_0\in\mathcal{U}_\vartheta} \mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C\rho_0^{1/2}}(\mathsf{h})\cap L_\infty(\frac{L_n}{\sqrt{n}})} \min_{w\in\mathbb{R}^n} h_{\eta;Z}(w,v) \ge \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) - \rho_0\Big) \le C\rho_0^{-3} n^{-1/6+3\vartheta}.$$

Here $L_n \equiv Cn^\vartheta$, *and* $\mathcal{U}_\vartheta$ *is specified as in Proposition 10.3.*

Proof. Fix $\varepsilon, \vartheta > 0$, $\mu_0 \in \mathcal{U}_\vartheta$ and $\xi \in \mathcal{E}_\vartheta$ as specified in Proposition 10.3. We define a renormalized version of $D_{\varepsilon;\eta}(\mathsf{h})$ as

$$\widetilde{D}_{\varepsilon;\eta}(\mathsf{h}) \equiv \{\widetilde{r} \in \mathbb{R}^m : |\mathsf{h}(\widetilde{r}/\sqrt{n}) - \mathbb{E}^\xi \mathsf{h}(\widetilde{r}_{\eta,*}/\sqrt{n})| \geq \varepsilon\},$$

where $\widetilde{r}_{\eta,*} = \sqrt{n} r_{\eta,*}$. Let $L_n = C_0 n^\vartheta$ and $Q(\widetilde{v}, \widetilde{w})$ be defined as in the proof of Theorem 10.4. Then we have,

$$\begin{aligned}
&\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} h_{\eta;Z}(w, v) \\
&= \max_{\widetilde{v} \in \widetilde{D}_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n)} \min_{\widetilde{w} \in L_\infty(L_n)} h_{\eta;Z}(\widetilde{w}/\sqrt{n}, \widetilde{v}/\sqrt{n}) \\
&= \max_{\widetilde{v} \in \widetilde{D}_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n)} \min_{\widetilde{w} \in L_\infty(L_n)} \left\{ \frac{1}{n^{3/2}} \langle \widetilde{v}, Z\widetilde{w} \rangle - \frac{1}{n} \langle \widetilde{v}, \xi \rangle + F(\widetilde{w}/\sqrt{n}) - \frac{\eta \|\widetilde{v}\|^2}{2n} \right\} \\
&= \max_{\widetilde{v} \in \widetilde{D}_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n)} \min_{\widetilde{w} \in L_\infty(L_n)} \left\{ \frac{1}{n^{3/2}} \langle \widetilde{v}, Z\widetilde{w} \rangle + Q(\widetilde{v}, \widetilde{w}) \right\}.
\end{aligned}$$

Using the comparison inequality in Theorem 10.2 and a similar calculation as in (10.13), with $s_n(\rho_0) \equiv \rho_0^{-3} n^{-1/6+3\vartheta}$,

$$\begin{aligned}
&\mathbb{P}^\xi \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} h_{\eta;Z}(w, v) \geq z - \rho_0 \Big) \\
&= \mathbb{P}^\xi \Big(\max_{\widetilde{v} \in \widetilde{D}_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n)} \min_{\widetilde{w} \in L_\infty(L_n)} \left\{ \frac{1}{n^{3/2}} \langle \widetilde{v}, Z\widetilde{w} \rangle + Q(\widetilde{v}, \widetilde{w}) \right\} \geq z - \rho_0 \Big) \\
&\leq \mathbb{P}^\xi \Big(\max_{\widetilde{v} \in \widetilde{D}_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n)} \min_{\widetilde{w} \in L_\infty(L_n)} \left\{ \frac{1}{n^{3/2}} \langle \widetilde{v}, G\widetilde{w} \rangle + Q(\widetilde{v}, \widetilde{w}) \right\} \geq z - 3\rho_0 \Big) + C s_n(\rho_0) \\
&= \mathbb{P}^\xi \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} h_{\eta;G}(w, v) \geq z - 3\rho_0 \Big) + C_1 s_n(\rho_0).
\end{aligned}$$

Using the convex Gaussian min-max theorem (cf. Theorem 6.1),

$$\begin{aligned}
&\mathbb{P}^\xi \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} h_{\eta;Z}(w, v) \geq z - \rho_0 \Big) \\
&\leq 2\,\mathbb{P} \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} \ell_\eta(w, v) \geq z - 3\rho_0 \Big) + C_1 s_n(\rho_0). \qquad (10.17)
\end{aligned}$$

On the other hand, using the definition of $w_{\eta,*}$ in (9.14), and the fact that for any $\mu_0 \in \mathcal{U}_\vartheta$, $\|\mathbb{E}\, w_{\eta,*}\|_\infty \leq L_n/\sqrt{n}$, we have $\mathbb{P}\left(\|w_{\eta,*}\|_\infty \geq L_n/\sqrt{n}\right) \leq Ce^{-n^{2\vartheta}/C}$. Combined with (10.17), we have

$$\begin{aligned}
&\mathbb{P}^\xi \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} h_{\eta;Z}(w, v) \geq z - \rho_0 \Big) \\
&\leq 2\,\mathbb{P} \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \ell_\eta(w_{\eta,*}, v) \geq z - 3\rho_0 \Big) + C_2 s_n(\rho_0).
\end{aligned}$$

In view of (9.45), now by choosing $z \equiv z_\eta \equiv \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)$ and $\varepsilon \equiv C_3\rho_0^{1/2}$, for $\rho_0 \ge C_4 n^{-1/2+\vartheta}$, $\xi \in \mathcal{E}_\vartheta \subset \mathscr{E}_{1,\xi}(\rho_0/C)$, it follows that

$$
\begin{aligned}
&\mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C_3\rho_0^{1/2}}(\mathsf{h})\cap L_\infty(L_n/\sqrt{n})}\min_{w\in\mathbb{R}^n} h_{\eta;Z}(w,v) \ge z_\eta - \rho_0\Big)\\
&\le \mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C_3\rho_0^{1/2}}(\mathsf{h})\cap L_\infty(L_n/\sqrt{n})}\min_{w\in L_\infty(L_n/\sqrt{n})} h_{\eta;Z}(w,v) \ge z_\eta - \rho_0\Big) \le C_4 s_n(\rho_0).
\end{aligned}
$$

The claim follows by adjusting constants. □

Proof of Theorem 2.4 for $\widehat{r}_{\eta;Z}$. Fix $\vartheta > 0$, $\mu_0 \in \mathcal{U}_\vartheta$ and $\xi \in \mathcal{E}_\vartheta$ as specified in Proposition 10.3. We continue writing $z_\eta \equiv \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)$ in the proof. Using the delocalization results in Proposition 10.3, on an event E_0 with $\mathbb{P}^\xi(E_0) \ge 1 - C_0 n^{-100}$, we have $\|\widehat{w}_{\eta;Z}\|_\infty \vee \|\widehat{r}_{\eta;Z}\|_\infty \le L_n/\sqrt{n}$ with $L_n = C_0 n^\vartheta$. Using Theorem 10.4, for $\rho_0 \le 1/C$, and $\eta \in \Xi_K$, by possibly adjusting $C_0 > 0$,

$$
\begin{aligned}
&\mathbb{P}^\xi\Big(\max_{v\in L_\infty(L_n/\sqrt{n})}\min_{w\in\mathbb{R}^m} h_{\eta;Z}(w,v) \le z_\eta - \rho_0/2\Big)\\
&\le \mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^m} H_{\eta;Z}(w) \le z_\eta - \rho_0/2\Big) + \mathbb{P}^\xi(E_0^c) \le C_0\rho_0^{-3}\cdot n^{-1/6+3\vartheta}. \qquad (10.18)
\end{aligned}
$$

Let us take $C_1 > 0$ to be the constant in Proposition 10.7. By noting that

$$
\begin{aligned}
&\Big\{\max_{v\in L_\infty(L_n/\sqrt{n})}\min_{w\in\mathbb{R}^m} h_{\eta;Z}(w,v) > z_\eta - \rho_0/2\Big\}\\
&\quad\cap\Big\{\max_{v\in D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h})\cap L_\infty(L_n/\sqrt{n})}\min_{w\in\mathbb{R}^n} h_{\eta;Z}(w,v) < z_\eta - \rho_0\Big\}\\
&\subset \{\widehat{v}_{\eta;Z} \notin D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})\},
\end{aligned}
$$

it follows from (10.18) and Proposition 10.7 that

$$
\begin{aligned}
&\mathbb{P}^\xi\Big(\widehat{v}_{\eta;Z} \in D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h})\Big)\\
&\le \mathbb{P}^\xi\Big(\widehat{v}_{\eta;Z} \in D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})\Big) + \mathbb{P}^\xi\Big(\widehat{v}_{\eta;Z} \notin L_\infty(L_n/\sqrt{n})\Big)\\
&\le \mathbb{P}^\xi\Big(\max_{v\in L_\infty(L_n/\sqrt{n})}\min_{w\in\mathbb{R}^m} h_{\eta;Z}(w,v) \le z_\eta - \rho_0/2\Big)\\
&\qquad + \mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h})\cap L_\infty(L_n/\sqrt{n})}\min_{w\in\mathbb{R}^n} h_{\eta;Z}(w,v) \ge z_\eta - \rho_0\Big) + \mathbb{P}^\xi(E_0^c)\\
&\le C\rho_0^{-3}\cdot n^{-1/6+3\vartheta}.
\end{aligned}
$$

Finally we only need to extend the above display to a uniform control over $\eta \in [1/K, K]$ by continuity arguments similar to Step 5 of the proof of Theorem 2.3 for $\widehat{r}_{\eta;G}$. By (9.50) (where G therein is replaced by Z) and (10.16), on an event E_1 with $\mathbb{P}^\xi(E_1) \ge 1 - Ce^{-n/C}$, for any $\eta_1, \eta_2 \in [1/K, K]$,

$$
\|\widehat{r}_{\eta_1;Z} - \widehat{r}_{\eta_2;Z}\| \le C|\eta_1 - \eta_2|.
$$

On the other hand, (9.52) remains valid, so we may proceed with an ε-net argument over $[1/K, K]$ to conclude. □

10.7. **Proof of Theorem 2.5.** To keep notation simple, we work with $\mathsf{A} = I_n$ and write $\Gamma^{I_n}_{\eta;(\Sigma,\|\mu_0\|)} = \Gamma_{\eta;(\Sigma,\|\mu_0\|)}$. The general case follows from minor modifications.

Lemma 10.8. *Suppose the conditions in Theorem 2.5 hold for some $K > 0$. Fix $q \in (0,\infty)$. There exists some constant $c = c(K,q) > 0$ such that $n^{\frac{1}{2}-\frac{1}{q}}\,\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0\|_q \geq c$ uniformly in $\eta \in \Xi_K$.*

Proof. We may write $\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0\|_q = \mathbb{E}\big(\sum_{j=1}^n |a_j + b_j g_j|^q\big)^{1/q}$ for some $a_j, b_j \in \mathbb{R}$ with $b_j \asymp 1$, and $g_j \sim \mathcal{N}(0,1/n)$ not necessarily independent of each other. So for some $c_j \in \mathbb{R}$,

$$\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0\|_q \gtrsim \mathbb{E}\Big(\sum_{j=1}^n |c_j + g_j|^q\Big)^{1/q}.$$

If $\sum_{j=1}^n |c_j|^q \geq C_0 \sum_{j=1}^n \mathbb{E}|g_j|^q$ for a large enough $C_0 > 0$, the lower bound follows trivially. Otherwise, with $Z \equiv \sum_{j=1}^n |c_j + g_j|^q$, we have $\mathbb{E}\,Z \geq \sum_{j=1}^n \inf_{c\in\mathbb{R}} \mathbb{E}|c+g_j|^q \gtrsim n^{1-q/2}$ and $\mathbb{E}\,Z^2 \lesssim \mathbb{E}\big(\sum_{j=1}^n (|g_j|^q + \mathbb{E}|g_j|^q)\big)^2 \lesssim (n^{1-q/2})^2$, so by Paley-Zygmund inequality, $\mathbb{P}(Z \geq \mathbb{E}\,Z/2) \geq (\mathbb{E}\,Z)^2/(4\,\mathbb{E}\,Z^2) \geq c_0$ for some $c_0 > 0$. In other words, on an event E_0 with $\mathbb{P}(E_0) \geq c_0$, $Z \geq c_0 n^{1-q/2}$. Using the above display, this means that $\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0\|_q \gtrsim \mathbb{E}\,Z^{1/q} \geq \mathbb{E}\,Z^{1/q}\mathbf{1}_{E_0} \gtrsim n^{1/q-1/2}$. □

Lemma 10.9. *Suppose the conditions in Theorem 2.5 hold for some $K > 0$. Fix $q \in (0,\infty)$. Then there exist constants $C = C(K,q) > 1$, $\vartheta = \vartheta(q) \in (0,1/50)$, and a measurable set $\mathcal{U}_\vartheta \subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \geq 1 - Ce^{-n^{\vartheta}/C}$, such that*

$$\sup_{\mu_0\in\mathcal{U}_\vartheta} n^{\frac{1}{2}-\frac{1}{q}}\Big|\,\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0\|_q - n^{-\frac{1}{2}}\|\mathrm{diag}(\Gamma_{\eta;(\Sigma,\|\mu_0\|)})\|_{q/2}^{\frac{1}{2}} M_q\Big| \leq Cn^{-\vartheta}.$$

Here $M_q = \mathbb{E}^{1/q}|\mathcal{N}(0,1)|^q$.

Proof. We write $\tau_{\eta,*} = \tau_\eta$, $\gamma_{\eta,*} = \gamma_\eta$ for notational simplicity in the proof. All the constants in $\lesssim, \gtrsim, \asymp$ below may depend on K, q. Recall the general fact $\|x\|_q \leq n^{-\frac{1}{2}+\frac{1}{q\wedge 2}}\|x\|^{\frac{2}{q\vee 2}}\|x\|_\infty^{1-\frac{2}{q\vee 2}}$ for $x \in \mathbb{R}^n$ and $q \in (0,\infty)$.

By Proposition 11.2 below, for any $\vartheta \in (0,1/2)$, there exists some $\mathcal{U}_\vartheta \subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \geq 1 - Ce^{-n^{1-2\vartheta}/C}$, such that $\sup_{\mu_0\in\mathcal{U}_\vartheta}\sup_{\eta\in\Xi_K}|\gamma_\eta^2 - \widetilde{\gamma}_\eta^2(\|\mu_0\|)| \leq n^{-\vartheta}$. Consequently, uniformly in $\mu_0 \in \mathcal{U}_\vartheta$ and $\eta \in \Xi_K$,

$$\begin{aligned}
&\Big|\,\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_\eta;\tau_\eta) - \mu_0\|_q - \mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\widetilde{\gamma}_\eta(\|\mu_0\|);\tau_\eta) - \mu_0\|_q\Big| \\
&\lesssim |\gamma_\eta - \widetilde{\gamma}_\eta(\|\mu_0\|)| \cdot n^{-1/2}\,\mathbb{E}\|(\Sigma+\tau_\eta I)^{-1}\Sigma^{1/2}g\|_q \\
&\leq n^{-\frac{1}{2}-\vartheta}\cdot n^{-\frac{1}{2}+\frac{1}{q\wedge 2}}\cdot\mathbb{E}\Big\{\|(\Sigma+\tau_\eta I)^{-1}\Sigma^{1/2}g\|^{\frac{2}{q\vee 2}}\cdot\|(\Sigma+\tau_\eta I)^{-1}\Sigma^{1/2}g\|_\infty^{1-\frac{2}{q\vee 2}}\Big\} \\
&\lesssim n^{-\frac{1}{2}+\frac{1}{q}-\vartheta}\big(\log n\big)^{\frac{1}{2}-\frac{1}{q\vee 2}}. \qquad (10.19)
\end{aligned}$$

For $g' \in \mathbb{R}^n$, let $\mathsf{f}(g') \equiv n^{-1/2}\|(\Sigma+\tau_\eta I)^{-1}g'\|_q$, and

$$\mathsf{F}_{\|\mu_0\|}(g') \equiv n^{-1/2}\Big\|(\Sigma+\tau_\eta I)^{-1}\big(-\tau_\eta\|\mu_0\|g' + \widetilde{\gamma}_\eta(\|\mu_0\|)\Sigma^{1/2}g\big)\Big\|_q,$$

$$\mathsf{F}_{\|\mu_0\|,0}(g') \equiv \Big\|(\Sigma+\tau_\eta I)^{-1}\Big(-\tau_\eta\|\mu_0\|\frac{g'}{\|g'\|}+\widetilde{\gamma}_\eta(\|\mu_0\|)\Sigma^{1/2}\frac{g}{\sqrt{n}}\Big)\Big\|_q.$$

Then for $g_1', g_2' \in \mathbb{R}^n$,

$$\big|\mathsf{F}_{\|\mu_0\|}(g_1')-\mathsf{F}_{\|\mu_0\|}(g_2')\big| \vee \big|\mathsf{f}(g_1')-\mathsf{f}(g_2')\big| \lesssim n^{-1+\frac{1}{q\wedge 2}}\|g_1'-g_2'\|.$$

By Gaussian concentration inequality, for any $\vartheta \in (0,1/2)$, we may find some $\mathcal{G}_{\vartheta,\|\mu_0\|} \subset \mathbb{R}^n$ with $\mathbb{P}(g_0 \in \mathcal{G}_{\vartheta,\|\mu_0\|}) \geq 1-Ce^{-n^{2\vartheta}/C}$, $g_0 \sim \mathcal{N}(0,I_n)$, such that uniformly in $g' \in \mathcal{G}_{\vartheta,\|\mu_0\|}$,

$$\begin{aligned}\max\Big\{\big|\|g'\|-\sqrt{n}\big|, n^{1-\frac{1}{q\wedge 2}}\big|\mathsf{F}_{\|\mu_0\|}(g')-\mathbb{E}_{g_0}\mathsf{F}_{\|\mu_0\|}(g_0)\big|,\\ n^{1-\frac{1}{q\wedge 2}}\big|\mathsf{f}(g')-\mathbb{E}\,\mathsf{f}(g_0)\big|\Big\} \leq n^{\vartheta}.\end{aligned} \tag{10.20}$$

As $\mathbb{E}\,\mathsf{f}(g_0) = n^{-1/2}\,\mathbb{E}\|(\Sigma+\tau_\eta I)g\|_q \lesssim n^{-\frac{1}{2}+\frac{1}{q}}(\log n)^{\frac{1}{2}-\frac{1}{q\vee 2}}$, for ϑ small enough, uniformly in $g' \in \mathcal{G}_{\vartheta,\|\mu_0\|}$,

$$\begin{aligned}&\big|\mathsf{F}_{\|\mu_0\|}(g')-\mathsf{F}_{\|\mu_0\|,0}(g')\big| \lesssim |\mathsf{f}(g')|\cdot\big|1-\sqrt{n}/\|g'\|\big|\\ &\lesssim \big(\mathbb{E}\,\mathsf{f}(g_0)+n^{-1+\frac{1}{q\wedge 2}+\vartheta}\big)\cdot n^{-\frac{1}{2}+\vartheta} \lesssim n^{-1+\frac{1}{q}+\vartheta}(\log n)^{\frac{1}{2}-\frac{1}{q\vee 2}}.\end{aligned} \tag{10.21}$$

Combining (10.20)-(10.21), for ϑ small enough,

$$\sup_{g' \in \mathcal{G}_{\vartheta,\|\mu_0\|}} n^{1-\frac{1}{q\wedge 2}}\big|\mathsf{F}_{\|\mu_0\|,0}(g')-\mathbb{E}_{g_0}\mathsf{F}_{\|\mu_0\|}(g_0)\big| \lesssim n^{\vartheta}. \tag{10.22}$$

Now let $\partial\mathcal{G}_{\vartheta,\|\mu_0\|} \equiv \{g'/\|g'\| : g' \in \mathcal{G}_{\vartheta,\|\mu_0\|}\} \subset \partial B_n(1)$. Using that $\{g_0 \in \mathcal{G}_{\vartheta,\|\mu_0\|}\} \subset \{g_0/\|g_0\| \in \partial\mathcal{G}_{\vartheta,\|\mu_0\|}\}$, we have $\mathbb{P}\big(g_0/\|g_0\| \in \partial\mathcal{G}_{\vartheta,\|\mu_0\|}\big) \geq \mathbb{P}(g_0 \in \mathcal{G}_{\vartheta,\|\mu_0\|}) \geq 1-Ce^{-n^{2\vartheta}/C}$. So with

$$\mathcal{V}_\vartheta \equiv \{\mu_0 = U_0 g' : U_0 \in [0,1], g' \in \partial\mathcal{G}_{\vartheta,U_0}\} \subset B_n(1),$$

we have $\mathbb{P}\big(\mathrm{Unif}(B_n(1)) \in \mathcal{V}_\vartheta\big) = \mathbb{E}_{U_0}\mathbb{P}_{g_0}\big(g_0/\|g_0\| \in \partial\mathcal{G}_{\vartheta,U_0}\big) \geq 1-Ce^{-n^{2\vartheta}/C}$. In other words, for this constructed set $\mathcal{V}_\vartheta$, we have the desired volume estimate $\mathrm{vol}(\mathcal{V}_\vartheta)/\mathrm{vol}(B_n(1)) \geq 1-Ce^{-n^{2\vartheta}/C}$, and by (10.22),

$$n^{1-\frac{1}{q\wedge 2}}\sup_{\mu_0\in\mathcal{V}_\vartheta}\Big|\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\widetilde{\gamma}_\eta(\|\mu_0\|);\tau_\eta)-\mu_0\|_q-\mathbb{E}_{g_0}\mathsf{F}_{\|\mu_0\|}(g_0)\Big| \lesssim n^{\vartheta}. \tag{10.23}$$

On the other hand, using the definition of $\Gamma_{\eta;(\Sigma,\|\mu_0\|)}$ in (2.7), we may compute

$$\mathbb{E}_{g_0}\mathsf{F}_{\|\mu_0\|}(g_0) = \mathbb{E}\,\big\|\Gamma^{1/2}_{\eta;(\Sigma,\|\mu_0\|)}g/\sqrt{n}\big\|_q. \tag{10.24}$$

Combining (10.19), (10.23) and (10.24), for ϑ chosen small enough,

$$\begin{aligned}&\sup_{\mu_0\in\mathcal{U}_\vartheta\cap\mathcal{V}_\vartheta} n^{1/2-1/q}\Big|\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_\eta;\tau_\eta)-\mu_0\|_q-\mathbb{E}\,\big\|\Gamma^{1/2}_{\eta;(\Sigma,\|\mu_0\|)}g/\sqrt{n}\big\|_q\Big|\\ &\lesssim n^{\frac{1}{2}-\frac{1}{q}}\cdot\Big(n^{-\frac{1}{2}+\frac{1}{q}-\vartheta}(\log n)^{\frac{1}{2}-\frac{1}{q\vee 2}}+n^{-1+\frac{1}{q\wedge 2}+\vartheta}\Big)\\ &= n^{-\vartheta}(\log n)^{\frac{1}{2}-\frac{1}{q\vee 2}}+n^{-\frac{1}{2}-\frac{1}{q}+\frac{1}{q\wedge 2}+\vartheta} \lesssim n^{-\vartheta/2}.\end{aligned}$$

The claim follows from Lemma B.2. □

Proof of Theorem 2.5. We write $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) = \widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}$ in the proof.

First we consider $0 < q \le 2$. This is the easy case, as $\mathsf{g}_q(x) \equiv \|x-\mu_0\|_q/n^{1/q-1/2}$ is 1-Lipschitz with respect to $\|\cdot\|$. So applying Theorems 2.3 and 2.4 verifies the existence of some small $\vartheta > 0$ such that for some $\mathcal{U}_\vartheta \subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \ge 1 - Ce^{-n^{\vartheta}/C}$,

$$\sup_{\mu_0\in\mathcal{U}_\vartheta} \mathbb{P}\Big(\sup_{\eta\in\Xi_K} n^{\frac{1}{2}-\frac{1}{q}}\Big|\|\widehat{\mu}_\eta-\mu_0\|_q - \mathbb{E}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_q\Big| \ge n^{-\vartheta}\Big) \le Cn^{-1/7}.$$

The ratio formulation follows from Lemmas 10.8 and 10.9 by further intersecting $\mathcal{U}_\vartheta$ and the set therein.

Next we consider $q \in (2,\infty)$. Let $L_n \equiv n^{\vartheta_1}$ for some ϑ_1 to be chosen later. Using Proposition 10.3 and its proofs below (10.8), for $\vartheta_1 > 0$ chosen small enough, we may find some $\mathcal{U}_{\vartheta_1} \subset B_n(1)$ with the desired volume estimate, such that $\sup_{\mu_0\in\mathcal{U}_{\vartheta_1}}\sup_{\eta\in\Xi_K}\|\mathbb{E}\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_\infty \le L_n/\sqrt{n}$, and

$$\sup_{\mu_0\in\mathcal{U}_{\vartheta_1}} \mathbb{P}\bigg(\sup_{\eta\in\Xi_K}\Big\{\|\widehat{\mu}_\eta-\mu_0\|_\infty \vee \|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_\infty\Big\} \ge \frac{L_n}{\sqrt{n}}\bigg) \le Cn^{-2D}, \tag{10.25}$$

where we choose $D > 0$ sufficiently large. Recall for $x \in \mathbb{R}^n$ and $q > 2$, $\|x\|_q \le \|x\|^{2/q}\|x\|_\infty^{1-2/q}$. This motivates the choice

$$\mathsf{g}_q(x) \equiv \bigg[\Big(\frac{L_n}{\sqrt{n}}\Big)^{\frac{2}{q}-1}\bigg\|(x-\mu_0)\wedge\Big(\frac{L_n}{\sqrt{n}}\Big)^{\frac{2}{q}-1}\vee\Big\{-\Big(\frac{L_n}{\sqrt{n}}\Big)^{\frac{2}{q}-1}\Big\}\bigg\|_q\bigg]^{\frac{q}{2}},$$

which verifies that g_q is 1-Lipschitz with respect to $\|\cdot\|$. Using (10.25),

$$\inf_{\mu_0\in\mathcal{U}_{\vartheta_1}} \mathbb{P}\Big(\mathsf{g}_q(\widehat{\mu}_\eta) = n^{(1-\frac{q}{2})\vartheta_1}\cdot\{n^{\frac{1}{2}-\frac{1}{q}}\|\widehat{\mu}_\eta-\mu_0\|_q\}^{\frac{q}{2}}, \forall \eta\in\Xi_K\Big) \ge 1 - Cn^{-D}, \tag{10.26}$$

and with $E_{\mu_0} \equiv \{\sup_{\eta\in\Xi_K}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_\infty \le L_n/\sqrt{n}\}$,

$$\begin{aligned}
&\sup_{\mu_0\in\mathcal{U}_{\vartheta_1}}\sup_{\eta\in\Xi_K}\Big|\mathbb{E}\,\mathsf{g}_q(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}) - n^{(1-\frac{q}{2})\vartheta_1}\,\mathbb{E}\,\{n^{\frac{1}{2}-\frac{1}{q}}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_q\}^{\frac{q}{2}}\Big|\\
&= n^{(1-\frac{q}{2})\vartheta_1}\sup_{\mu_0\in\mathcal{U}_{\vartheta_1}}\sup_{\eta\in\Xi_K}\mathbb{E}\,\{n^{\frac{1}{2}-\frac{1}{q}}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_q\}^{\frac{q}{2}}\mathbf{1}_{E^c_{\mu_0}}\\
&\qquad + \sup_{\mu_0\in\mathcal{U}_{\vartheta_1}}\sup_{\eta\in\Xi_K}\mathbb{E}\,\mathsf{g}_q(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)})\mathbf{1}_{E^c_{\mu_0}} \lesssim n^{-D}.
\end{aligned}\tag{10.27}$$

As the map $g \mapsto \|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_q = \|(\Sigma+\tau_{\eta,*}I)^{-1}(-\tau_{\eta,*}\mu_0+\gamma_{\eta,*}\Sigma^{1/2}g/\sqrt{n})\|_q$ is $Cn^{-1/2}$-Lipschitz with respect to $\|\cdot\|$, Gaussian concentration yields

$$\mathbb{P}\Big(n^{1/2}\Big|\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_q - \mathbb{E}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_q\Big| \ge n^{\vartheta_1}\Big) \le Cn^{-2D}.$$

Using the Lipschitz property of the maps, we may strengthen the above inequality to a uniform control over $\eta \in \Xi_K$. This means uniformly in $\mu_0 \in \mathcal{U}_{\vartheta_1}, \eta\in\Xi_K$,

$$\Big|\mathbb{E}\Big\{n^{\frac{1}{2}-\frac{1}{q}}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_q\Big\}^{\frac{q}{2}} - \Big\{n^{\frac{1}{2}-\frac{1}{q}}\,\mathbb{E}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}-\mu_0\|_q\Big\}^{\frac{q}{2}}\Big| \lesssim n^{-\frac{1}{q}+\vartheta_1}. \tag{10.28}$$

Combining (10.27)-(10.28), we have uniformly in $\mu_0 \in \mathcal{U}_{\vartheta_1}, \eta \in \Xi_K$,

$$\Big|\mathbb{E}\,\mathsf{g}_q(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}) - n^{(1-\frac{q}{2})\vartheta_1}\big\{n^{\frac{1}{2}-\frac{1}{q}}\,\mathbb{E}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_q\big\}^{q/2}\Big| \le Cn^{(2-\frac{q}{2})\vartheta_1-\frac{1}{q}}. \quad (10.29)$$

Combining (10.26) and (10.29) proves the existence of some small ϑ_2 and some $\mathcal{U}_{\vartheta_2} \subset B_n(1)$ with the desired volume estimate, such that

$$\sup_{\mu_0\in\mathcal{U}_{\vartheta_2}} \mathbb{P}\Big(n^{\frac{1}{2}-\frac{1}{q}} \sup_{\eta\in\Xi_K} \big|\|\widehat{\mu}_\eta - \mu_0\|_q - \mathbb{E}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_q\big| \ge n^{-\vartheta_2}\Big) \le Cn^{-1/7}.$$

The ratio formulation follows again from Lemmas 10.8 and 10.9. □

11. Proofs for Section 3

11.1. **Proof of Theorem 3.1.** For ϑ chosen small enough, we fix $\mu_0 \in \mathcal{U}_\vartheta$, where $\mathcal{U}_\vartheta$ is specified in Theorem 2.4. We omit the subscripts in $R^{\#}_{(\Sigma,\mu_0)}(\eta) = R^{\#}(\eta)$, $\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta) = \bar{R}^{\#}(\eta)$, and write $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) = \widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}$ in the proof. All the constants in $\lesssim, \gtrsim, \asymp$ and $\mathcal{O}$ below may possibly depend on K.
(1). Consider the case $\# = \mathsf{pred}$. We omit the superscript pred as well. Using Theorem 2.4-(1) with $\mathsf{g}(x) = \|\Sigma^{1/2}(x-\mu_0)\|$, on an event E_0 with $\mathbb{P}(E_0^c) \le C_0\varepsilon^{-c_0}n^{-1/6.5}$,

$$\sup_{\eta\in\Xi_K} \Big|\sqrt{R(\eta)} - \mathbb{E}\|\Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0)\|\Big| \le \varepsilon.$$

By Gaussian-Poincaré inequality, $0 \le \bar{R}(\eta) - \big(\mathbb{E}\|\Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0)\|\big)^2 = \mathrm{Var}\big(\|\Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0)\|\big) \lesssim n^{-1}$. As $\bar{R}(\eta) \asymp 1$ uniformly in $\eta \in \Xi_K$, on E_0,

$$\sup_{\eta\in\Xi_K} \big|R^{1/2}(\eta) - \bar{R}^{1/2}(\eta)\big| \le \varepsilon + C_0' n^{-1}. \quad (11.1)$$

On the other hand, using both the standard form $\widehat{\mu}_\eta = n^{-1}(X^\top X/n + \eta I_n)^{-1}X^\top Y$ and the alternative form $\widehat{\mu}_\eta = n^{-1}X^\top(XX^\top/n + \eta I_m)^{-1}Y$, we have

$$\sup_{\eta\in\Xi_K} \|\widehat{\mu}_\eta\| \lesssim \Big(\|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}\mathbf{1}^{-1}_{\phi^{-1}\ge 1+1/K} \wedge 1\Big)\cdot\Big(1 + \frac{\|Z\|_{\mathrm{op}} + \|\xi\|}{\sqrt{n}}\Big)^2. \quad (11.2)$$

Consequently, on an event E_1 with $\mathbb{P}(E_1^c) \le C_1 e^{-n/C_1}$,

$$\sup_{\eta\in\Xi_K} \|\widehat{\mu}_\eta\| \le C_1. \quad (11.3)$$

Finally, using (11.1) and (11.3), on $E_0 \cap E_1$,

$$\sup_{\eta\in\Xi_K} |R(\eta) - \bar{R}(\eta)| \lesssim \sup_{\eta\in\Xi_K} |R^{1/2}(\eta) - \bar{R}^{1/2}(\eta)|\Big(1 + \sup_{\eta\in\Xi_K}\|\widehat{\mu}_\eta\|\Big) \lesssim \varepsilon + n^{-1}.$$

The claim follows. The case $\# = \mathsf{est}$ follows from minor modifications so will be omitted.
(2). Consider the case $\# = \mathsf{res}$. We omit the superscript res as well. Further fix $\xi \in \mathcal{E}_\vartheta$ as specified in Theorem 2.4 (the concrete form of $\mathcal{E}_\vartheta$ is given in Proposition 10.3). Using the same Theorem 2.4-(2) with $\mathsf{h}(x) = \|x\|$,

$$\mathbb{P}^\xi\Big(\sup_{\eta\in[1/K,K]} \big|\|\widehat{r}_\eta\| - \mathbb{E}^\xi\|r_{\eta,*}\|\big| \ge \varepsilon\Big) \le C\varepsilon^{-c_0}\cdot n^{-1/6.5}.$$

By Gaussian-Poincaré inequality, $0 \leq \mathbb{E}^{\xi}\|r_{\eta,*}\|^2 - \big(\mathbb{E}^{\xi}\|r_{\eta,*}\|\big)^2 = \mathrm{Var}^{\xi}\big(\|r_{\eta,*}\|\big) \lesssim 1/n$. Combined with the fact that $\mathbb{E}^{\xi}\|r_{\eta,*}\|^2 = (\eta\gamma_{\eta,*}/\tau_{\eta,*})^2 + \mathcal{O}(\big|\|\xi\|^2/m - \sigma_\xi^2\big|)$, for $\eta \in [1/K, K]$, using the stability estimate in Proposition 8.1-(3),

$$\Big|\mathbb{E}^{\xi}\|r_{\eta,*}\| - \eta\gamma_{\eta,*}/\tau_{\eta,*}\Big| \lesssim \Big|\big(\mathbb{E}^{\xi}\|r_{\eta,*}\|\big)^2 - (\eta\gamma_{\eta,*}/\tau_{\eta,*})^2\Big| \lesssim n^{-1/2+\vartheta}.$$

So for $\varepsilon \in (Cn^{-1/2+\vartheta}, 1/C]$,

$$\mathbb{P}^{\xi}\Big(\sup_{\eta\in[1/K,K]}\Big|\|\widehat{r}_\eta\| - \eta\gamma_{\eta,*}/\tau_{\eta,*}\Big| \geq \varepsilon\Big) \leq C\varepsilon^{-c_0}\cdot n^{-1/6.5}.$$

Now taking expectation over ξ, for the same range of ε,

$$\mathbb{P}\Big(\sup_{\eta\in[1/K,K]}\Big|\|\widehat{r}_\eta\| - \eta\gamma_{\eta,*}/\tau_{\eta,*}\Big| \geq \varepsilon\Big) \leq C\varepsilon^{-c_0}\cdot n^{-1/6.5}. \tag{11.4}$$

On the other hand, using (11.2),

$$\sup_{\eta\in[1/K,K]}\|\widehat{r}_\eta\| \lesssim \Big(\|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}\mathbf{1}^{-1}_{\phi^{-1}\geq 1+1/K}\wedge 1\Big)\cdot\Big(1+\frac{\|Z\|_{\mathrm{op}}+\|\xi\|}{\sqrt{n}}\Big)^3.$$

Consequently, on an event E_3 with $\mathbb{P}(E_3^c) \leq C_3 e^{-n/C_3}$, $\sup_{\eta\in[1/K,K]}\|\widehat{r}_\eta\| \leq C_3$, and therefore

$$\sup_{\eta\in[1/K,K]}\Big|\|\widehat{r}_\eta\|^2 - (\eta\gamma_{\eta,*}/\tau_{\eta,*})^2\Big| \leq C_3\cdot\sup_{\eta\in[1/K,K]}\Big|\|\widehat{r}_\eta\| - \eta\gamma_{\eta,*}/\tau_{\eta,*}\Big|.$$

The claim follows. The case $\# = \mathsf{in}$ proceeds similarly, but with the function now taken as $\mathsf{h}(x) = \|x - \xi/\sqrt{n}\|$, and the claim follows by computing that

$$\begin{aligned}\mathbb{E}^{\xi}\|r_{\eta,*} - \xi/\sqrt{n}\|^2 &= \phi\cdot\Big\{\Big(\frac{\eta}{\phi\tau_{\eta,*}}\Big)^2(\phi\gamma_{\eta,*}^2 - \sigma_\xi^2) + \frac{\|\xi\|^2}{m}\cdot\Big(\frac{\eta}{\phi\tau_{\eta,*}} - 1\Big)^2\Big\}\\ &= \Big(\frac{\eta\gamma_{\eta,*}}{\tau_{\eta,*}}\Big)^2 + \phi\sigma_\xi^2\cdot\Big[\Big(\frac{\eta}{\phi\tau_{\eta,*}} - 1\Big)^2 - \Big(\frac{\eta}{\phi\tau_{\eta,*}}\Big)^2\Big] + \mathcal{O}(\big|\|\xi\|^2/m - \sigma_\xi^2\big|).\end{aligned}$$

The proof is complete. □

11.2. **Proof of Theorem 3.2.**

Lemma 11.1. *Suppose $1/K \leq \phi^{-1} \leq K$, and $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \leq K$ for some $K > 0$. Then with $g \sim \mathcal{N}(0, I_n)$, there exists some $C = C(K) > 0$ such that for $\varepsilon \in (0,1)$, and $q \in \{0, 1/2\}$,*

$$\mathbb{P}\Big(\sup_{\eta\in\Xi_K}\Big|\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^q g/\|g\|\|^2 - n^{-1}\,\mathrm{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma^{2q}\big)\Big| > \varepsilon\Big) \leq C\varepsilon^{-1}e^{-n\varepsilon^2/C}.$$

Proof. We only prove the case $q = 1/2$. All the constants in $\lesssim, \gtrsim, \asymp$ below may depend on K. We write $A_\eta \equiv (\Sigma + \tau_{\eta,*}I)^{-2}\Sigma$ for notational simplicity. Note that

$$\begin{aligned}&\Big|\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^{1/2}g/\|g\|\|^2 - n^{-1}\,\mathrm{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big)\Big|\\ &= n^{-1}\Big|e_g^{-2}\|A_\eta^{1/2}g\|^2 - \mathbb{E}\|A_\eta^{1/2}g\|^2\Big|\\ &\lesssim e_g^{-2}\cdot n^{-1}\Big|\|A_\eta^{1/2}g\|^2 - \mathbb{E}\|A_\eta^{1/2}g\|^2\Big| + |e_g^{-2} - 1|.\end{aligned}$$

Here in the last inequality we used $\mathbb{E}\|A_\eta^{1/2}g\|^2 \lesssim n$. As

- $\|A_\eta\|_F^2 = \mathrm{tr}\big((\Sigma+\tau_{\eta,*}I)^{-4}\Sigma^2\big) \lesssim n(1\wedge\tau_{\eta,*})^{-4} \asymp n$, and

- $\|A_\eta\|_F^2 \gtrsim \mathrm{tr}(\Sigma^2)\cdot(1\vee\tau_{\eta,*})^{-4} \gtrsim n$,

we have uniformly in $\eta\in[0,K]$, $\|A_\eta\|_F \asymp \sqrt{n}$. It is easy to see that $\|A_\eta\|_{\mathrm{op}} \asymp 1$. So by Hanson-Wright inequality, there exists some constant $C_1 = C_1(K)$ such that for $\varepsilon\in(0,1)$,

$$
\begin{aligned}
&\mathbb{P}\Big(\big|\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^{1/2}g/\|g\|\|^2 - n^{-1}\,\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma)\big| > \varepsilon\Big)\\
&\leq \mathbb{P}\Big(\big|n^{-1}(\|A_\eta^{1/2}g\|^2 - \mathbb{E}\|A_\eta^{1/2}g\|^2)\big| > \varepsilon/4\Big) + \mathbb{P}\left(|e_g^{-2}-1| > \varepsilon/2\right) + \mathbb{P}(e_g^2 \leq 1/2)\\
&\leq C_1 e^{-n\varepsilon^2/C_1}.
\end{aligned}
$$

On the other hand, for any $\eta_1,\eta_2\in\Xi_K$, using Proposition 8.1-(3),

$$
\begin{aligned}
\big|\|(\Sigma+\tau_{\eta_1,*}I)^{-1}\Sigma^{1/2}g/\|g\|\|^2 - \|(\Sigma+\tau_{\eta_2,*}I)^{-1}\Sigma^{1/2}g/\|g\|\|^2\big| &\lesssim |\eta_1-\eta_2|,\\
n^{-1}\big|\,\mathrm{tr}\,((\Sigma+\tau_{\eta_1,*}I)^{-2}\Sigma) - \mathrm{tr}\,((\Sigma+\tau_{\eta_2,*}I)^{-2}\Sigma)\big| &\lesssim |\eta_1-\eta_2|,
\end{aligned}
$$

so we may conclude by a standard discretization and union bound argument. □

Proposition 11.2. *The following hold with $\mathfrak{m}_\eta \equiv \mathfrak{m}(-\eta/\phi)$, $\mathfrak{m}'_\eta \equiv \mathfrak{m}'(-\eta/\phi)$.*

(1) $\tau_{\eta,*} = 1/\mathfrak{m}_\eta$ and $\partial_\eta\tau_{\eta,*} = \mathfrak{m}'_\eta/(\phi\mathfrak{m}_\eta^2)$.

(2) It holds that

$$
\begin{aligned}
\frac{1}{n}\,\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma) &= \frac{\phi\mathfrak{m}_\eta^2}{\mathfrak{m}'_\eta}(\mathfrak{m}_\eta - (\eta/\phi)\mathfrak{m}'_\eta),\\
\frac{1}{n}\,\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-2}) &= \frac{\phi\mathfrak{m}_\eta^2}{\mathfrak{m}'_\eta}((\phi^{-1}-1)\mathfrak{m}'_\eta + 2(\eta/\phi)\cdot\mathfrak{m}_\eta\mathfrak{m}'_\eta - \mathfrak{m}_\eta^2).
\end{aligned}
$$

(3) Suppose $1/K \leq \phi^{-1} \leq K$, and $\|\Sigma\|_{\mathrm{op}}\vee\mathcal{H}_\Sigma \leq K$ for some $K>0$. There exists some constant $C = C(K) > 0$ such that the following hold. For any $\varepsilon\in(0,1/2]$, for some $\mathcal{U}_\varepsilon\subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_\varepsilon)/\mathrm{vol}(B_n(1)) \geq 1 - C\varepsilon^{-1}e^{-n\varepsilon^2/C}$,

$$
\sup_{\mu_0\in\mathcal{U}_\varepsilon}\sup_{\eta\in\Xi_K}\left|\gamma_{\eta,*}^2 - \frac{\sigma_\xi^2\mathfrak{m}'_\eta + \|\mu_0\|^2(\phi\mathfrak{m}_\eta - \eta\mathfrak{m}'_\eta)}{\phi\mathfrak{m}_\eta^2}\right| \leq \varepsilon.
$$

When $\Sigma = I_n$, we may take $\mathcal{U}_\varepsilon = B_n(1)$ and the above inequality holds with $\varepsilon = 0$.

Proof. (1) follows from definition so we focus on (2)-(3).
(2). Differentiating both sides of (2.3) with respect to η yields that

$$
-n^{-1}\,\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma)\cdot\partial_\eta\tau_{\eta,*} = -(\mathfrak{m}_\eta - (\eta/\phi)\mathfrak{m}'_\eta).
$$

Now using $\partial_\eta\tau_{\eta,*} = \mathfrak{m}'_\eta/(\phi\mathfrak{m}_\eta^2)$ to obtain the formula for $n^{-1}\,\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma)$.

Next, using that $\phi - \frac{\eta}{\tau_{\eta,*}} = n^{-1}\,\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-1}\Sigma) = 1 - \tau_{\eta,*}\cdot n^{-1}\,\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-1})$, we may solve

$$
n^{-1}\,\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-1}) = \mathfrak{m}_\eta(1-\phi+\eta\cdot\mathfrak{m}_\eta).
$$

Differentiating with respect to η on both sides of the above display, we obtain

$$
\begin{aligned}
-n^{-1}\,\mathrm{tr}\,((\Sigma+\tau_{\eta,*}I)^{-2})\cdot\partial_\eta\tau_{\eta,*} &= -\phi^{-1}\mathfrak{m}'_\eta(1-\phi+\eta\cdot\mathfrak{m}_\eta) + \mathfrak{m}_\eta\cdot(\mathfrak{m}_\eta - (\eta/\phi)\mathfrak{m}'_\eta)\\
&= -(\phi^{-1}-1)\mathfrak{m}'_\eta - 2(\eta/\phi)\cdot\mathfrak{m}_\eta\mathfrak{m}'_\eta + \mathfrak{m}_\eta^2,
\end{aligned}
$$

proving the second identity.

(3). Let $\mu_0 \equiv U_0 g_0/\|g_0\|$, where $U_0 \sim \mathrm{Unif}[0,1]$ and $g_0 \sim \mathcal{N}(0, I_n)$ are independent variables. Then μ_0 is uniformly distributed on $B_n(1)$. For some $\varepsilon > 0$ to be chosen later, let

$$\mathcal{G}_\varepsilon \equiv \Big\{g \in \mathbb{R}^n : \sup_{\eta\in\Xi_K} \Big|\Big\|(\Sigma + \tau_{\eta,*}I)^{-1}\Sigma^{1/2}\frac{g}{\|g\|}\Big\|^2 - \frac{1}{n}\operatorname{tr}\big((\Sigma + \tau_{\eta,*}I)^{-2}\Sigma\big)\Big| \leq \varepsilon\Big\}. \quad (11.5)$$

Let $\mathcal{U}_\varepsilon \equiv \{Ug/\|g\| : U \in [0,1], g \in \mathcal{G}_\varepsilon\} \subset B_n(1)$. Using Lemma 11.1, there exists some constant $C_0 = C_0(K) > 0$ such that $\mathrm{vol}(\mathcal{U}_\varepsilon)/\mathrm{vol}(B_n(1)) = \mathbb{P}_{\mu_0}(\mu_0 \in \mathcal{U}_\varepsilon) \geq 1 - C_0\varepsilon^{-1}e^{-n\varepsilon^2/C_0}$, and moreover,

$$\sup_{\mu_0\in\mathcal{U}_\varepsilon}\sup_{\eta\in\Xi_K}\Big|\|(\Sigma + \tau_{\eta,*}I)^{-1}\Sigma^{1/2}\mu_0\|^2 - \|\mu_0\|^2 \cdot n^{-1}\operatorname{tr}\big((\Sigma + \tau_{\eta,*}I)^{-2}\Sigma\big)\Big| \leq \varepsilon.$$

Note that when $\Sigma = I_n$, the above estimate holds for all $\mu_0 \in B_n(1)$ with $\varepsilon = 0$.

Combining the above display with the formula (8.3) for $\gamma_{\eta,*}^2$, and the fact that the denominator therein is of order 1 (depending on K), we have

$$\sup_{\mu_0\in\mathcal{U}_\varepsilon}\sup_{\eta\in\Xi_K}\left|\gamma_{\eta,*}^2 - \frac{\sigma_\xi^2 + \|\mu_0\|^2\tau_{\eta,*}^2 \cdot \frac{1}{n}\operatorname{tr}\big((\Sigma + \tau_{\eta,*}I)^{-2}\Sigma\big)}{\frac{\eta}{\tau_{\eta,*}} + \tau_{\eta,*}\cdot\frac{1}{n}\operatorname{tr}\big((\Sigma + \tau_{\eta,*}I)^{-2}\Sigma\big)}\right| \leq C_1\varepsilon.$$

Now using (2), the second term in the above display equals to

$$\frac{\sigma_\xi^2 + \|\mu_0\|^2 \cdot \frac{\phi}{\mathfrak{m}_\eta'}(\mathfrak{m}_\eta - \frac{\eta}{\phi}\mathfrak{m}_\eta')}{\eta\mathfrak{m}_\eta + \frac{\phi\mathfrak{m}_\eta}{\mathfrak{m}_\eta'}(\mathfrak{m}_\eta - \frac{\eta}{\phi}\mathfrak{m}_\eta')} = \frac{\sigma_\xi^2\mathfrak{m}_\eta' + \|\mu_0\|^2(\phi\mathfrak{m}_\eta - \eta\mathfrak{m}_\eta')}{\phi\mathfrak{m}_\eta^2}.$$

The claim follows by adjusting constants. □

Proof of Theorem 3.2. As $\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) = \phi\gamma_{\eta,*}^2 - \sigma_\xi^2$, directly invoking Proposition 11.2-(3) yields the claim for $\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$.

Next we handle $\bar{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta)$. Note that

$$\bar{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) = \tau_{\eta,*}^2\|(\Sigma + \tau_{\eta,*}I)^{-1}\mu_0\|^2 + \gamma_{\eta,*}^2 \cdot n^{-1}\operatorname{tr}\big((\Sigma + \tau_{\eta,*}I)^{-2}\Sigma\big). \quad (11.6)$$

Using a similar construction as in the proof of Proposition 11.2 via the help of Lemma 11.1, this time with $q = 0$ therein, we may find some $\mathcal{U}_\varepsilon \subset B_n(1)$ with the desired volume estimate, such that both Proposition 11.2-(3) and

$$\sup_{\mu_0\in\mathcal{U}_\varepsilon}\sup_{\eta\in\Xi_K}\Big|\|(\Sigma + \tau_{\eta,*}I)^{-1}\mu_0\|^2 - \|\mu_0\|^2 \cdot n^{-1}\operatorname{tr}\big((\Sigma + \tau_{\eta,*}I)^{-2}\big)\Big| \leq \varepsilon \quad (11.7)$$

hold. Combining (11.6)-(11.7), we may set

$$\begin{aligned}\mathscr{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) &\equiv \tau_{\eta,*}^2\|\mu_0\|^2 \cdot n^{-1}\operatorname{tr}\big((\Sigma + \tau_{\eta,*}I)^{-2}\big)\\ &\qquad + (\phi\mathfrak{m}_\eta^2)^{-1}\Big(\sigma_\xi^2\mathfrak{m}_\eta' + \|\mu_0\|^2(\phi\mathfrak{m}_\eta - \eta\mathfrak{m}_\eta')\Big) \cdot n^{-1}\operatorname{tr}\big((\Sigma + \tau_{\eta,*}I)^{-2}\Sigma\big)\\ &\equiv R_{2,1} + R_{2,2}.\end{aligned}$$

By Proposition 11.2-(2), we may compute $R_{2,1}, R_{2,2}$ separately:

$$R_{2,1} = \|\mu_0\|^2 \cdot \frac{\phi}{\mathfrak{m}_\eta'} \cdot \Big((\phi^{-1} - 1)\mathfrak{m}_\eta' + 2(\eta/\phi)\cdot\mathfrak{m}_\eta\mathfrak{m}_\eta' - \mathfrak{m}_\eta^2\Big)$$

$$= \|\mu_0\|^2(1-\phi) + \Big\{2\|\mu_0\|^2\eta\mathfrak{m}_\eta - \|\mu_0\|^2\phi\cdot\frac{\mathfrak{m}_\eta^2}{\mathfrak{m}_\eta'}\Big\},$$

$$R_{2,2} = \frac{1}{\mathfrak{m}_\eta'}\Big(\sigma_\xi^2\mathfrak{m}_\eta' + \|\mu_0\|^2(\phi\mathfrak{m}_\eta - \eta\mathfrak{m}_\eta')\Big)\cdot(\mathfrak{m}_\eta - (\eta/\phi)\mathfrak{m}_\eta')$$

$$= \sigma_\xi^2(\mathfrak{m}_\eta - (\eta/\phi)\mathfrak{m}_\eta') + \phi^{-1}\|\mu_0\|^2\eta^2\mathfrak{m}_\eta' - \Big\{2\|\mu_0\|^2\eta\mathfrak{m}_\eta - \|\mu_0\|^2\phi\cdot\frac{\mathfrak{m}_\eta^2}{\mathfrak{m}_\eta'}\Big\}.$$

Consequently,

$$\begin{aligned}\mathscr{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) &= \|\mu_0\|^2(1-\phi) + \sigma_\xi^2(\mathfrak{m}_\eta - (\eta/\phi)\mathfrak{m}_\eta') + \phi^{-1}\|\mu_0\|^2\eta^2\mathfrak{m}_\eta'\\ &= \sigma_\xi^2\cdot\Big\{\mathsf{SNR}_{\mu_0}(1-\phi) + \mathfrak{m}_\eta + (\eta/\phi)(\eta\cdot\mathsf{SNR}_{\mu_0}-1)\mathfrak{m}_\eta'\Big\}.\end{aligned}$$

The claims for $\mathscr{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta)$ and $\mathscr{R}^{\mathsf{res}}_{(\Sigma,\mu_0)}(\eta)$ follow from Proposition 11.2-(3). □

11.3. **Proof of Proposition 3.3.** We will prove the following version of Proposition 3.3, where $\mathfrak{M}^{\#}$ is represented via $\tau_{\eta,*}$ instead of $\mathfrak{m}$. In the proof below, we will also verify the representation of $\mathfrak{M}^{\#}$ via $\mathfrak{m}$ as stated in Proposition 3.3.

Proposition 11.3. *Recall* $\mathsf{SNR}_{\mu_0} = \|\mu_0\|^2/\sigma_\xi^2$. *Then for* $\# \in \{\mathsf{pred},\mathsf{est},\mathsf{in}\}$,

$$\partial_\eta\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) = \sigma_\xi^2\cdot\mathfrak{M}^{\#}(\eta)\cdot(\eta\cdot\mathsf{SNR}_{\mu_0}-1).$$

Here with $T_{-p,q}(\eta) \equiv n^{-1}\operatorname{tr}\big((\Sigma+\tau_*(\eta)I)^{-p}\Sigma^q\big)$ *for* $p,q\in\mathbb{N}$,

$$\mathfrak{M}^{\#}(\eta) \equiv \begin{cases}\phi\big(-\tau_*''(\eta)\big), & \# = \mathsf{pred};\\ 2(\tau_*'(\eta))^2\big(T_{-3,1}(\eta) + \tau_*'(\eta)T_{-2,1}(\eta)T_{-3,2}(\eta)\big), & \# = \mathsf{est};\\ \frac{2(\tau_*'(\eta))^2}{\tau_*^2(\eta)}\Big(\eta^2\tau_*'(\eta)T_{-3,2}(\eta) + \tau_*^3(\eta)T_{-2,1}^2(\eta)\Big), & \# = \mathsf{in}.\end{cases}$$

Suppose further $1/K \le \phi^{-1} \le K$ *and* $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ *for some* $K>0$. *Then there exists some* $C = C(K) > 0$ *such that uniformly in* $\eta\in\Xi_K$ *and for all* $\#\in\{\mathsf{pred},\mathsf{est},\mathsf{in}\}$,

(1) $1/C \le \mathfrak{M}^{\#}(\eta) \le C$, *and*

(2) if additionally $\eta_* \equiv \mathsf{SNR}_{\mu_0}^{-1} \in \Xi_K$,

$$1/C \le \frac{|\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta_*)|}{\|\mu_0\|^2(\eta-\eta_*)^2} \le C.$$

Proof. In the proof we write $\tau_{\eta,*} = \tau_\eta$. Recall the notation $\mathfrak{m}_\eta = \mathfrak{m}(-\eta/\phi)$, $\mathfrak{m}_\eta' = \mathfrak{m}'(-\eta/\phi)$, and we naturally write $\mathfrak{m}_\eta'' \equiv \mathfrak{m}''(-\eta/\phi)$. By differentiating with respect to η for both sides of $\mathfrak{m}_\eta = 1/\tau_\eta$, with some calculations we have

$$\mathfrak{m}_\eta = \tau_\eta^{-1},\quad \mathfrak{m}_\eta' = \phi\tau_\eta'/\tau_\eta^2,\quad \mathfrak{m}_\eta'' = -\phi^2(\tau_\eta''\tau_\eta - 2(\tau_\eta')^2)/\tau_\eta^3. \tag{11.8}$$

Using ρ, we may also write $\mathfrak{m}_\eta^{(q)} = q!\int\frac{\rho(\mathrm{d}x)}{(x+\eta/\phi)^{q+1}}$ for $q\in\mathbb{N}$. Here by convention $0! = 1$.

(1). Using the formula for $\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}$,

$$\partial_\eta\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) = \sigma_\xi^2\cdot\partial_\eta\Big\{\mathfrak{m}_\eta^{-2}(\phi\cdot\mathsf{SNR}_{\mu_0}\mathfrak{m}_\eta - (\eta\cdot\mathsf{SNR}_{\mu_0}-1)\mathfrak{m}_\eta')\Big\}$$

$$= \phi^{-1}\sigma_\xi^2 \cdot \mathfrak{m}_\eta^{-3}(\mathfrak{m}_\eta \mathfrak{m}_\eta'' - 2(\mathfrak{m}_\eta')^2) \cdot (\eta \cdot \mathsf{SNR}_{\mu_0} - 1).$$

Some calculations show that

$$\mathfrak{m}_\eta \mathfrak{m}_\eta'' - 2(\mathfrak{m}_\eta')^2 = \tau_\eta^{-3}\phi^2(-\tau_\eta'') = 2\Big\{\int \frac{\rho(\mathrm{d}x)}{(x+\eta/\phi)}\int \frac{\rho(\mathrm{d}x)}{(x+\eta/\phi)^3} - \Big(\int \frac{\rho(\mathrm{d}x)}{(x+\eta/\phi)^2}\Big)^2\Big\},$$

so the identity follows.
(2). Using the formula for $\mathscr{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}$,

$$\begin{aligned}\partial_\eta \mathscr{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) &= \sigma_\xi^2 \cdot \partial_\eta\big(\mathsf{SNR}_{\mu_0}(1-\phi) + \mathfrak{m}_\eta + (\eta/\phi)(\eta\cdot\mathsf{SNR}_{\mu_0} - 1)\mathfrak{m}_\eta'\big)\\ &= \phi^{-1}\sigma_\xi^2\cdot(2\mathfrak{m}_\eta' - (\eta/\phi)\mathfrak{m}_\eta'')\cdot(\eta\cdot\mathsf{SNR}_{\mu_0} - 1). \end{aligned} \tag{11.9}$$

To compute the second term in the above display, recall the identity for τ_η', τ_η'' in (8.5)-(8.6). Also recall $G_0(\eta) = \eta + \tau_\eta^2 T_{-2,1}(\eta) = \tau_\eta/\tau_\eta'$ defined in (8.5). Then

$$\begin{aligned} 2\mathfrak{m}_\eta' - \frac{\eta}{\phi}\mathfrak{m}_\eta'' &= \frac{\phi}{\tau_\eta}\Big\{\frac{2\tau_\eta'}{\tau_\eta}\Big(1 - \frac{\eta\tau_\eta'}{\tau_\eta}\Big) + \frac{\eta\tau_\eta''}{\tau_\eta}\Big\}\\ &= \frac{\phi}{\tau_\eta}\Big\{\frac{2\tau_\eta'}{\tau_\eta}\frac{\tau_\eta^2 T_{-2,1}(\eta)}{G_0(\eta)} - \frac{2\eta\tau_\eta\tau_\eta'}{G_0^2(\eta)}T_{-3,2}(\eta)\Big\}\\ &= \frac{2\phi\tau_\eta'}{G_0(\eta)}\Big(T_{-2,1}(\eta) - \frac{\eta}{G_0(\eta)}T_{-3,2}(\eta)\Big)\\ &\overset{(*)}{=} \frac{2\phi\tau_\eta'}{G_0(\eta)}\Big\{\tau_\eta T_{-3,1}(\eta) + \Big(1 - \frac{\eta}{G_0(\eta)}\Big)T_{-3,2}(\eta)\Big\}\\ &= 2\phi(\tau_\eta')^2(T_{-3,1}(\eta) + \tau_\eta' T_{-2,1}(\eta)T_{-3,2}(\eta)). \end{aligned}$$

Here in $(*)$ we used $T_{-2,1}(\eta) - T_{-3,2}(\eta) = \tau_\eta T_{-3,1}(\eta)$. The claimed identity follows by combining the above display and (11.9). Using ρ, we may write

$$2\mathfrak{m}_\eta' - (\eta/\phi)\mathfrak{m}_\eta'' = 2\int \frac{x}{(x+\eta/\phi)^3}\,\rho(\mathrm{d}x).$$

(3). Using the formula for $\mathscr{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}$,

$$\begin{aligned}\partial_\eta \mathscr{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta) &= \sigma_\xi^2\cdot\partial_\eta\Big\{\phi^{-1}\eta^2(\phi\cdot\mathsf{SNR}_{\mu_0}\mathfrak{m}_\eta - (\eta\cdot\mathsf{SNR}_{\mu_0} - 1)\mathfrak{m}_\eta') + (\phi - 2\eta\mathfrak{m}_\eta)\Big\}\\ &= \sigma_\xi^2\cdot(2\mathfrak{m}_\eta - 4(\eta/\phi)\mathfrak{m}_\eta' + \phi^{-2}\eta^2\mathfrak{m}_\eta'')\cdot(\eta\cdot\mathsf{SNR}_{\mu_0} - 1). \end{aligned}\tag{11.10}$$

The second term in the above display requires some non-trivial calculations:

$$\begin{aligned} 2\mathfrak{m}_\eta - \frac{4\eta}{\phi}\mathfrak{m}_\eta' + \frac{\eta^2}{\phi^2}\mathfrak{m}_\eta'' &= \frac{1}{\tau_\eta}\Big\{2 - 4\eta\frac{\tau_\eta'}{\tau_\eta} - \eta^2\frac{\tau_\eta''}{\tau_\eta} + 2\eta^2\Big(\frac{\tau_\eta'}{\tau_\eta}\Big)^2\Big\}\\ &= \frac{1}{\tau_\eta}\Big\{2 - \frac{4\eta}{G_0(\eta)} + \eta^2\frac{2\tau_\eta\tau_\eta' T_{-3,2}(\eta)}{G_0^2(\eta)} + \frac{2\eta^2}{G_0^2(\eta)}\Big\}\\ &= \frac{2}{\tau_\eta G_0^2(\eta)}\{G_0^2(\eta) - 2\eta G_0(\eta) + \eta^2\tau_\eta\tau_\eta' T_{-3,2}(\eta) + \eta^2\}. \end{aligned}$$

Expanding the $G_0(\eta)$ terms in the bracket using $G_0(\eta) = \eta + \tau_\eta^2 T_{-2,1}(\eta)$, with some calculations we arrive at

$$2\mathfrak{m}_\eta - \frac{4\eta}{\phi}\mathfrak{m}'_\eta + \frac{\eta^2}{\phi^2}\mathfrak{m}''_\eta = \frac{2}{G_0^2(\eta)}\Big(\eta^2\tau'_\eta T_{-3,2}(\eta) + \tau_\eta^3 T_{-2,1}^2(\eta)\Big).$$

The claimed identity follows by combining the above display and (11.10). Using ρ, we may write

$$2\mathfrak{m}_\eta - \frac{4\eta}{\phi}\mathfrak{m}'_\eta + \frac{\eta^2}{\phi^2}\mathfrak{m}''_\eta = 2\int \frac{x^2}{(x+\eta/\phi)^3}\,\rho(\mathrm{d}x).$$

Finally, the claimed first two-sided bound on $\mathfrak{M}^{\#}$ follows from Proposition 8.1, and the second bound follows by using the fundamental theorem of calculus. □

11.4. **Proof of Theorem 3.4.** The following lemma gives a technical extension of Theorem 3.1 for $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$ under $\sigma_\xi^2 \approx 0$ when $\phi^{-1} > 1$. For $\# = \mathsf{in}$, the extension also allows uniform control over $\eta \approx 0$ under both the above small variance scenario with $\phi^{-1} > 1$, and under the original conditions.

Lemma 11.4. *Suppose Assumption A holds and the following hold for some $K > 0$.*

- $1 + 1/K \le \phi^{-1} \le K$, $\|\Sigma^{-1}\|_{\mathrm{op}} \vee \|\Sigma\|_{\mathrm{op}} \le K$.
- *Assumption B with $\sigma_\xi^2 \in [0, K]$.*

Fix a small enough $\vartheta \in (0, 1/50)$. Then there exist a constant $C = C(K, \vartheta) > 1$, and a measurable set $\mathcal{U}_\vartheta \subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \ge 1 - Ce^{-n^{\vartheta}/C}$, such that for any $\varepsilon \in (0, 1/2]$, and $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}, \mathsf{res}\}$,

$$\sup_{\mu_0 \in \mathcal{U}_\vartheta} \mathbb{P}\bigg(\sup_{\eta \in \Xi_K} |R^{\#}_{(\Sigma,\mu_0)}(\eta, \sigma_\xi) - \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta, \sigma_\xi)| \ge \varepsilon\bigg) \le C \cdot \begin{cases} ne^{-n\varepsilon^{c_0}/C}, & Z = G; \\ \varepsilon^{-c_0} n^{-1/6.5}, & \textit{otherwise}. \end{cases}$$

Proof. All the constants in $\lesssim, \gtrsim, \asymp$ below may possibly depend on K.
(**Part 1**). We shall first extend the claim of Theorem 3.1 for $\# = \mathsf{pred}$ to $\sigma_\xi^2 \in [0, K]$ in the case $\phi^{-1} \ge 1 + 1/K$. Note that uniformly in $\eta \in [0, K]$, for $\sigma_\xi, \sigma'_\xi \in [0, K]$,

$$\|\widehat{\mu}_\eta(\sigma_\xi) - \widehat{\mu}_\eta(\sigma'_\xi)\| \lesssim |\sigma_\xi - \sigma'_\xi| \cdot n^{-1}\|Z\|_{\mathrm{op}}\|\xi_0\| \cdot \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}. \tag{11.11}$$

Using the estimate (11.2), uniformly in $\eta \in [0, K]$, for all $\sigma_\xi, \sigma'_\xi \in [0, K]$,

$$\begin{aligned}
&|R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma_\xi) - R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma'_\xi)| \\
&\lesssim \|\widehat{\mu}_\eta(\sigma_\xi) - \widehat{\mu}_\eta(\sigma'_\xi)\| \cdot \big(\|\widehat{\mu}_\eta(\sigma_\xi)\| + \|\widehat{\mu}_\eta(\sigma'_\xi)\| + \|\mu_0\|\big) \\
&\lesssim |\sigma_\xi - \sigma'_\xi| \cdot \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}^2 \cdot \Big(1 + \frac{\|Z\|_{\mathrm{op}} + \|\xi_0\|}{\sqrt{n}}\Big)^4.
\end{aligned}$$

So on an event E_1 with $\mathbb{P}(E_1) \ge 1 - C_1 e^{-n/C_1}$, for $\sigma_\xi, \sigma'_\xi \in [0, K]$,

$$\sup_{\eta \in [0,K]} |R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma_\xi) - R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma'_\xi)| \le C_1 \cdot |\sigma_\xi - \sigma'_\xi|.$$

On the other hand, using Lemma 11.5-(2),

$$\sup_{\eta \in [0,K]} |\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma_\xi) - \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma'_\xi)| \le C_1 \cdot |\sigma_\xi - \sigma'_\xi|.$$

Using the above two displays, for any $\varepsilon > 0$, by choosing $\sigma'_\xi \equiv \varepsilon/(2C_1)$, we have for any $\sigma_\xi \le \sigma'_\xi$,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{\eta\in[0,K]} |R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi) - \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi)| \ge 2\varepsilon\Big)\\ &\le \mathbb{P}\Big(\sup_{\eta\in[0,K]} |R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta,\sigma'_\xi) - \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta,\sigma'_\xi)| \ge \varepsilon\Big) + C_1 e^{-n/C_1}. \qquad (11.12)\end{aligned}$$

The first term on the right hand side of the above display can be handled by the proven claim in Theorem 3.1, upon noting that (i) the constant C therein depends on K polynomially, and here we choose K to be larger than $2C_1/\varepsilon$; (ii) $(n/\varepsilon)^{C'} e^{-n\varepsilon^{C'}} \wedge 1 \le ne^{-n\varepsilon^{C''}}$ holds for C'' chosen much larger than C'.

The extension of the claim of Theorem 3.1 for $\# = \mathsf{est}$ to $\sigma_\xi^2 \in [0,K]$ follows a similar proof with minor modifications, so we omit the details.

(**Part 2**). Next we consider the case $\# = \mathsf{in}$. We need to extend the corresponding claim of Theorem 3.1 to both $\sigma_\xi^2 \in [0,K]$ and $\eta \in [0,K]$.

We first verify the (high probability) Lipschitz continuity of the maps $\sigma_\xi \mapsto R^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi), \bar{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi)$. Note that uniformly in $\eta \in [0,K]$, by virtue of (11.11), for any $\sigma_\xi, \sigma'_\xi \in [0,K]$,

$$\begin{aligned}&|R^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi) - R^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta,\sigma'_\xi)|\\ &\lesssim \Big(1 + \frac{\|Z\|_{\mathrm{op}}}{\sqrt{n}}\Big)^2 \cdot \|\widehat{\mu}_\eta(\sigma_\xi) - \widehat{\mu}_\eta(\sigma'_\xi)\| \cdot \big(\|\widehat{\mu}_\eta(\sigma_\xi)\| + \|\widehat{\mu}_\eta(\sigma'_\xi)\| + \|\mu_0\|\big)\\ &\lesssim |\sigma_\xi - \sigma'_\xi| \cdot \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}^2 \cdot \Big(1 + \frac{\|Z\|_{\mathrm{op}} + \|\xi_0\|}{\sqrt{n}}\Big)^6.\end{aligned}$$

This verifies the high probability Lipschitz property of $\sigma_\xi \mapsto R^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi)$. The Lipschitz property of $\sigma_\xi \mapsto \bar{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi)$ is easily verified. From here we may use a similar argument to (11.12) to conclude the extension of the claim of Theorem 3.1 for $\# = \mathsf{in}$ to $\sigma_\xi^2 \in [0,K]$.

Finally we verify the (high probability) Lipschitz continuity of the maps $\eta \mapsto R^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi), \bar{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi)$. Using the estimates (9.29) (with G replaced by Z) and (11.2), uniformly in $\sigma_\xi \in [0,K]$ and $\eta_1,\eta_2 \in [0,K]$,

$$\begin{aligned}&|R^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta_1,\sigma_\xi) - R^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta_2,\sigma_\xi)|\\ &\lesssim \Big(1 + \frac{\|Z\|_{\mathrm{op}}}{\sqrt{n}}\Big)^2 \cdot \|\widehat{\mu}_{\eta_1}(\sigma_\xi) - \widehat{\mu}_{\eta_2}(\sigma_\xi)\| \cdot \big(\|\widehat{\mu}_{\eta_1}(\sigma_\xi)\| + \|\widehat{\mu}_{\eta_2}(\sigma_\xi)\| + \|\mu_0\|\big)\\ &\lesssim \Big(1 + \frac{\|Z\|_{\mathrm{op}} + \|\xi_0\|}{\sqrt{n}}\Big)^6 \cdot \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}^3 \cdot |\eta_1 - \eta_2|.\end{aligned}$$

The Lipschitz property of $\sigma_\xi \mapsto \bar{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta,\sigma_\xi)$ is again easily verified. Again from here we may argue similarly to (11.12) to extend the claim of Theorem 3.1 for $\# = \mathsf{in}$ to $\eta \in [0,K]$. The case for $\# = \mathsf{res}$ is similar so we omit repetitive details. □

Lemma 11.5. *Suppose $\phi^{-1} > 1$. The following hold.*

(1) The system of equations

$$\begin{cases}\phi\gamma^2 = \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau),\\ \phi - \frac{\eta}{\tau} = \gamma^{-2}\,\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau) = \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau I)^{-1}\Sigma\big)\end{cases}$$

admit a unique solution $(\gamma_{\eta,*}(0), \tau_{\eta,*}(0)) \in [0,\infty)\times(0,\infty)$.

(2) It holds that $\tau_{\eta,*}(0) = \tau_{\eta,*}(\sigma_\xi)$. *If furthermore* $1 + 1/K \leq \phi^{-1} \leq K$ *and* $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \leq K$ *for some* $K > 0$, *then there exists some* $C = C(K) > 0$ *such that* $|\gamma^2_{\eta,*}(\sigma_\xi) - \gamma^2_{\eta,*}(0)| \leq C\sigma_\xi^2$.

Proof. The claim (1) follows verbatim from the proof of Proposition 8.1-(1) by setting $\sigma_\xi^2 = 0$ therein. The claim (2) follows by using the formula (8.3). □

Proof of Theorem 3.4. Let $\mathcal{U}_\vartheta \subset B_n(1)$ be as specified in Theorem 3.1 or 2.4. In view of its explicit form given in Proposition 10.3, with $\mathcal{U}_{\delta,\vartheta} \equiv \mathcal{U}_\vartheta \cap (B_n(1)\setminus B_n(\delta))$, the volume estimates $\min\{\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)), \mathrm{vol}(\mathcal{U}_{\delta,\vartheta})/\mathrm{vol}(B_n(1)\setminus B_n(\delta))\} \geq 1 - Ce^{-n^\vartheta/C}$ hold.

On the other hand, using the construction around (11.5), we may find some $\mathcal{V}_\varepsilon \subset B_n(1), \mathcal{V}_{\varepsilon,\delta} \subset B_n(1)\setminus B_n(\delta)$ (for the latter, we take $U_0 \sim \mathrm{Unif}(\delta, 1)$ therein) with $\min\{\mathrm{vol}(\mathcal{V}_\varepsilon)/\mathrm{vol}(B_n(1)), \mathrm{vol}(\mathcal{V}_{\varepsilon,\delta})/\mathrm{vol}(B_n(1)\setminus B_n(\delta))\} \geq 1 - C\varepsilon^{-1}e^{-n\varepsilon^2/C}$, such that for $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$,

$$\sup_{\mu_0\in\{\mathcal{V}_\varepsilon,\mathcal{V}_{\varepsilon,\delta}\}}\sup_{\eta\in\Xi_L}|\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)| \leq \varepsilon. \tag{11.13}$$

Now let

$$\mathcal{W}_{\varepsilon,\vartheta} \equiv \mathcal{U}_\vartheta \cap \mathcal{V}_\varepsilon, \quad \mathcal{W}_{\varepsilon,\delta,\vartheta} \equiv \mathcal{U}_{\delta,\vartheta} \cap \mathcal{V}_{\varepsilon,\delta}. \tag{11.14}$$

Then we have the volume estimates $\min\{\mathrm{vol}(\mathcal{W}_{\varepsilon,\vartheta})/\mathrm{vol}(B_n(1)), \mathrm{vol}(\mathcal{W}_{\varepsilon,\delta,\vartheta})/\mathrm{vol}(B_n(1)\setminus B_n(\delta))\} \geq 1 - C\varepsilon^{-1}e^{-n\varepsilon^2/C} - Ce^{-n^\vartheta/C}$.

Moreover, by Proposition 11.3, provided $\eta_* \equiv \sigma_\xi^2/\|\mu_0\|^2 = \mathsf{SNR}^{-1}_{\mu_0} \in \Xi_L$,

$$\|\mu_0\|^2/C_0 \leq \frac{|\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta_*)|}{(\eta-\eta_*)^2} \leq C_0\|\mu_0\|^2 \tag{11.15}$$

holds uniformly in $\eta \in \Xi_L$ for some $C_0 > 0$.

(**Noisy case** $\sigma_\xi^2 \in [1/K, K]$). Fix $\mu_0 \in \mathcal{W}_{\varepsilon,\delta,\vartheta}$. Under the assumed conditions, $\eta_* \in \Xi_L$. So using the estimates (11.13) and (11.15), for any $\eta' \geq \eta_*$,

$$\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta') - \inf_{\eta\in\Xi_L}\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta) \geq \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta') - \inf_{\eta\in\Xi_L}\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - 2\varepsilon \geq \frac{\delta^2(\eta'-\eta_*)^2}{C_0} - 2\varepsilon.$$

Combined with a similar inequality for $\eta' \leq \eta_*$, we conclude that for any $\mu_0 \in \mathcal{W}_{\varepsilon,\delta,\vartheta}$ and $\eta' \in \Xi_L$,

$$\Big|\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta') - \inf_{\eta\in\Xi_L}\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)\Big| \geq \frac{\delta^2(\eta'-\eta_*)^2}{C_0} - 2\varepsilon.$$

Now for $|\eta'-\eta_*| \geq \Delta$, choosing $\varepsilon \equiv \varepsilon_0 \equiv \delta^2\Delta^2/(4C_0)$, we have

$$\inf_{\mu_0\in\mathcal{W}_{\varepsilon,\delta,\vartheta}}\inf_{\eta'\in\Xi_L:|\eta'-\eta_*|\geq\Delta}\Big|\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta') - \inf_{\eta\in\Xi_L}\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)\Big| \geq \frac{\delta^2\Delta^2}{2C_0}.$$

From here the claim follows from Theorem 3.1.
(**Noiseless case** $\sigma_\xi^2 = 0$). In this case, (11.15) implies that the map $\eta \mapsto \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ attains global minimum at $\eta = 0$. So together with (11.13), it implies that uniformly in $\mu_0 \in \mathcal{W}_{\varepsilon,\vartheta}$,

$$\Big|\min_{\eta\in[0,K]} \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \bar{R}^{\#}_{(\Sigma,\mu_0)}(0)\Big| \leq \varepsilon.$$

From here the claim follows from Lemma 11.4 that holds for $\sigma_\xi = 0$. □

12. Proofs for Section 4

12.1. Proof of Theorem 4.1.

All the constants in $\lesssim, \gtrsim, \asymp$ may depend on K.

Proof of Theorem 4.1 for $\widehat{\tau}_\eta$. Let κ_0 be defined in the same way as in the proof of Proposition 10.3. Using a similar local law and continuity argument as in the proof of that proposition, on an event E_0 with $\mathbb{P}(E_0) \geq 1 - Cn^{-D}$,

$$\sup_{\eta\in\Xi_K} \Big|m^{-1}\operatorname{tr}\big(\check{\Sigma} + (\eta/\phi)I\big)^{-1} - \mathfrak{m}(-\eta/\phi)\Big| \lesssim \kappa_0^{-1} n^{-1/2+\varepsilon}.$$

So on $E_0 \cap \mathscr{E}(C_1)$, where $\mathscr{E}(C_1) \equiv \{\|Z\|_{\mathrm{op}}/\sqrt{n} \leq C_1\}$ with $\mathbb{P}(\mathscr{E}(C_1)) \geq 1 - Ce^{-n/C}$, uniformly in $\eta \in \Xi_K$,

$$|\widehat{\tau}_\eta - \tau_{\eta,*}| \leq \frac{\big|\frac{1}{m}\operatorname{tr}\big(\check{\Sigma} + \frac{\eta}{\phi}I\big)^{-1} - \mathfrak{m}\big(-\frac{\eta}{\phi}\big)\big|}{\frac{1}{m}\operatorname{tr}\big(\check{\Sigma} + \frac{\eta}{\phi}I\big)^{-1}\cdot \mathfrak{m}\big(-\frac{\eta}{\phi}\big)} \lesssim \Big\{C_1^2 \mathbf{1}^{-1}_{\phi^{-1}\geq 1+1/K} \wedge \eta^{-1}\Big\}\cdot \kappa_0^{-1} n^{-1/2+\varepsilon}.$$

Here in the last inequality, we use the following estimate for $\mathfrak{m}(z)$: As $\mathfrak{m}$ is the Stieltjes transform of ρ (cf. [KY17, Lemma 2.2]), $\mathfrak{m}(z) \geq 0$ for $z \leq 0$, and

$$\frac{1}{\mathfrak{m}(z)} = (-z) + \frac{1}{m}\operatorname{tr}\Big(\big(I + \Sigma\mathfrak{m}(z)\big)^{-1}\Sigma\Big) \lesssim 1 + |z|.$$

The claim follows. □

Proof of Theorem 4.1 for $\widehat{\gamma}_\eta$. Using Theorem 3.1, the stability of $\tau_{\eta,*}$ in Proposition 8.1, and the proven fact in (1) on $\widehat{\tau}_\eta$, it holds for $\varepsilon \in (0, 1/2]$ that

$$\mathbb{P}\Big(\sup_{\eta\in[1,K,K]} \Big|\eta^{-1}\widehat{\tau}_\eta\|\widehat{r}_\eta(\sigma_\xi)\| - \gamma_{\eta,*}(\sigma_\xi)\Big| \geq \varepsilon\Big) \leq C_1\varepsilon^{-c_0} n^{-6.5}. \tag{12.1}$$

Next we consider extension to $\eta \in [0, K]$ in the regime $\phi^{-1} \geq 1 + 1/K$. By KKT condition, we have $n^{-1}X^\top(Y - X\widehat{\mu}_\eta) = \eta\widehat{\mu}_\eta$, so a.s. $\widehat{r}_\eta/\eta = (Y - X\mu_\eta)/(\sqrt{n}\eta) = \sqrt{n}(XX^\top)^{-1}X\widehat{\mu}_\eta$ for any $\eta > 0$. So we only need to verify the high probability Lipschitz continuity for $\eta \mapsto \sqrt{n}\widehat{\tau}_\eta(XX^\top)^{-1}X\widehat{\mu}_\eta$: for any $\eta_1, \eta_2 \in [0, K]$, using the estimate (9.29) (with G replaced by Z) we obtain, for some universal $c_0 > 1$,

$$\begin{aligned}
&\Big|\sqrt{n}\widehat{\tau}_{\eta_1}\big\|(XX^\top)^{-1}X\widehat{\mu}_{\eta_1}\big\| - \sqrt{n}\widehat{\tau}_{\eta_2}\big\|(XX^\top)^{-1}X\widehat{\mu}_{\eta_2}\big\|\Big|\\
&\lesssim \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}\cdot(\|Z\|_{\mathrm{op}}/\sqrt{n})\cdot\Big(|\widehat{\tau}_{\eta_1} - \widehat{\tau}_{\eta_2}|\cdot\|\widehat{\mu}_{\eta_1}\| + |\widehat{\tau}_{\eta_2}|\cdot\|\widehat{\mu}_{\eta_1} - \widehat{\mu}_{\eta_2}\|\Big)\\
&\lesssim \Big(1 + \frac{\|Z\|_{\mathrm{op}} + \|\xi_0\|}{\sqrt{n}} + \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}\Big)^{c_0}\cdot|\eta_2 - \eta_1|.
\end{aligned}$$

Finally we consider extension to $\sigma_\xi^2 \in [0,K]$ in the same regime $\phi^{-1} \geq 1 + 1/K$ by verifying a similar high probability uniform-in-η Lipschitz continuity property for $\sigma_\xi \mapsto \sqrt{n}(XX^\top)^{-1}X\widehat{\mu}_\eta(\sigma_\xi)$: for any $\sigma_\xi, \sigma_\xi' \in [0,K]$, using the estimate (11.11),

$$\begin{aligned}
&\sup_{\eta\in[0,K]} \Big| \sqrt{n}\widehat{\tau}_\eta \big\|(XX^\top)^{-1}X\widehat{\mu}_\eta(\sigma_\xi)\big\| - \sqrt{n}\widehat{\tau}_\eta \big\|(XX^\top)^{-1}X\widehat{\mu}_\eta(\sigma_\xi')\big\|\Big| \\
&\lesssim \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}^2 \cdot (\|Z\|_{\mathrm{op}}/\sqrt{n}) \cdot \sup_{\eta\in[0,K]} \|\widehat{\mu}_\eta(\sigma_\xi) - \widehat{\mu}_\eta(\sigma_\xi')\| \\
&\lesssim \Big(1 + \frac{\|Z\|_{\mathrm{op}} + \|\xi_0\|}{\sqrt{n}} + \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}\Big)^{c_0} \cdot |\sigma_\xi - \sigma_\xi'|.
\end{aligned}$$

The claimed bound follows. □

12.2. **Proof of Theorem 4.2.** Recall we have $\gamma_{\eta,*}^2 = \phi^{-1}(\sigma_\xi^2 + \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta))$. For both the case $\sigma_\xi^2 \in [1/K, K]$ and $\sigma_\xi^2 \in [0,K]$ with $\phi^{-1} \geq 1 + 1/K$, we take $\mathcal{W}_{\varepsilon,\delta,\vartheta} \subset B_n(1) \setminus B_n(\delta)$ as constructed in (11.14) in the proof of Theorem 3.4, with $\varepsilon \equiv \varepsilon_n \equiv n^{-\vartheta}$. Fix $\mu_0 \in \mathcal{W}_{\varepsilon,\delta,\vartheta}$, then $\eta_* = \mathsf{SNR}_{\mu_0}^{-1} \in \Xi_L$. Using Theorems 3.2 and 4.1, on an event E_0 with $\mathbb{P}(E_0^c) \leq Cn^{-1/7}$,

$$\sup_{\eta\in\Xi_L} \Big|\widehat{\gamma}_\eta^2 - \phi^{-1}(\sigma_\xi^2 + \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta))\Big| \leq \varepsilon. \tag{12.2}$$

This in particular implies that on E_0, both the following inequalities hold:

$$\begin{aligned}
\phi\widehat{\gamma}^2_{\widehat{\eta}^{\mathsf{GCV}}} - \sigma_\xi^2 - \phi\varepsilon \leq \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) \leq \phi\widehat{\gamma}^2_{\widehat{\eta}^{\mathsf{GCV}}} - \sigma_\xi^2 + \phi\varepsilon,\\
\phi \min_{\eta\in\Xi_L}\widehat{\gamma}_\eta^2 - \sigma_\xi^2 - \phi\varepsilon \leq \min_{\eta\in\Xi_L}\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) \leq \phi \min_{\eta\in\Xi_L}\widehat{\gamma}_\eta^2 - \sigma_\xi^2 + \phi\varepsilon.
\end{aligned} \tag{12.3}$$

Using the definition of $\widehat{\eta}^{\mathsf{GCV}}$ which gives $\widehat{\gamma}^2_{\widehat{\eta}^{\mathsf{GCV}}} = \min_{\eta\in\Xi_L}\widehat{\gamma}_\eta^2$, the above two displays can be used to relate $\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}})$ and $\min_{\eta\in\Xi_L}\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$: on the event E_0,

$$\Big|\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \min_{\eta\in\Xi_L}\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\Big| \leq 2\phi\varepsilon. \tag{12.4}$$

As $\eta_* \in \Xi_L$, $\min_{\eta\in\Xi_L}\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) = \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta_*)$ for $\# \in \{\mathsf{pred},\mathsf{est},\mathsf{in}\}$. Consequently, by the second inequality in Proposition 11.3, we have on the event E_0,

$$|\widehat{\eta}^{\mathsf{GCV}} - \eta_*| \leq \frac{C}{\|\mu_0\|}\Big|\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta_*)\Big|^{1/2} \leq C_1\varepsilon^{1/2}. \tag{12.5}$$

This means on E_0, for both $\# \in \{\mathsf{est},\mathsf{in}\}$,

$$\Big|\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \min_{\eta\in\Xi_L}\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)\Big| \leq C_2\varepsilon.$$

We may conclude from here by virtues of Theorems 3.1 and 3.2, together with Lemma 11.4. □

12.3. **Proof of Theorem 4.3.**

Lemma 12.1. *Consider the following version of (2.1) with sample size $m - m_\ell$:*

$$\begin{cases} \frac{m-m_\ell}{n} \cdot \gamma^2 = \sigma_\xi^2 + \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau), \\ (\frac{m-m_\ell}{n} - \frac{\eta}{\tau}) \cdot \gamma^2 = \mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau). \end{cases} \tag{12.6}$$

(1) *The fixed point equation (12.6) admits a unique solution $(\gamma_{\eta,*}^{(\ell)}, \tau_{\eta,*}^{(\ell)}) \in (0,\infty)^2$, for all $(m,n) \in \mathbb{N}^2$ when $\eta > 0$ and $m < n$ when $\eta = 0$.*
(2) *Further suppose $1/K \le \phi^{-1}, \sigma_\xi^2 \le K$, $m_\ell/n \le 1/(2K)$ and $\|\Sigma^{-1}\|_{\mathrm{op}} \vee \|\Sigma\|_{\mathrm{op}} \le K$ for some $K > 10$. Then there exists some $C = C(K) > 1$ such that uniformly in $\eta \in \Xi_K$, $1/C \le \gamma_{\eta,*}^{(\ell)}, \tau_{\eta,*}^{(\ell)} \le C$. Moreover,*

$$|\gamma_{\eta,*}^{(\ell)} - \gamma_{\eta,*}| \vee |\tau_{\eta,*}^{(\ell)} - \tau_{\eta,*}| \le \frac{Cm_\ell}{n}.$$

Proof. All the constants in $\lesssim, \gtrsim, \asymp$ below may depend on K. We only need to prove (2). The method of proof is similar to that of Proposition 8.1-(3). Instead of considering (12.6), we shall consider the system of equations

$$\begin{cases} \phi - \alpha = \frac{1}{\gamma^2}(\sigma_\xi^2 + \tau^2 \|(\Sigma + \tau I)^{-1}\Sigma^{1/2}\mu_0\|^2) + \frac{1}{n}\operatorname{tr}((\Sigma + \tau I)^{-2}\Sigma^2), \\ \phi - \alpha = \frac{1}{n}\operatorname{tr}((\Sigma + \tau I)^{-1}\Sigma) + \frac{\eta}{\tau}, \end{cases} \tag{12.7}$$

indexed by $\alpha \ge 0$. For $\alpha \in [0, 1/(2K)]$, the solution $(\gamma_{\eta,*}(\alpha), \tau_{\eta,*}(\alpha))$ exists uniquely for $\eta > 0$ and also for $\eta = 0$ if additionally $m < n$. Moreover, using the apriori estimate in Proposition 8.1-(2), we have uniformly in $\eta \in \Xi_K$ and $\alpha \in [0, 1/(2K)]$, $\gamma_{\eta,*}(\alpha), \tau_{\eta,*}(\alpha) \asymp 1$. Now differentiating on both sides of the second equation in (12.7) with respect to α, we obtain

$$1 = \Big(n^{-1}\operatorname{tr}((\Sigma + \tau_{\eta,*}(\alpha)I)^{-2}\Sigma) + \eta\tau_{\eta,*}^{-2}(\alpha)\Big) \cdot \tau_{\eta,*}'(\alpha).$$

This means uniformly in $\eta \in \Xi_K$ and $\alpha \in [0, 1/(2K)]$, $\tau_{\eta,*}'(\alpha) \asymp 1$. Next, using the first equation in (12.7), we obtain

$$\gamma_{\eta,*}^2(\alpha) = \frac{\sigma_\xi^2 + \tau_{\eta,*}^2(\alpha)\|(\Sigma + \tau_{\eta,*}(\alpha)I)^{-1}\Sigma^{1/2}\mu_0\|^2}{\phi - \alpha - \frac{1}{n}\operatorname{tr}((\Sigma + \tau_{\eta,*}(\alpha)I)^{-2}\Sigma^2)} \equiv \frac{G_{1,\eta}(\alpha)}{G_{2,\eta}(\alpha)}.$$

Using similar calculations as in (8.9)-(8.10), we have uniformly in $\eta \in \Xi_K$ and $\alpha \in [0, 1/(2K)]$, $G_{1,\eta}(\alpha), G_{2,\eta}(\alpha) \asymp 1$, and $|G_{1,\eta}'(\alpha)| \vee |G_{2,\eta}'(\alpha)| \lesssim 1$. This concludes the claim. □

Proof of Theorem 4.3. All the constants in $\lesssim, \gtrsim, \asymp$ below may depend on K, L.

As $\|Y^{(\ell)} - X^{(\ell)}\widehat{\mu}_\eta^{(\ell)}\|^2 = \|Z^{(\ell)}\Sigma^{1/2}(\mu_0 - \widehat{\mu}_\eta^{(\ell)}) + \xi^{(\ell)}\|^2$ and $\widehat{\mu}_\eta^{(\ell)}$ is independent of $(Z^{(\ell)}, \xi^{(\ell)})$, by using Lemma B.3 first conditionally on $(Z^{(-\ell)}, \xi^{(-\ell)})$ and then further taking expectation over $(Z^{(-\ell)}, \xi^{(-\ell)})$, we have for $0 < \varrho \le 1$,

$$\begin{aligned}\mathbb{P}\Big(E_{0,\ell}^c(\eta) \equiv \Big\{\Big|m_\ell^{-1}\|Y^{(\ell)} - X^{(\ell)}\widehat{\mu}_\eta^{(\ell)}\|^2 - (\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)} - \mu_0)\|^2 + \sigma_\xi^2)\Big| \\ \ge C_0(\sigma_\xi^2 \vee \|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)} - \mu_0)\|^2) m_\ell^{-(1-\varrho)/2}\Big\}\Big) \le C_0 e^{-m_\ell^\varrho/C_0}.\end{aligned}$$

Here $C_0 > 0$ is a universal constant. Using similar arguments as in (11.3) (by noting that the normalization in $\widehat{\mu}_\eta^{(\ell)}$ is still n), there exists some constant $C_1 > 0$ such that for any $\ell \in [k]$, on an event $E_{1,\ell}$ with $\mathbb{P}(E_{1,\ell}^c) \leq C_1 e^{-m_\ell/C_1}$, $\sup_{\eta\in\Xi_L}\|\widehat{\mu}_\eta^{(\ell)}\| \leq C_1$. This means that for any $\eta \in \Xi_L$, on the event $\cap_{\ell\in[k]}(E_{0,\ell}(\eta) \cap E_{1,\ell})$,

$$\max_{\ell\in[k]} m_\ell^{(1-\varrho)/2} \cdot \Big|m_\ell^{-1}\|Y^{(\ell)} - X^{(\ell)}\widehat{\mu}_\eta^{(\ell)}\|^2 - (\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)} - \mu_0)\|^2 + \sigma_\xi^2)\Big| \leq C_1'. \tag{12.8}$$

On the other hand, using Theorem 3.1, we may find some $\mathcal{U}_{\vartheta;\ell} \subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_{\vartheta;\ell})/\mathrm{vol}(B_n(1)) \geq 1 - C_2 e^{-n^\vartheta/C_2}$, such that for $\varepsilon \in (0, 1/2]$, on an event $E_{2,\ell}(\varepsilon)$ with $\mathbb{P}(E_{2,\ell}^c(\varepsilon)) \leq C_2(ne^{-n\varepsilon^4/C_2} + \varepsilon^{-c_0}n^{-1/6.5}\mathbf{1}_{Z\neq G})$, for $\mu_0 \in \mathcal{U}_{\vartheta;\ell}$,

$$\sup_{\eta\in\Xi_L}\Big|\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)} - \mu_0)\|^2 - \Big\{\frac{m - m_\ell}{n}(\gamma_{\eta,*}^{(\ell)})^2 - \sigma_\xi^2\Big\}\Big| \leq \varepsilon.$$

Here $\gamma_{\eta,*}^{(\ell)}$ is taken from Lemma 12.1, and we extend the definition to $\ell = 0$ with $\widehat{\mu}_\eta^{(0)} \equiv \widehat{\mu}_\eta$ and $\gamma_{\eta,*}^{(0)} \equiv \gamma_{\eta,*}$. Using the statement (2) of the same Lemma 12.1, on the event $E_{2,\ell}(\varepsilon)$, we then have

$$\sup_{\eta\in\Xi_L}\Big|\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)} - \mu_0)\|^2 - \{\phi\gamma_{\eta,*}^2 - \sigma_\xi^2\}\Big| \leq \varepsilon + \frac{C_2 m_\ell}{n}.$$

Replacing $\phi\gamma_{\eta,*}^2 - \sigma_\xi^2$ by $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) = \|\Sigma^{1/2}(\widehat{\mu}_\eta - \mu_0)\|^2$ yields that, on $\cap_{\ell\in[0:k]}E_{2,\ell}(\varepsilon)$,

$$\sup_{\eta\in\Xi_L}\Big|\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)} - \mu_0)\|^2 - R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\Big| \leq 2\varepsilon + \frac{C_2 m_\ell}{n}. \tag{12.9}$$

Combining (12.8)-(12.9), for $\mu_0 \in \mathcal{U}_\vartheta \equiv \cap_{\ell\in[0:k]}\mathcal{U}_{\vartheta;\ell}$, $\varepsilon \in (0, 1/2]$ and $\eta \in \Xi_L$,

$$\begin{aligned}
&\mathbb{P}\bigg(\Big|R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta) - (R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) + \sigma_\xi^2)\Big| \geq C_2' \cdot \Big\{\frac{1}{k}\sum_{\ell\in[k]}\frac{1}{m_\ell^{(1-\varrho)/2}} + \frac{1}{k} + \varepsilon\Big\}\bigg) \\
&\leq C_2' \cdot \begin{cases} \sum_{\ell\in[k]} e^{-m_\ell^\varrho/C_0} + kne^{-n\varepsilon^4/C_2}, & Z = G; \\ \sum_{\ell\in[k]} e^{-m_\ell^\varrho/C_0} + \varepsilon^{-c_0} \cdot kn^{-1/6.5}, & \text{otherwise}. \end{cases}
\end{aligned} \tag{12.10}$$

Now we strengthen the estimate (12.10) into a uniform version. It is easy to verify that on an event $E_{3,\ell}$ with $\mathbb{P}(E_{3,\ell}^c) \leq C_3 e^{-m_\ell/C_3}$, $\|Z^{(\ell)}\|_{\mathrm{op}} \leq C_3(\sqrt{m_\ell} + \sqrt{n})$, $\|\xi^{(\ell)}\| \leq C_3\sqrt{m_\ell}$, and for $\eta_1, \eta_2 \in \Xi_L$, $\|\widehat{\mu}_{\eta_1}^{(\ell)} - \widehat{\mu}_{\eta_2}^{(\ell)}\| \leq C_3|\eta_1 - \eta_2|$. So on $\cap_{\ell\in[k]}(E_{1,\ell} \cap E_{3,\ell})$, for $\eta_1, \eta_2 \in \Xi_L$,

$$\begin{aligned}
\Big|R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta_1) - R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta_2)\Big| &\lesssim \frac{1}{k}\sum_{\ell\in[k]}\frac{1}{m_\ell}\Big|\|Z^{(\ell)}\|_{\mathrm{op}}\|\widehat{\mu}_{\eta_1}^{(\ell)} - \widehat{\mu}_{\eta_2}^{(\ell)}\| \cdot (\|Z^{(\ell)}\|_{\mathrm{op}} + \|\xi^{(\ell)}\|)\Big| \\
&\lesssim \frac{1}{k}\sum_{\ell\in[k]}\frac{m_\ell + n}{m_\ell} \cdot |\eta_1 - \eta_2| \leq C_3' \cdot \frac{1}{k}\sum_{\ell\in[k]}\frac{n}{m_\ell} \cdot |\eta_1 - \eta_2|,
\end{aligned}$$

and

$$\Big|R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta_1) - R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta_2)\Big| \leq C_3'|\eta_1 - \eta_2|.$$

From here, using (i) (12.10) along with a discretization and union bound that strengthens (12.10) to a uniform control, and (ii) Theorem 3.1 which replaces

$R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$ by $\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$, we obtain for $\mu_0 \in \mathcal{U}_\vartheta$ and $\varepsilon \in (0, 1/2]$,

$$\mathbb{P}\bigg(\sup_{\eta\in\Xi_L}\big|R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta) - (\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) + \sigma_\xi^2)\big| \geq C_3'' \cdot \bigg\{\frac{1}{k}\sum_{\ell\in[k]}\frac{1}{m_\ell^{(1-\varrho)/2}} + \frac{1}{k} + \varepsilon\bigg\} \equiv \varepsilon_{\{m_\ell\}}\bigg)$$

$$\leq \mathfrak{p}_0 \equiv \frac{C_3''}{\varepsilon k}\sum_{\ell\in[k]}\frac{n}{m_\ell}\cdot\begin{cases}\sum_{\ell\in[k]} e^{-m_\ell^{\varrho}/C_0} + kne^{-n\varepsilon^4/C_2}, & Z = G;\\ \sum_{\ell\in[k]} e^{-m_\ell^{\varrho}/C_0} + \varepsilon^{-c_0}\cdot kn^{-1/6.5}, & \text{otherwise.}\end{cases}$$

Now with the same $\mathcal{W}_{\varepsilon,\delta,\vartheta}$ as in (11.14) using $\varepsilon \equiv \varepsilon_n \equiv n^{-\vartheta}$, for any $\mu_0 \in \mathcal{U}_\vartheta \cap \mathcal{W}_{\varepsilon,\delta,\vartheta}$, we may further replace $\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$ by $\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$ in the above display (with a possibly slightly larger $\varepsilon_{\{m_\ell\}}$, but for notational simplicity we abuse this notation). In summary, for any $\mu_0 \in \mathcal{U}_\vartheta \cap \mathcal{W}_{\varepsilon,\delta,\vartheta}$, on an event E_4 with $\mathbb{P}(E_4^c) \leq \mathfrak{p}_0$,

$$\sup_{\eta\in\Xi_L}\big|R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta) - (\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) + \sigma_\xi^2)\big| \leq \varepsilon_{\{m_\ell\}}.$$

From here, using similar arguments as in (12.3)-(12.4), on the event E_4,

$$\big|\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{CV}}) - \min_{\eta\in\Xi_L}\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big| \leq 2\varepsilon_{\{m_\ell\}}.$$

Similar to (12.5), on the event E_4, we have

$$|\widehat{\eta}^{\mathsf{CV}} - \eta_*| \leq C_4 \cdot \varepsilon_{\{m_\ell\}}^{1/2}. \tag{12.11}$$

From here we may argue along the same lines as those following (12.5) in the proof of Theorem 4.2 to conclude with probability estimated at $\mathfrak{p}_0$, by further noting that $\mathcal{U}_\vartheta \cap \mathcal{W}_{\varepsilon,\delta,\vartheta}$ satisfies the desired volume estimate. Under the further condition $\min_{\ell\in[k]} m_\ell \geq \log^{2/\delta} m$, by taking $\varrho = \delta$, $\mathfrak{p}_0$ simplifies as indicated in the statement of the theorem for n large. □

12.4. **Proof of Theorem 4.4.** We only prove the case for $\# = \mathsf{GCV}$; the other case is similar. All constants in $\lesssim, \gtrsim, \asymp$ and $\mathcal{O}$ may possibly depend on K, L. Let $\mathcal{W}_{\varepsilon,\delta,\vartheta} \subset B_n(1) \setminus B_n(\delta)$ be as constructed in (11.14) with $\varepsilon \equiv \varepsilon_n \equiv n^{-\vartheta}$.

(1). We first prove the statement for the length of the CI. Note that $|\mathrm{CI}_j(\eta)| = 2\widehat{\gamma}_\eta(\Sigma^{-1})_{jj}^{1/2} z_{\alpha/2}/\sqrt{n}$. By Theorem 4.1-(2), on an event E_0 with the probability indicated therein,

$$\max_{j\in[n]}\sup_{\eta\in\Xi_L}\bigg||\mathrm{CI}_j(\eta)| - 2\gamma_{\eta,*}(\Sigma^{-1})_{jj}^{1/2}\frac{z_{\alpha/2}}{\sqrt{n}}\bigg| \leq \frac{2\|\Sigma^{-1}\|_{\mathrm{op}}^{1/2} z_{\alpha/2}}{\sqrt{n}}\sup_{\eta\in\Xi_L}|\widehat{\gamma}_\eta - \gamma_{\eta,*}| \lesssim \frac{z_{\alpha/2}}{\sqrt{n}}\cdot\varepsilon.$$

Consequently, on the event E_0, for any $\mu_0 \in \mathcal{W}_{\varepsilon,\delta,\vartheta}$,

$$\begin{aligned}
&\sqrt{n}z_{\alpha/2}^{-1}\cdot\max_{j\in[n]}\big||\mathrm{CI}_j(\widehat{\eta}^{\mathsf{GCV}})| - \min_{\eta\in\Xi_L}|\mathrm{CI}_j(\eta)|\big|\\
&\lesssim |\gamma_{\widehat{\eta}^{\mathsf{GCV}},*} - \min_{\eta\in\Xi_L}\gamma_{\eta,*}| + \varepsilon\\
&\lesssim |\gamma^2_{\widehat{\eta}^{\mathsf{GCV}},*} - \min_{\eta\in\Xi_L}\gamma^2_{\eta,*}| + \varepsilon \quad \text{(using Proposition 8.1-(3))}\\
&\asymp \big|\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \min_{\eta\in\Xi_L}\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big| + \varepsilon \quad \text{(using definition of } \gamma^2_{\eta,*}\text{)}\\
&\lesssim \big|\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \min_{\eta\in\Xi_L}\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big| + \varepsilon \quad \text{(using Theorem 3.2).}
\end{aligned}$$

As in the proof of Theorem 4.2, for $\sigma_\xi^2 \leq K$, $\eta_* = \mathsf{SNR}_{\mu_0}^{-1} \in \Xi_L$, so by using Proposition 11.3-(2), on the event E_0, for any $\mu_0 \in \mathcal{W}_{\varepsilon,\delta,\vartheta}$,

$$\begin{aligned}&\sqrt{n}z_{\alpha/2}^{-1} \cdot \max_{j\in[n]} \Big||\mathrm{CI}_j(\widehat{\eta}^{\mathsf{GCV}})| - \min_{\eta\in\Xi_L} |\mathrm{CI}_j(\eta)|\Big| \\ &\lesssim |\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta_*)| + \varepsilon \lesssim |\widehat{\eta}^{\mathsf{GCV}} - \eta_*|^2 + \varepsilon.\end{aligned}$$

The above reasoning also proves that on the same event E_0, for any $\mu_0 \in \mathcal{W}_{\varepsilon,\delta,\vartheta}$,

$$|\gamma_{\widehat{\eta}^{\mathsf{GCV}},*} - \gamma_{\eta_*,*}| \lesssim |\widehat{\eta}^{\mathsf{GCV}} - \eta_*|^2 + \varepsilon.$$

From here, in view of (12.5), by adjusting constants, on an event E_1 with $\mathbb{P}(E_1^c) \leq C_1 n^{-1/7}$, it holds that

$$\begin{aligned}\sqrt{n}z_{\alpha/2}^{-1} \cdot \max_{j\in[n]} &\Big||\mathrm{CI}_j(\widehat{\eta}^{\mathsf{GCV}})| - \min_{\eta\in\Xi_L} |\mathrm{CI}_j(\eta)|\Big| \\ &\vee |\gamma_{\widehat{\eta}^{\mathsf{GCV}},*} - \gamma_{\eta_*,*}| \vee |\widehat{\eta}^{\mathsf{GCV}} - \eta_*|^2 \leq \varepsilon. \end{aligned} \tag{12.12}$$

This proves the claim for the length of the CI.

(2). Next we prove the statement for the coverage. We note that a similar Lipschitz continuity argument as in the proof of Lemma 11.4 shows that for any 1-Lipschitz $\mathsf{g} : \mathbb{R}^n \to \mathbb{R}$, on an event $E_2(\mathsf{g})$ with $\mathbb{P}(E_2(\mathsf{g})^c) \leq Cn^{-1/7}$,

$$\sup_{\eta\in\Xi_L} \Big|\mathsf{g}(\widehat{\mu}_\eta^{\mathsf{dR}}) - \mathbb{E}\,\mathsf{g}(\mu_0 + \gamma_{\eta,*}\Sigma^{-1/2}g/\sqrt{n})\Big| \leq \varepsilon. \tag{12.13}$$

On the other hand, using the Lipschitz continuity of $\eta \mapsto \tau_{\eta,*}$ in Proposition 8.1-(3),

$$\begin{aligned}\Big|\mathsf{g}(\widehat{\mu}^{\mathsf{dR}}_{\widehat{\eta}^{\mathsf{GCV}}}) - \mathsf{g}(\widehat{\mu}^{\mathsf{dR}}_{\eta_*})\Big| &\leq \|\widehat{\mu}^{\mathsf{dR}}_{\widehat{\eta}^{\mathsf{GCV}}} - \widehat{\mu}^{\mathsf{dR}}_{\eta_*}\| \\ &\lesssim |\tau_{\widehat{\eta}^{\mathsf{GCV}},*} - \tau_{\eta_*,*}| \sup_{\eta\in\Xi_L} \|\widehat{\mu}_\eta\| + \|\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}}} - \widehat{\mu}_{\eta_*}\| \\ &\lesssim \|\widehat{\eta}^{\mathsf{GCV}} - \eta_*\| \sup_{\eta\in\Xi_L} \|\widehat{\mu}_\eta\| + \|\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}}} - \widehat{\mu}_{\eta_*}\|.\end{aligned}$$

So by enlarging C_1 if necessary, we may assume without loss of generality that on $E_1 \cap E_2(\mathsf{g})$, we have

$$\Big|\mathsf{g}(\widehat{\mu}^{\mathsf{dR}}_{\widehat{\eta}^{\mathsf{GCV}}}) - \mathsf{g}(\widehat{\mu}^{\mathsf{dR}}_{\eta_*})\Big| \leq C_1\varepsilon^{1/2}. \tag{12.14}$$

Now we shall make a good choice of g in (12.12). Let $\Delta \in (0,1)$ and $\mathsf{g}_{0,\Delta} : \mathbb{R} \to [0,1]$ be a function such that $\mathsf{g}_{0,\Delta} = 1$ on $[-1,1]$, $\mathsf{g}_{0,\Delta} = 0$ on $\mathbb{R} \setminus (-1-\Delta, 1+\Delta)$, and linearly interpolated in $(-1-\Delta,-1) \cup (1, 1+\Delta)$. Let

$$\mathsf{g}(u) \equiv \frac{\Delta}{n}\sum_{j=1}^n \mathsf{g}_{0,\Delta}\bigg(\frac{u_j - \mu_{0,j}}{(\gamma_{\eta_*,*} + \varepsilon)(\Sigma^{-1})_{jj}^{1/2} z_{\alpha/2}/\sqrt{n}}\bigg). \tag{12.15}$$

It is easy to verify the Lipschitz property of g: for any $u_1, u_2 \in \mathbb{R}^n$, $|\mathsf{g}(u_1) - \mathsf{g}(u_2)| \lesssim n^{-1/2}\Delta\|\mathsf{g}_{0,\Delta}\|_{\mathrm{Lip}} \sum_{j=1}^n |u_{1,j} - u_{2,j}| \lesssim \|u_1 - u_2\|$. Consequently, we may apply (12.13) with g defined in (12.15) to obtain that on the event $E_1 \cap E_2(\mathsf{g})$,

$$\mathscr{C}^{\mathsf{dR}}(\widehat{\eta}^{\mathsf{GCV}}) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}\Big(\widehat{\mu}^{\mathsf{dR}}_{\widehat{\eta}^{\mathsf{GCV}},j} \in \Big[\mu_{0,j} \pm \widehat{\gamma}_{\widehat{\eta}^{\mathsf{GCV}}}(\Sigma^{-1})_{jj}^{1/2}\frac{z_{\alpha/2}}{\sqrt{n}}\Big]\Big)$$

$$
\begin{aligned}
&\le \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\Big(\widehat{\mu}^{\mathsf{dR}}_{\widehat{\eta}^{\mathsf{GCV}},j} \in \Big[\mu_{0,j} \pm (\gamma_{\eta_*,*}+\varepsilon)(\Sigma^{-1})^{1/2}_{jj}\frac{z_{\alpha/2}}{\sqrt{n}}\Big]\Big)\\
&\le \Delta^{-1}\cdot \mathsf{g}(\widehat{\mu}^{\mathsf{dR}}_{\widehat{\eta}^{\mathsf{GCV}}}) \quad (\text{using } \mathbf{1}_{[-1,1]} \le \mathsf{g}_{0,\Delta})\\
&\le \Delta^{-1}\cdot \mathsf{g}(\widehat{\mu}^{\mathsf{dR}}_{\eta_*}) + \mathcal{O}(\varepsilon^{1/2}/\Delta) \quad (\text{by (12.14)})\\
&\le \Delta^{-1}\cdot \mathbb{E}\,\mathsf{g}(\mu_0 + \gamma_{\eta_*,*}\Sigma^{-1/2}g/\sqrt{n}) + \mathcal{O}(\varepsilon^{1/2}/\Delta). \qquad (12.16)
\end{aligned}
$$

Now using $\mathsf{g}_{0,\Delta} \le \mathbf{1}_{[-1-\Delta,1+\Delta]}$ and the anti-concentration of the standard normal random variable, we may compute

$$
\begin{aligned}
&\Delta^{-1}\cdot \mathbb{E}\,\mathsf{g}(\mu_0 + \gamma_{\eta_*,*}\Sigma^{-1/2}g/\sqrt{n}) = \mathbb{E}\,\mathsf{g}_{0,\Delta}\Big(\frac{\gamma_{\eta_*,*}}{\gamma_{\eta_*,*}+\varepsilon}\cdot\frac{g}{z_{\alpha/2}}\Big)\\
&\le \mathbb{P}\Big(\mathcal{N}(0,1) \in \Big[\pm z_{\alpha/2}\cdot(1+\varepsilon/\gamma_{\eta_*,*})\cdot(1+\Delta)\Big]\Big) \le 1-\alpha+\mathcal{O}(\varepsilon+\Delta). \qquad (12.17)
\end{aligned}
$$

Combining the above two displays (12.16)-(12.17), on the event $E_1 \cap E_2(\mathsf{g})$,

$$
\mathscr{C}^{\mathsf{dR}}(\widehat{\eta}^{\mathsf{GCV}}) \le 1-\alpha+\mathcal{O}(\varepsilon+\Delta+\varepsilon^{1/2}/\Delta).
$$

Finally choosing $\Delta = \varepsilon^{1/4}$ to conclude the upper control. The lower control can be proved similarly so we omit the details. □

Appendix A. Further results on $\mathsf{OPT}^{\#}_{(\Sigma,\mu_0)}$

For $\# \in \{\mathsf{pred},\mathsf{est},\mathsf{in}\}$, recall $\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ defined in Theorem 3.2, and the optimally tuned risks $\mathsf{OPT}^{\#}_{(\Sigma,\mu_0)}$ defined as $\mathsf{OPT}^{\#}_{(\Sigma,\mu_0)} = \min_{\eta\ge 0}\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)/\sigma_\xi^2 = \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta_*)/\sigma_\xi^2$ with $\eta_* = \mathsf{SNR}_{\mu_0}^{-1}$.

Proposition A.1. *We have*

$$
\mathsf{OPT}^{\mathsf{pred}}_{(\Sigma,\mu_0)} = \frac{\phi\tau_{\eta_*,*}}{\eta_*} - 1,\ \mathsf{OPT}^{\mathsf{est}}_{(\Sigma,\mu_0)} = \frac{1-\phi}{\eta_*} + \frac{1}{\tau_{\eta_*,*}},\ \mathsf{OPT}^{\mathsf{in}}_{(\Sigma,\mu_0)} = -\frac{\eta_*}{\tau_{\eta_*,*}} + \phi.
$$

Consequently,

(1) $\mathsf{OPT}^{\mathsf{est}}_{(\Sigma,\mu_0)} = \mathsf{SNR}_{\mu_0}\big(1-\phi+\frac{\phi}{\mathsf{OPT}^{\mathsf{pred}}_{(\Sigma,\mu_0)}+1}\big)$*, and* $\mathsf{OPT}^{\mathsf{in}}_{(\Sigma,\mu_0)} = \phi\cdot\frac{\mathsf{OPT}^{\mathsf{pred}}_{(\Sigma,\mu_0)}}{\mathsf{OPT}^{\mathsf{pred}}_{(\Sigma,\mu_0)}+1}$*;*

(2) $\partial_\phi\,\mathsf{OPT}^{\mathsf{pred}}_{(\Sigma,\mu_0)} \le 0$*,* $\partial_\phi\,\mathsf{OPT}^{\mathsf{est}}_{(\Sigma,\mu_0)} \le 0$*, and* $\partial_\phi\,\mathsf{OPT}^{\mathsf{in}}_{(\Sigma,\mu_0)} \ge 0$.

Proof. We only need to verify (2). This follows from the formula for $\mathsf{OPT}^{\#}_{(\Sigma,\mu_0)}$ and the following simple consequences of the second equation of (2.1):

- $\partial_\phi\tau_{\eta_*,*} \le 0$,
- $\partial_\phi(\phi\tau_{\eta_*,*}) = n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta_*,*}I)^{-2}\Sigma^2\big)\cdot\partial_\phi\tau_{\eta_*,*} \le 0$,
- $-\eta_*^{-1} + \partial_\phi\tau_{\eta_*,*}^{-1} = \eta_*^{-1}n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta_*,*}I)^{-2}\Sigma\big)\cdot\partial_\phi\tau_{\eta_*,*} \le 0$.

The proof is complete. □

The first claim in (1) above provides a non-asymptotic version of [DW18, Corollary 2.2], while the first claim in (2) gives a non-asymptotic analogue of the monotonicity result obtained in [PD23, Theorem 6]. Interestingly, (2) also asserts the monotonicity of $\mathsf{OPT}^{\mathsf{est}}_{(\Sigma,\mu_0)}, \mathsf{OPT}^{\mathsf{in}}_{(\Sigma,\mu_0)}$ with respect to ϕ^{-1}.

Appendix B. Auxiliary results

Proposition B.1. *Let $H : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be a non-negative, differentiable function. Suppose there exists some deterministic $\Gamma > 0$ such that $\|\nabla H(g)\|^2 \leq \Gamma^2 H(g)$ almost surely for $g \sim \mathcal{N}(0, I_n)$. Then there exists some universal constant $C > 0$ such that for all $t \geq 0$,*

$$\mathbb{P}\Big(|H(g) - \mathbb{E}\, H(g)|/C \geq \Gamma\, \mathbb{E}^{1/2}\, H(g) \cdot \sqrt{t} + \Gamma^2 \cdot t\Big) \leq Ce^{-t/C}.$$

Proof. The method of proof via the Gaussian log-Sobolev inequality and the Herbst's argument is well known. We give some details for the convenience of the reader. Let $Z \equiv H(g) - \mathbb{E}\, H(g)$ be the centered version of H, and $G(g) \equiv \lambda Z = \lambda(H(g) - \mathbb{E}\, H(g))$. Then $\|\nabla G(g)\|^2 = \lambda^2 \|\nabla H(g)\|^2 \leq \lambda^2\Gamma^2 \cdot H(g) = \lambda^2\Gamma^2 \cdot (Z + \mathbb{E}\, H(g))$. By the Gaussian log-Sobolev inequality (see e.g., [BLM13, Theorem 5.4], or [GN16, Theorem 2.5.6]),

$$\mathrm{Ent}(e^{\lambda Z}) = \mathbb{E}[\lambda Z e^{\lambda Z}] - \mathbb{E}\, e^{\lambda Z} \log \mathbb{E}\, e^{\lambda Z} \leq \frac{1}{2}\, \mathbb{E}\, [\lambda^2\Gamma^2(Z + \mathbb{E}\, H(g))e^{\lambda Z}].$$

With $m_Z(\lambda) \equiv \mathbb{E}\, e^{\lambda Z}$ denoting the moment generation function of Z, the above inequality is equivalent to

$$\lambda m_Z'(\lambda) - m_Z(\lambda) \log m_Z(\lambda) \leq \frac{\Gamma^2\lambda^2}{2}\Big(m_Z'(\lambda) + \mathbb{E}\, H(g) \cdot m_Z(\lambda)\Big).$$

Now dividing $\lambda^2 m_\lambda(Z)$ on both sides of the above display, we have $(\log m_Z(\lambda)/\lambda)' \leq \frac{\Gamma^2}{2}\big(\log m_Z(\lambda) + \lambda\, \mathbb{E}\, H(g)\big)'$. Integrating both sides with the condition $\lim_{\lambda \downarrow 0}(\log m_Z(\lambda)/\lambda) = 0$ and $\log m_Z(\lambda) = 0$, we arrive at $\log m_Z(\lambda) \leq \frac{\Gamma^2}{2}(\lambda \log m_Z(\lambda) + \lambda^2\, \mathbb{E}\, H(g))$. Solving for $\log m_Z(\lambda)$ and using the standard method to convert to tail bound yield the claimed inequality. □

Lemma B.2. *Let $\Sigma \in \mathbb{R}^{n\times n}$ be an invertible covariance matrix with $\|\Sigma\|_{\mathrm{op}} \vee \|\Sigma^{-1}\|_{\mathrm{op}} \leq K$ for some $K > 0$. Then for any $q \in (0, \infty)$, there exists some $C = C(K, q) > 0$ such that*

$$\bigg|\frac{\mathbb{E}\|\mathcal{N}(0, \Sigma)\|_q}{\|\mathrm{diag}(\Sigma)\|_{q/2}^{1/2} M_q} - 1\bigg| \leq Cn^{-\frac{1}{q\vee 2}}\sqrt{\log n}.$$

where $M_q \equiv \mathbb{E}^{1/q}|\mathcal{N}(0, 1)|^q = 2^{1/2}\{\Gamma((q + 1)/2)/\sqrt{\pi}\}^{1/q}$.

Proof. Let $g \sim \mathcal{N}(0, I_n)$. We first prove that for some $C_0 > 1$,

$$n^{\frac{1}{q\vee 2}}/C_0 \leq \mathbb{E}\|\Sigma^{1/2} g\|_q \leq C_0 n^{\frac{1}{q}}. \tag{B.1}$$

The upper bound in the above display is trivial. For the lower bound, using $\|x\| \leq n^{\frac{1}{2}-\frac{1}{q\vee 2}}\|x\|_q$, we find $\mathbb{E}\|\Sigma^{1/2} g\|_q \geq n^{-\frac{1}{2}+\frac{1}{q\vee 2}}\, \mathbb{E}\|\Sigma^{1/2} g\| \gtrsim n^{\frac{1}{q\vee 2}}$. This proves (B.1).

As $\|x\|_q \leq n^{-\frac{1}{2}+\frac{1}{q\wedge 2}}\|x\|$, the map $g \mapsto \|\Sigma^{1/2} g\|_q$ is $\|\Sigma\|_{\mathrm{op}}^{1/2} n^{-\frac{1}{2}+\frac{1}{q\wedge 2}}$-Lipschitz with respect to $\|\cdot\|$. So by Gaussian concentration, for any $t \geq 0$,

$$\mathbb{P}\Big(E(t)^c \equiv \Big\{n^{\frac{1}{2}-\frac{1}{q\wedge 2}}\Big|\|\Sigma^{1/2} g\|_q - \mathbb{E}\|\Sigma^{1/2} g\|_q\Big| \geq C\sqrt{t}\Big\}\Big) \leq Ce^{-t/C}.$$

Consequently, using the above concentration and (B.1),

$$\begin{aligned}\mathbb{E}\|\Sigma^{1/2}g\|_q^q &\leq \mathbb{E}\|\Sigma^{1/2}g\|_q^q \mathbf{1}_{E(t)} + \mathbb{E}^{1/2}\|\Sigma^{1/2}g\|_q^{2q} \cdot \mathbb{P}^{1/2}(E(t)^c)\\ &\leq (\mathbb{E}\|\Sigma^{1/2}g\|_q + C\sqrt{t})^q + C \cdot n^{1/q}\,\mathbb{P}^{1/2}(E(t)^c)\\ &\leq (\mathbb{E}\|\Sigma^{1/2}g\|_q)^q \cdot \{(1 + Cn^{-\frac{1}{q\sqrt{2}}}\sqrt{t})^q + C \cdot n^{\frac{1}{q}-\frac{1}{q\sqrt{2}}}\,\mathbb{P}^{1/2}(E(t)^c)\}.\end{aligned}$$

By choosing $t = C_1 \log n$ for some sufficiently large $C_1 > 0$, we have

$$\frac{\mathbb{E}\|\mathcal{N}(0,\Sigma)\|_q}{\|\mathrm{diag}(\Sigma)\|_{q/2}^{1/2} M_q} = \frac{\mathbb{E}\|\Sigma^{1/2}g\|_q}{\mathbb{E}^{1/q}\|\Sigma^{1/2}g\|_q^q} \geq (1 - Cn^{-\frac{1}{q\sqrt{2}}}\sqrt{\log n})_+.$$

The upper bound follows similarly. □

Lemma B.3. *Let $Z \in \mathbb{R}^{m\times n}$ be a random matrix with independent, mean-zero, unit variance, uniformly sub-gaussian components. Suppose the coordinates of ξ are i.i.d. mean zero and uniformly subgaussian with variance $\sigma_\xi^2 > 0$, and are independent of Z. Then there exists some universal constant $C > 0$ such that for any $b \in \mathbb{R}^n$ and $0 < \varrho \leq 1$, with probability at least $1 - Ce^{-m^\varrho/C}$,*

$$\left| m^{-1}\|Zb + \xi\|^2 - (\|b\|^2 + \sigma_\xi^2)\right| \leq C \cdot (\sigma_\xi^2 \vee \|b\|^2) \cdot m^{-(1-\varrho)/2}.$$

Proof. Let $Z_1, \ldots, Z_m \in \mathbb{R}^n$ be the rows of Z. Then

$$\frac{1}{m}\|Zb + \xi\|^2 = \|b\|^2 \frac{1}{m}\sum_{i=1}^m \left\langle Z_i, \frac{b}{\|b\|}\right\rangle^2 + \frac{2\sigma_\xi\|b\|}{m}\sum_{i=1}^n \frac{\xi_i}{\sigma_\xi}\left\langle Z_i, \frac{b}{\|b\|}\right\rangle + \sigma_\xi^2 \frac{\|\xi/\sigma_\xi\|^2}{m}.$$

Using standard concentration estimates, with probability at least $1 - Ce^{-m^\varrho/C}$,

- $\left|\|b\|^2 \frac{1}{m}\sum_{i=1}^m \langle Z_i, \frac{b}{\|b\|}\rangle^2 - \|b\|^2\right| \leq C\|b\|^2 \cdot m^{-(1-\varrho)/2}$,
- $\left|\frac{2\sigma_\xi\|b\|}{m}\sum_{i=1}^n \frac{\xi_i}{\sigma_\xi}\langle Z_i, \frac{b}{\|b\|}\rangle\right| \leq C\sigma_\xi\|b\| \cdot m^{-(1-\varrho)/2}$,
- $\left|\sigma_\xi^2 \frac{\|\xi/\sigma_\xi\|^2}{m} - \sigma_\xi^2\right| \leq C\sigma_\xi^2 \cdot m^{-(1-\varrho)/2}$.

Collecting the bounds to conclude. □

References

[ASS20] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky, *High-dimensional dynamics of generalization error in neural networks*, Neural Networks **132** (2020), 428–446.

[AZ20] Morgane Austern and Wenda Zhou, *Asymptotics of cross-validation*, arXiv preprint arXiv:2001.11111 (2020).

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song, *A convergence theory for deep learning via over-parameterization*, International Conference on Machine Learning, PMLR, 2019, pp. 242–252.

[Bel20] Pierre C. Bellec, *Out-of-sample error estimate for robust m-estimators with convex penalty*, arXiv preprint arXiv:2008.11840 (2020).

[BEM13] Mohsen Bayati, Murat A Erdogdu, and Andrea Montanari, *Estimating lasso risk and noise level*, Advances in Neural Information Processing Systems **26** (2013).

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, *Reconciling modern machine-learning practice and the classical bias-variance trade-off*, Proc. Natl. Acad. Sci. USA **116** (2019), no. 32, 15849–15854.

[BHX20] Mikhail Belkin, Daniel Hsu, and Ji Xu, *Two models of double descent for weak features*, SIAM J. Math. Data Sci. **2** (2020), no. 4, 1167–1180.

[BLLT20] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler, *Benign overfitting in linear regression*, Proc. Natl. Acad. Sci. USA **117** (2020), no. 48, 30063–30070.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford University Press, Oxford, 2013.

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

[BS10] Zhidong Bai and Jack W. Silverstein, *Spectral analysis of large dimensional random matrices*, second ed., Springer Series in Statistics, Springer, New York, 2010.

[BZ23] Pierre C. Bellec and Cun-Hui Zhang, *Debiasing convex regularized estimators and interval estimation in linear models*, Ann. Statist. **51** (2023), no. 2, 391–436.

[CDK22] Chen Cheng, John Duchi, and Rohith Kuditipudi, *Memorize to generalize: on the necessity of interpolation in high dimensional linear regression*, Conference on Learning Theory, PMLR, 2022, pp. 5528–5560.

[CM22] Chen Cheng and Andrea Montanari, *Dimension free ridge regression*, arXiv preprint arXiv:2210.08571 (2022).

[CMW22] Michael Celentano, Andrea Montanari, and Yuting Wei, *The lasso with general gaussian designs with applications to hypothesis testing*, arXiv preprint arXiv:2007.13716v2 (2022).

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach, *On lazy training in differentiable programming*, Advances in Neural Information Processing Systems **32** (2019).

[CW79] Peter Craven and Grace Wahba, *Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numer. Math. **31** (1978/79), no. 4, 377–403.

[Dic16] Lee H. Dicker, *Ridge regression and asymptotic minimax estimation over spheres of growing dimension*, Bernoulli **22** (2016), no. 1, 1–37.

[DKT22] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis, *A model of double descent for high-dimensional binary linear classification*, Inf. Inference **11** (2022), no. 2, 435–495.

[DSH23] Alexis Derumigny and Johannes Schmidt-Hieber, *On lower bounds for the bias-variance trade-off*, Ann. Statist., to appear. Available at arXiv:2006.00278 (2023+).

[DvdL05] Sandrine Dudoit and Mark J. van der Laan, *Asymptotics of cross-validated risk estimation in estimator selection and performance assessment*, Stat. Methodol. **2** (2005), no. 2, 131–154.

[DW18] Edgar Dobriban and Stefan Wager, *High-dimensional asymptotics of prediction: ridge regression and classification*, Ann. Statist. **46** (2018), no. 1, 247–279.

[DZPS18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh, *Gradient descent provably optimizes over-parameterized neural networks*, arXiv preprint arXiv:1810.02054 (2018).

[Efr04] Bradley Efron, *The estimation of prediction error: covariance penalties and cross-validation*, J. Amer. Statist. Assoc. **99** (2004), no. 467, 619–642, With comments and a rejoinder by the author.

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

[EK18] ———, *On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators*, Probab. Theory Related Fields **170** (2018), no. 1-2, 95–175.

[GHW79] Gene H. Golub, Michael Heath, and Grace Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics **21** (1979), no. 2, 215–223.

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer Series in Statistics, Springer-Verlag, New York, 2002.

[GN16] Evarist Giné and Richard Nickl, *Mathematical foundations of infinite-dimensional statistical models*, Cambridge Series in Statistical and Probabilistic Mathematics, [40], Cambridge University Press, New York, 2016.

[Gor85] Yehoram Gordon, *Some inequalities for Gaussian processes and applications*, Israel J. Math. **50** (1985), no. 4, 265–289.

[Gor88] ______, *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbf{R}^n$, Geometric aspects of functional analysis (1986/87), Lecture Notes in Math., vol. 1317, Springer, Berlin, 1988, pp. 84–106.

[Han22] Qiyang Han, *Noisy linear inverse problems under convex constraints: Exact risk asymptotics in high dimensions*, arXiv preprint arXiv:2201.08435 (2022).

[HK70] Arthur E. Hoerl and Robert W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), no. 1, 55–67.

[HKZ14] Daniel Hsu, Sham M. Kakade, and Tong Zhang, *Random design analysis of ridge regression*, Found. Comput. Math. **14** (2014), no. 3, 569–600.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, Ann. Statist. **50** (2022), no. 2, 949–986.

[HS22] Qiyang Han and Yandi Shen, *Universality of regularized regression estimators in high dimensions*, arXiv preprint arXiv:2206.07936 (2022).

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, Advances in Neural Information Processing Systems **31** (2018).

[JWHT21] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An introduction to statistical learning—with applications in R*, Springer Texts in Statistics, Springer, New York, [2021] ©2021, Second edition [of 3100153].

[KL22] Nicholas Kissel and Jing Lei, *On high-dimensional gaussian comparisons for cross-validation*, arXiv preprint arXiv:2211.04958 (2022).

[KLS20] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez, *The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization*, J. Mach. Learn. Res. **21** (2020), Paper No. 169, 16.

[KY17] Antti Knowles and Jun Yin, *Anisotropic local laws for random matrices*, Probab. Theory Related Fields **169** (2017), no. 1-2, 257–352.

[KZSS21] Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro, *Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting*, Advances in Neural Information Processing Systems **34** (2021), 20657–20668.

[LD19] Sifan Liu and Edgar Dobriban, *Ridge regression: Structure, cross-validation, and sketching*, arXiv preprint arXiv:1910.02373 (2019).

[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** (2021), 18137–18151.

[Li85] Ker-Chau Li, *From Stein's unbiased risk estimates to the method of generalized cross validation*, Ann. Statist. **13** (1985), no. 4, 1352–1377.

[Li86] ______, *Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing*, Ann. Statist. **14** (1986), no. 3, 1101–1112.

[Li87] ______, *Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set*, Ann. Statist. **15** (1987), no. 3, 958–975.

[LR20] Tengyuan Liang and Alexander Rakhlin, *Just interpolate: kernel "ridgeless" regression can generalize*, Ann. Statist. **48** (2020), no. 3, 1329–1347.

[LS22] Tengyuan Liang and Pragya Sur, *A precise high-dimensional asymptotic theory for boosting and minimum-$\ell 1$-norm interpolated classifiers*, Ann. Statist. **50** (2022), no. 3, 1669–1695.

[MM21] Léo Miolane and Andrea Montanari, *The distribution of the Lasso: uniform control over sparse balls and adaptive parameter tuning*, Ann. Statist. **49** (2021), no. 4, 2313–2335.

[MM22] Song Mei and Andrea Montanari, *The generalization error of random features regression: precise asymptotics and the double descent curve*, Comm. Pure Appl. Math. **75** (2022), no. 4, 667–766.

[MMM22] Song Mei, Theodor Misiakiewicz, and Andrea Montanari, *Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration*, Appl. Comput. Harmon. Anal. **59** (2022), 3–84.

[MRSY23] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan, *The generalization error of max-margin linear classifiers: Benign overfitting and high-dimensional asymptotics in the overparametrized regime*, arXiv preprint arXiv:1911.01544v3 (2023).

[MZ22] Andrea Montanari and Yiqiao Zhong, *The interpolation phase transition in neural networks: memorization and generalization under lazy training*, Ann. Statist. **50** (2022), no. 5, 2816–2847.

[PD23] Pratik Patil and Jin-Hong Du, *Generalized equivalences between subsampling and ridge regularization*, arXiv preprint arXiv:2305.18496 (2023).

[PWRT21] Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan Tibshirani, *Uniform consistency of cross-validation estimators for high-dimensional ridge regression*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3178–3186.

[RMR21] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco, *Asymptotics of ridge (less) regression under general source condition*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3889–3897.

[RV09] Mark Rudelson and Roman Vershynin, *Smallest singular value of a random rectangular matrix*, Comm. Pure Appl. Math. **62** (2009), no. 12, 1707–1739.

[SAH19] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi, *The impact of regularization on high-dimensional logistic regression*, Advances in Neural Information Processing Systems **32** (2019).

[Ste81] Charles M. Stein, *Estimation of the mean of a multivariate normal distribution*, Ann. Statist. **9** (1981), no. 6, 1135–1151.

[Sto74] M. Stone, *Cross-validatory choice and assessment of statistical predictions*, J. Roy. Statist. Soc. Ser. B **36** (1974), 111–147.

[Sto77] ———, *Asymptotics for and against cross-validation*, Biometrika **64** (1977), no. 1, 29–35.

[TAH18] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi, *Precise error analysis of regularized M-estimators in high dimensions*, IEEE Trans. Inform. Theory **64** (2018), no. 8, 5592–5628.

[TB22] Alexander Tsigler and Peter L. Bartlett, *Benign overfitting in ridge regression*, arXiv preprint arXiv:2009.14286v2 (2022).

[TOH15] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi, *Regularized linear regression: A precise analysis of the estimation error*, Conference on Learning Theory, PMLR, 2015, pp. 1683–1709.

[TV+04] Antonia M Tulino, Sergio Verdú, et al., *Random matrix theory and wireless communications*, Foundations and Trends® in Communications and Information Theory **1** (2004), no. 1, 1–182.

[vdVW96] Aad van der Vaart and Jon A. Wellner, *Weak Convergence and Empirical Processes*, Springer Series in Statistics, Springer-Verlag, New York, 1996.

[WWM22] Shuaiwen Wang, Haolei Weng, and Arian Maleki, *Does SLOPE outperform bridge regression?*, Inf. Inference **11** (2022), no. 1, 1–54.

[WX20] Denny Wu and Ji Xu, *On the optimal weighted ℓ_2 regularization in overparameterized linear regression*, Advances in Neural Information Processing Systems **33** (2020), 10112–10123.

[ZBH+21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, *Understanding deep learning (still) requires rethinking generalization*, Communications of the ACM **64** (2021), no. 3, 107–115.

[ZKS+22] Lijia Zhou, Frederic Koehler, Pragya Sur, Danica J Sutherland, and Nathan Srebro, *A non-asymptotic moreau envelope theory for high-dimensional generalized linear models*, arXiv preprint arXiv:2210.12082 (2022).

[ZZY22] Xianyang Zhang, Huijuan Zhou, and Hanxuan Ye, *A modern theory for high-dimensional Cox regression models*, arXiv preprint arXiv:2204.01161 (2022).

(Q. Han) Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA.
Email address: `qh85@stat.rutgers.edu`

(X. Xu) Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong.
Email address: `xxuay@connect.ust.hk`