Journal of Machine Learning Research 27 (2026) 1-94 Submitted 3/25; Revised 11/25; Published 1/26

The Distribution of Ridgeless Least Squares Interpolators

Qiyang Han QH85@STAT.RUTGERS.EDU
Department of Statistics,
Rutgers University,
Piscataway, NJ 08854, USA

Xiaocong Xu XUXIAOCO@MARSHALL.USC.EDU
Data Sciences and Operations Department
University of Southern California
Los Angeles, CA 90089, USA

Editor: Mahdi Soltanolkotabi

Abstract

The Ridgeless minimum ℓ_2-norm interpolator in overparametrized linear regression has attracted considerable attention in recent years in both machine learning and statistics communities. While it seems to defy conventional wisdom that overfitting leads to poor prediction, recent theoretical research on its ℓ_2-type risks reveals that its norm minimizing property induces an 'implicit regularization' that helps prediction in spite of interpolation.

This paper takes a further step that aims at understanding its precise stochastic behavior 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 provides a precise quantification of the prescribed implicit regularization in the most general distributional sense. Our distributional characterizations hold for general non-Gaussian random designs and extend uniformly to positively regularized Ridge estimators.

As a direct application, we obtain a complete characterization for a general class of weighted ℓ_q risks of the Ridge(less) estimators that are previously only known for $q = 2$ by random matrix methods. These weighted ℓ_q risks not only include the standard prediction and estimation errors, but also include the non-standard covariate shift settings. Our uniform characterizations further reveal a surprising feature of the commonly used generalized and k-fold cross-validation schemes: tuning the estimated ℓ_2 prediction risk by these methods alone lead to simultaneous optimal ℓ_2 in-sample, prediction and estimation risks, as well as the optimal length of debiased confidence intervals.

Keywords: comparison inequality, cross validation, minimum norm interpolator, random matrix theory, ridge regression, universality

1. Introduction

1.1 Overview

Consider the standard linear regression model

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

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), 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. Hoerl and Kennard (1970)) with regularization $\eta > 0$,

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

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

$$\widehat{\mu}_0 = \mathop{\arg\min}_{\mu \in \mathbb{R}^n} \left\{ \|\mu\|^2 : Y = X\mu \right\} = (X^\top X)^- X^\top Y, \tag{3}$$

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 (James et al., 2021; Derumigny and Schmidt-Hieber, 2023). 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 (Jacot et al., 2018; Du et al., 2019; Allen-Zhu et al., 2019; Belkin et al., 2019; Chizat et al., 2019; Zhang et al., 2021).

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 an exclusive focus on their prediction risks, cf. Tulino et al. (2004); El Karoui (2013); Hsu et al. (2014); Dicker (2016); Dobriban and Wager (2018); El Karoui (2018); Advani et al. (2020); Belkin et al. (2020); Muthukumar et al. (2020); Wu and Xu (2020); Bartlett et al. (2020, 2021); Chang et al. (2021); Koehler et al. (2021); Richards et al. (2021); Hastie et al. (2022); Tsigler and Bartlett (2023); Cheng and Montanari (2024); Zhou et al. (2024). The readers are referred to Tsigler and Bartlett (2023, Sections 1.2 & 9) for a thorough review on the relation between various ℓ_2 risk results for the Ridge(less) estimator. A unique insight from these works is the existence of 'implicit regularization' within the Ridgeless interpolator $\widehat{\mu}_0$, so that for certain scenarios of (Σ, μ_0), the prediction risk of $\widehat{\mu}_0$ could be small (i.e., benign overfitting) or even optimal (Kobak et al., 2020; Hastie et al., 2022; Tsigler and Bartlett, 2023).

Despite substantial progress in understanding the ℓ_2 risk behavior of the Ridgeless estimator $\widehat{\mu}_0$, our understanding of its stochastic behavior as a statistical estimator remains limited. This gap is particularly important if we aim to consider $\widehat{\mu}_0$ also as a 'good' estimator that can be applied in a broader context of statistical inference tasks, rather than merely viewing it as a theoretical proxy for modern interpolating learning algorithms.

The main goal of this paper is to advance our understanding of the precise stochastic behavior of the Ridge(less) estimator $\widehat{\mu}_\eta$. We achieve this by developing a high-dimensional *distributional* characterization in the so-called proportional regime where m and n is of the same order. This approach allows us to move beyond the exclusive focus in the existing literature on ℓ_2-type risks of $\widehat{\mu}_\eta$. As will be clear, the distributional characterization of the Ridge(less) estimator $\widehat{\mu}_\eta$ not only provides a precise quantitative understanding of the 'implicit regularization' phenomenon for the Ridgeless interpolator $\widehat{\mu}_0$ in the most general distributional sense, but also unveils major new insights on the utility of the widely used cross-validation schemes in machine learning/statistics practice.

1.2 Distribution of Ridge(less) estimators

Before formally describing our high dimensional distributional characterization, it is insightful to consider the low dimensional regime $\phi^{-1} \ll 1$ where the sample size m far exceeds the signal dimension n. In this regime, with $(\bar{\eta}, \bar{\sigma}_\xi^2) \equiv (\eta, \sigma_\xi^2)/\phi$, using the closed form of (2) and the fact that $m^{-1}X^\top X \approx \Sigma$, we may safely regard $\widehat{\mu}_\eta \approx \big(\Sigma + \bar{\eta} I_n\big)^{-1}\big(\Sigma\mu_0 + m^{-1}X^\top\xi\big)$. Using central limit theorem for $m^{-1}X^\top\xi \overset{d}{\approx} \bar{\sigma}_\xi \cdot n^{-1/2}\Sigma^{1/2}g$ where $g \sim \mathcal{N}(0, I_n)$, we have

$$\widehat{\mu}_\eta \overset{d}{\approx} (\Sigma + \bar{\eta} I_n)^{-1}\Sigma^{1/2}\big(\Sigma^{1/2}\mu_0 + n^{-1/2} \cdot \bar{\sigma}_\xi g\big), \quad \phi^{-1} \ll 1. \tag{4}$$

A principled way to understand the above formula (4) is to consider an 'effective regression problem' in the *Gaussian sequence model*. Suppose for a given pair of (Σ, μ_0) and a noise level $\gamma > 0$, we observe

$$y^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma) \equiv \Sigma^{1/2}\mu_0 + n^{-1/2} \cdot \gamma g, \quad g \sim \mathcal{N}(0, I_n). \tag{5}$$

The Ridge estimator $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)$ with regularization $\tau \geq 0$ in the Gaussian sequence model (5) is defined as

$$\begin{aligned}\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) &\equiv \mathop{\arg\min}_{\mu\in\mathbb{R}^n} \left\{\frac{1}{2}\|\Sigma^{1/2}\mu - y^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma)\|^2 + \frac{\tau}{2}\|\mu\|^2\right\} \\ &= (\Sigma + \tau I_n)^{-1}\Sigma^{1/2}\big(\Sigma^{1/2}\mu_0 + n^{-1/2} \cdot \gamma g\big).\end{aligned} \tag{6}$$

Here, the subscript (Σ, μ_0) emphasizes the dependence on the underlying Gaussian sequence model with covariance Σ and signal μ_0. Comparing (4) and (6), it is clear that we may interpret (4) as $\widehat{\mu}_\eta \overset{d}{\approx} \widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\bar{\sigma}_\xi;\bar{\eta})$. In the proportional regime $\phi^{-1} \asymp 1$, the aforementioned interpretation still applies, but a crucial modification will be needed: the pair of the (scaled) original noise and regularization $(\bar{\sigma}_\xi, \bar{\eta})$ must be replaced by a pair of '*effective noise and regularization*'

$$(\gamma_{\eta,*}, \tau_{\eta,*}) \equiv \text{unique solution of the fixed point equation (13)} \tag{7}$$

when $\eta > 0$ and when $\eta = 0$ in the overparametrized regime (cf. Proposition 2).

More precisely, in the overparametrized regime $\phi^{-1} > 1$, 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 in Theorems 3 and 4 that the distribution $\widehat{\mu}_\eta$ can be characterized via $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*}; \tau_{\eta,*})$ in the following sense: for any 1-Lipschitz function $\mathsf{g} : \mathbb{R}^n \to \mathbb{R}$ and any $K > 0$, with high probability,

$$\sup_{\eta \in [0,K]} \left| \mathsf{g}(\widehat{\mu}_\eta) - \mathbb{E}\, \mathsf{g}\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*}; \tau_{\eta,*})\big) \right| \approx 0. \tag{8}$$

A particularly important technical aspect of (8) is that the distributional approximation (8) holds uniformly down to the interpolation regime $\eta = 0$ for $\phi^{-1} > 1$. This uniform guarantee will prove essential in the results ahead.

Interestingly, the distributional characterization (8) offers a principled approach to understand the 'implicit regularization' phenomenon for the Ridgeless interpolator $\widehat{\mu}_0$, through the lens of its distributionally equivalent Ridge estimator $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{0,*}, \tau_{0,*})$ in the Gaussian sequence model (5). Specifically, the prescribed implicit regularization can be directly attributed to the quantity $\tau_{0,*} > 0$ that can be solved as the unique positive solution to the equation

$$\phi = \frac{1}{n} \operatorname{tr}\big((\Sigma + \tau_{0,*} I_n)^{-1} \Sigma\big). \tag{9}$$

While this interpretation has been suggested in the context of ℓ_2 risks (Hastie et al., 2022; Cheng and Montanari, 2024) via a-posterior calculations, our theory (8) provides a formal justification for this equivalent understanding of the implicit regularization phenomenon for $\widehat{\mu}_0$ via $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}$, in the most precise and general distributional sense. The readers are referred to Section 1.5 for a more detailed comparison on the relation of our characterization of the implicit regularization via $\tau_{0,*}$ and a different line of interpretation in Bartlett et al. (2020, 2021); Tsigler and Bartlett (2023).

1.3 General ℓ_q-type risk formulae

As mentioned above, most prior works on the risk properties of Ridge(less) estimators $\widehat{\mu}_\eta$ have focused exclusively on ℓ_2-type risks, leveraging random matrix theory (RMT) (Tulino et al., 2004; El Karoui, 2013; Dicker, 2016; Dobriban and Wager, 2018; El Karoui, 2018; Advani et al., 2020; Wu and Xu, 2020; Bartlett et al., 2021; Richards et al., 2021; Hastie et al., 2022; Cheng and Montanari, 2024). This RMT approach is viable due to a direct reduction of ℓ_2-type risks of $\widehat{\mu}_\eta$ to the spectrum of X. In contrast, the more general ℓ_q risks depend not only on the spectrum but also on the structure of X's singular vectors in a highly nontrivial manner; therefore, the feasibility of a similar RMT-based analysis is in question.

Our uniform distributional theory in (8) is strong enough to characterize all ℓ_q risks of the Ridge(less) estimator. Specifically, for any $q \in [1, \infty)$ and a p.s.d. matrix $\mathsf{A} \in \mathbb{R}^{n \times n}$, with high probability,

$$\frac{\|\mathsf{A}\widehat{\mu}_\eta - \mu_0\|_q}{n^{-1/2} \|\operatorname{diag}\big(\Gamma^{\mathsf{A}}_{\eta;(\Sigma,\|\mu_0\|)}\big)\|^{1/2}_{q/2} M_q} \approx 1, \tag{10}$$

where $M_q = \mathbb{E}^{1/q}|\mathcal{N}(0,1)|^q$ and $\Gamma^{\mathsf{A}}_{\eta;(\Sigma,\|\mu_0\|)} = \mathsf{A}(\Sigma+\tau_{\eta,*}I_n)^{-1}(\widetilde{\gamma}^2_{\eta,*}(\|\mu_0\|)\Sigma+\tau^2_{\eta,*}\|\mu_0\|^2 I_n)(\Sigma+\tau_{\eta,*}I_n)^{-1}\mathsf{A}$; see Theorem 6 for the precise definition of $\widetilde{\gamma}^2_{\eta,*}(\|\mu_0\|)$ and the formal statement of the above result (10).

Beyond providing a precise characterization of all ℓ_q risks, the uniform nature of (8) 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 (8), we show that for 'most' μ_0's, the global optimum of $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}(\eta)$ for all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$ will be achieved approximately *at the same point* $\eta_* = \mathsf{SNR}^{-1}_{\mu_0}$ with high probability[1] (cf. Theorem 54).

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 49). 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 8), 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 $\eta_* = \mathsf{SNR}^{-1}_{\mu_0}$ (cf. Proposition 9), at least for 'most' signal μ_0's.

1.4 Cross-validation: optimality beyond prediction

The discussion in Section 1.3 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 (22) and (24) for precise definitions and literature review of $\widehat{\eta}^{\mathsf{GCV}}, \widehat{\eta}^{\mathsf{CV}}$ in the context of Ridge regression.

By design, both methods $\widehat{\eta}^{\mathsf{GCV}}, \widehat{\eta}^{\mathsf{CV}}$ are intended to estimate the prediction risk, so it is natural to expect that they perform well for the task of prediction. Interestingly, the insight from Section 1.3 suggests a far broader utility of these adaptive tuning methods. Indeed, as all the empirical risk curves $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}(\eta)$ are approximately minimized at the same point $\eta_* = \mathsf{SNR}^{-1}_{\mu_0}$, it is reasonable to conjecture that $\widehat{\eta}^{\mathsf{GCV}}, \widehat{\eta}^{\mathsf{CV}}$ could also yield optimal

1. Here $\mathsf{SNR}_{\mu_0} = \|\mu_0\|^2/\sigma^2_\xi$ is the usual notion of signal-to-noise ratio; when $\mu_0 \neq 0$ and $\sigma^2_\xi = 0$, we shall interpret $\mathsf{SNR}^{-1}_{\mu_0} = 0$.

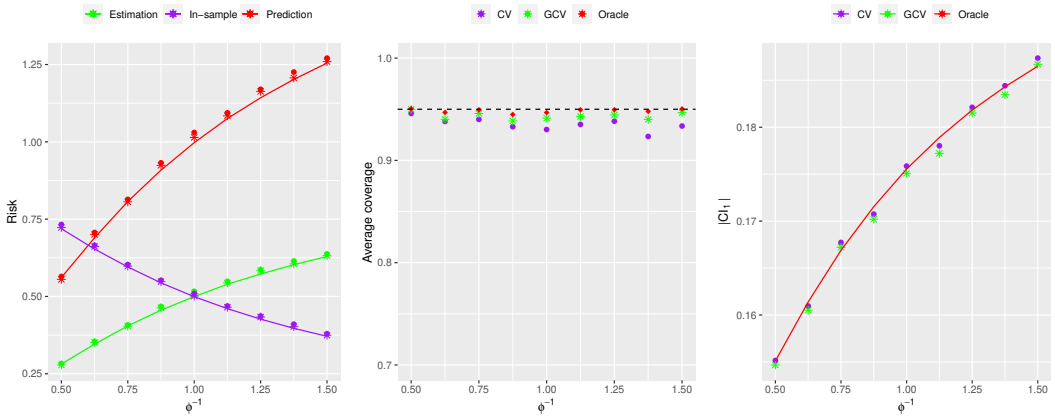

Figure 1: *Left panel*: Comparison between empirical risks and theoretical risks for $* = \mathsf{GCV}$ and $\bullet = \mathsf{CV}$ with $k = 5$. *Middle panel*: Averaged CI coverage $\mathscr{C}^{\mathsf{dR}}(\widehat{\eta}^{\#})$ for $\# \in \{\mathsf{GCV}, \mathsf{CV}\}$ and the oracle $\mathscr{C}^{\mathsf{dR}}(\eta_*)$. *Right panel*: CI length of $\mathrm{CI}_1(\widehat{\eta}^{\#})$ for $\# \in \{\mathsf{GCV}, \mathsf{CV}\}$ and the oracle CI length. See Section 4 for the precise definitions.

performance for estimation and in-sample risks. We show in Theorems 10 and 12 that this is indeed the case: for 'most' signal μ_0's and all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$, with high probability,

$$R^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}), R^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{CV}}) \approx \min_{\eta \in [0,K]} R^{\#}_{(\Sigma,\mu_0)}(\eta). \tag{11}$$

A typical simulation for this phenomenon is reported in the left panel of Figure 1 above, where empirical risks tuned by $\widehat{\eta}^{\mathsf{GCV}}$ (in $*$) and $\widehat{\eta}^{\mathsf{CV}}$ (in $\bullet$) achieve optimal theoretical risks (in solid lines) for estimation and in-sample risks as well.

Even more surprisingly, the optimality of $\widehat{\eta}^{\mathsf{GCV}}, \widehat{\eta}^{\mathsf{CV}}$ extends to the much more challenging task of statistical inference. In fact, we show in Theorem 13 that within 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. This is numerically validated in the middle and right panels of Figure 1.

To the best of our knowledge, theoretical optimality properties for the cross-validation schemes beyond the realm of prediction accuracy has not been established in the literature, either for Ridge regression or for other regularized regression estimators.

On the other hand, in the related Lasso setting, some numerical evidence for the broader utility of cross-validation and other adaptive tuning methods is reported in Miolane and Montanari (2021, Figure 1). There it is shown that the SURE method, which is designed to tune in-sample risk, nearly matches the performance of k-fold cross-validation in prediction tasks, despite not being expected to perform well in prediction apriori. Our findings here in the context of Ridge regression can therefore be viewed as a first step toward understanding the broader potential of cross validation and other adaptive tuning schemes for a wider range of statistical inference problems.

1.5 Further literature

1.5.1 Relation to mean-field asymptotics

Our distributional theory (8) for the Ridge(less) estimator is closely related to a recent line of research that examines the mean-field behavior of statistical estimators in the proportional regime $m \asymp n$, see, e.g. Bayati and Montanari (2012); Donoho and Montanari (2016); Thrampoulidis et al. (2018); Sur and Candès (2019); Li and Wei (2021); Miolane and Montanari (2021); Liang and Sur (2022); Celentano et al. (2023); Han and Shen (2023) for an incomplete list and many more references can be found therein.

A key feature of this line of works is the use of a simplified 'effective' regression problem to understand the complicated behavior of the original statistical estimator. For instance, in the closely related Lasso setting, the 'equivalence' between the Lasso estimator $\widehat{\mu}_\eta^{\mathsf{L}}$ in the linear model and a corresponding Lasso estimator in the sequence model $\widehat{\mu}^{\mathsf{seq},\mathsf{L}}_{\Sigma,\mu_0}(\gamma^{\mathsf{L}}_{\eta,*}, \tau^{\mathsf{L}}_{\eta,*})$ has been established under Gaussian designs with positive regularization. This equivalence was first shown for ℓ_2-type risks in Bayati and Montanari (2012), and later in the distributional sense akin to (8) in Miolane and Montanari (2021); Celentano et al. (2023). Such equivalence for Lasso is further extended to the interpolating regime in Li and Wei (2021) for the ℓ_2 risk under a standard Gaussian isotropic design. Our theory (8) here can thus be placed into a similar position as the progress made in Miolane and Montanari (2021); Celentano et al. (2023) over the Lasso risk characterization in Bayati and Montanari (2012), but now in the context of Ridge(less) estimator beyond a purely ℓ_2 risk as obtained in the references cited above.

While we have developed our distributional theory (8) primarily in the proportional regime $m \asymp n$, we conjecture that our theory (8) remains valid in the full nonparametric regime in which the ℓ_2 risk of the Ridge(less) estimator exceeds $\mathcal{O}(m^{-1/2})$. Some progress in this direction is made in Han (2023) in a related context of convex-constrained least squares estimator under a Gaussian design.

1.5.2 Relation to existing interpretation of 'implicit regularization'

A separate line of research (Bartlett et al., 2020, 2021; Tsigler and Bartlett, 2023) offers a different perspective on the implicit regularization phenomenon within the Ridgeless interpolator $\widehat{\mu}_0$. Specifically, by writing $X = [X_{\le k}, X_{>k}]$ with 'effective dimension' k and expressing the Ridgeless interpolator as $\widehat{\mu}_0 = X^\top (X_{\le k} X_{\le k}^\top + X_{>k} X_{>k}^\top)^{-1} Y$, this line of research identifies covariance structures Σ for which $X_{>k} X_{>k}^\top$ scales proportionally to the identity matrix (in a suitable sense). This implies that $X_{>k} X_{>k}^\top$ qualitatively plays the same role as if positive Ridge regularization were applied to the effective data $X_{\le k}$. In particular, this line of theory suggests that the prediction risk of $\widehat{\mu}_0$ can indeed vanish (i.e., benign overfitting), provided that the eigen-decay of Σ is neither too fast nor too slow.

While this approach is insightful, it falls short of providing an *exact* understanding for the emergence of the implicit regularization phenomenon. This is so, as this approach seeks sufficient conditions for $X_{>k} X_{>k}^\top \asymp I$, and produces risk bounds for $\widehat{\mu}_0$ modulo unspecified multiplicative constants. In contrast, our characterization of the implicit regularization via (9) is exact up to the leading constant order, and is susceptible to be also exact in other

regimes as well; see Cheng and Montanari (2024) for some recent partial progress along this line.

Furthermore, both the approaches of Bartlett et al. (2020, 2021); Tsigler and Bartlett (2023) and Cheng and Montanari (2024) rely heavily on the closed form of the Ridgeless interpolator and thus do not generalize to more general interpolators. In contrast, a significant technical advantage of our characterization of implicit regularization via (9) lies in its natural connection to the mean-field theory for general regression estimators. This suggests a general paradigm to quantify the implicit regularization for a large class of interpolators via mean-field asymptotics. For instance, the minimum ℓ_1-norm interpolator studied in Li and Wei (2021) demonstrates implicit regularization in the prediction risk that can be characterized via $\tau_{0,*}^{\mathsf{L}}$ as the 'Lasso' version of (9). Our approach developed here for Ridgeless interpolator is expected to be useful for quantifying the implicit regularization phenomenon for a more general class of interpolators.

1.6 Organization

The rest of the paper is organized as follows. In Section 2, we present our main results on the distributional characterizations (8) of the Ridge(less) estimator $\widehat{\mu}_\eta$. In Section 3, we provide a number of approximate ℓ_q risk formulae, and derive the optimal regularization for ℓ_2 risks. In Section 4, we give a formal validation for the two cross validation schemes mentioned above, both in terms of (11) and statistical inference via the debiased Ridge estimator. Due to the high technicalities involved in the proof of (8), a proof outline will be given in Section 5. All the proof details are then presented in Appendices A to G.

1.7 Notation

For any positive integer n, let $[n] = [1 : n]$ denote the set $\{1, \dots, n\}$. 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 $a = \mathcal{O}_x(b), a = \Omega_x(b)$ 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 any $\tau > 0$ are defined by $\mathsf{e}_f(x; \tau) \equiv \min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2\tau} \|x - z\|^2 + f(z) \right\}$ and $\mathsf{prox}_f(x; \tau) \equiv \arg\min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2\tau} \|x - z\|^2 + f(z) \right\}$.

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

$$\Xi_K \equiv [\mathbf{1}_{\phi^{-1}<1+1/K} K^{-1}, K]. \tag{12}$$

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 (6), 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

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

The $\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau)$ defined above is naturally related to the usual notion of degrees-of-freedom (cf. Stein (1981); Efron (2004)) for $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)$, in the sense that $\mathrm{df}\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)\big) \equiv \sum_{j=1}^n \frac{1}{\gamma^2/n} \mathrm{Cov}\left((\Sigma^{1/2}\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)})_j, y^{\mathsf{seq}}_{(\Sigma,\mu_0),j}\right) = \frac{n}{\gamma^2}\,\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau)$.

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. van der Vaart and Wellner (1996, 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.

Remark 1. *The requirement on the noise level σ_ξ^2 will be specified in concrete results below. We assume sub-gaussian noise for simplicity, but our proofs use it only through the high probability events in Section A.2. With easy modifications, all results extend to more general noise distributions, including certain heavy-tailed or weakly dependent cases for which these events still hold.*

2.3 The fixed point equation

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

$$\begin{cases} \phi\gamma^2 = \sigma_\xi^2 + \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau), \\ \phi - \frac{\eta}{\tau} = \frac{1}{n}\,\mathrm{tr}\left((\Sigma + \tau I_n)^{-1}\Sigma\right) = \frac{1}{\gamma^2}\,\mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau). \end{cases} \tag{13}$$

Fixed point equations of the type described above have appeared in the general mean-field theory for high dimensional regularized least squares estimators (LSEs), see, e.g. Bayati

and Montanari (2012); Bu et al. (2021); Li and Wei (2021); Han (2023); Celentano et al. (2023) for a sample of this type of equations in the i.i.d. sampling setting, and Bao et al. (2025) in the i.n.i.d. sampling setting. A common theme of these works characterizes the behavior of the regularized LSE in the linear model—at various levels of generality—via a regularized LSE in the equivalent sequence model, whose 'effective noise' and 'effective regularization' are determined by the solution pair to the fixed point equation.

In the context of Ridge regression, the form of the fixed point equation (13) appeared in Bartlett et al. (2021); Cheng and Montanari (2024) for the purpose of characterizing ℓ_2 risks for the Ridge(less) estimator $\widehat{\mu}_\eta$. It is now well understood that for the purpose of distributional characterizations of $\widehat{\mu}_\eta$, further stability properties for the solution pair to the fixed point equation will be needed (Miolane and Montanari, 2021; Celentano et al., 2023; Han, 2023). We establish these properties for the solution to (13) in the following proposition. Recall Ξ_K from (12).

Proposition 2. *Recall $\mathcal{H}_\Sigma = \mathrm{tr}(\Sigma^{-1})/n$. The following hold.*

1. *The fixed point equation (13) 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 $1/K \le \phi^{-1} \le K$ and $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ for some $K > 1$. Then there exists some $C = C(K) > 1$ such that uniformly in $\eta \in \Xi_K$,*

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

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

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

3. *Suppose $1/K \le \phi^{-1} \le K$ and $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ 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 $\mathrm{vol}(\mathcal{U}_\varepsilon)/\mathrm{vol}(B_n(1)) \ge 1 - C\varepsilon^{-1}e^{-n\varepsilon^2/C}$,*

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

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 23 and 52.

As an important qualitative consequence of (2), under the condition $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$, 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 (Bartlett et al., 2020, 2021; Hastie et al., 2022; Tsigler and Bartlett, 2023; Cheng and Montanari, 2024)—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 understanding some qualitative aspects of the risk curves in Section 3.2 ahead.

2.4 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

$$\widehat{r}_\eta \equiv \frac{1}{\sqrt{n}}\big(Y - X\widehat{\mu}_\eta\big). \tag{14}$$

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

$$r_{\eta,*} \equiv \frac{\eta}{\phi\tau_{\eta,*}}\bigg(- \sqrt{\phi\gamma_{\eta,*}^2 - \sigma_\xi^2}\cdot\frac{h}{\sqrt{n}} + \frac{\xi}{\sqrt{n}}\bigg). \tag{15}$$

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 $\mathcal{H}_\Sigma = \mathrm{tr}(\Sigma^{-1})/n$ and Ξ_K from (12).

Theorem 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 $\varepsilon \in (0, 1/2]$,*

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

   Here we recall that $\widehat{\mu}_{\eta;G}$ is defined in (2) with $X = G\Sigma^{1/2}$, and that $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}$ with effective noise and regularization pair $(\gamma_{\eta,*}, \tau_{\eta,*})$ is defined in (6)-(7).

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

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

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., (Bartlett et al., 2021, Assumption 4.12) or a slight variant in (Montanari et al., 2025, 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 4. *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 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* $\min\{\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)), \mathbb{P}(\xi \in \mathcal{E}_\vartheta)\} \ge 1 - Ce^{-n^{2\vartheta}/C}$, *such that the following hold.*

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

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

   Here we recall that $\widehat{\mu}_\eta$ *is defined in* (2), *and that* $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}$ *with effective noise and regularization pair* $(\gamma_{\eta,*}, \tau_{\eta,*})$ *is defined in* (6)-(7).

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*),*

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

Concrete forms of $\mathcal{U}_\vartheta, \mathcal{E}_\vartheta$ *are specified in Proposition 40.*

Remark 5. *Compared with the Gaussian case (Theorem 3), which admits exponential tails via Gaussian concentration and a direct CGMT argument, the sub-Gaussian universality in Theorem 4 yields only polynomial rates. This stems from the quantitative comparison inequalities Han and Shen (2023) employed in the universality step. Extending these bounds to exponential decay would likely require methods beyond the comparison framework, which is beyond the scope of the present paper.*

Theorems 3 and 4 are proved in Section C and Section D, respectively. Due to the high technicalities in the proof, a sketch is outlined in Section 5. These distributional results are the main input for all the applications developed in the subsequent sections. In particular, the flexibility in the choice of the test functions g and h allows us to obtain a variety of functionals of interest. By choosing the test functions appropriately, we derive (i) in Section 3 the ℓ_q-risk asymptotics for general $q \in [1, \infty)$, extending the classical ℓ_2-risk formulas that are typically accessible via random matrix theory, and (ii) in Section 4 the optimality of cross-validation tuning rules.

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 3 in hope of proving universality of $\widehat{r}_\eta$ in Theorem 4.

In the context of distributional characterizations for regularized regression estimators in the proportional regime, results in similar vein to Theorem 3 have been obtained in the closely related Lasso setting for isotropic $\Sigma = I_n$ in Miolane and Montanari (2021), and for general Σ in Celentano et al. (2023), both under Gaussian designs and with strictly non-vanishing regularization. A substantially simpler, isotropic ($\Sigma = I_n$) version of Theorem 4 is obtained in Han and Shen (2023) that holds pointwise in non-vanishing regularization level $\eta > 0$. As will be clear from the proof sketch in Section 5, in addition to the complications due to the implicit nature of the solution to the fixed point equation (13) for general Σ, the major difficulty in proving Theorems 3 and 4 rests in handling the singularity of the optimization problem (2) as $\eta \downarrow 0$.

3. General ℓ_q-type risk formulae

As a demonstration of the analytic power of Theorems 3 and 4, this section will be devoted to a detailed study for the ℓ_q-type risks for Ridgeless interpolators. We then conduct a more in-depth study of ℓ_2 risks, where techniques from RMT lead to a detailed characterization of the optimal regularization strategy for risk minimization.

3.1 Weighted ℓ_q risks and delocalization

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 $q \in [1, \infty)$. Recall Ξ_K from (12).

Theorem 6. *Suppose the same conditions in Theorem 4 hold for some $K > 0$. Fix $q \in [1, \infty)$ and a p.s.d. matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$ with $\|\mathsf{A}\|_{\mathrm{op}} \vee \|\mathsf{A}^{-1}\|_{\mathrm{op}} \le K$. 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 $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \ge 1 - Ce^{-n^\vartheta/C}$, such that*

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

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

$$\Gamma^{\mathsf{A}}_{\eta;(\Sigma,\|\mu_0\|)} \equiv \mathsf{A}(\Sigma + \tau_{\eta,*} I_n)^{-1}\Big(\widetilde{\gamma}^2_{\eta,*}(\|\mu_0\|)\Sigma + \tau^2_{\eta,*}\|\mu_0\|^2 I_n\Big)(\Sigma + \tau_{\eta,*} I_n)^{-1}\mathsf{A}, \tag{16}$$

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

The proof of the above theorem can be found in Section E. 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. Tulino et al. (2004); El Karoui (2013); Dicker (2016); Dobriban and Wager (2018); El Karoui (2018); Advani et al. (2020); Wu and Xu (2020); Bartlett et al. (2021); Richards et al. (2021); Hastie et al. (2022); Cheng and Montanari (2024).

For $\mathsf{A} \neq \Sigma$, Theorem 6 above characterizes the out-of-distribution ℓ_q risk for the Ridge(less) estimators. This setting is naturally related to the covariate shift setting, where

ℓ_2-type risks are studied in Patil et al. (2024); Tang et al. (2024) using random matrix methods in slightly different specific settings.

Let us remark that obtaining ℓ_q risks for $q \in [1,2]$ via our Theorems 3 and 4 is relatively easy, as $x \mapsto \|x\|_q/n^{1/q-1/2}$ is 1-Lipschitz with respect to $\|\cdot\|$ for $q \in [1,2]$. 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 7. *Suppose the same conditions as in Theorem 6 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 40, proved via the anisotropic local laws developed in Knowles and Yin (2017). In essence, delocalization allows us to apply Theorems 3 and 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 4; the readers are referred to Section 5 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 6 can be replaced by n^{-D} for any $D > 0$.
2. n^{-100} in Proposition 7 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.

3.2 ℓ_2 risk formulae and optimal regularization

In this subsection, we will study in some detail the behavior of various ℓ_2 risks associated with $\widehat{\mu}_\eta$. As will be clear below, a major analytic advantage of studying ℓ_2 risks is their close connection to techniques from RMT.

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

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

We also define the residual and its theoretical version as

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

From Theorems 3 and 4, it is natural to expect that for $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}, \mathsf{res}\}$,

$$\sup_{\eta \in \Xi^{\#}} |R^{\#}_{(\Sigma,\mu_0)}(\eta) - \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)| \approx 0 \text{ with high probability.} \tag{17}$$

A rigorous statement of (17) is deferred to Theorem 49; its proof and the proofs for all other results in this section can be found in Section F.

Using the so-called Stieltjes transformation $\mathfrak{m}(\cdot)$ in the RMT literature (defined formally via (41) in Section A.3 ahead), the following theorem provides an efficient RMT representation of $\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ that holds for 'most' μ_0's. Recall Ξ_K from (12).

Theorem 8. *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 $\mathrm{vol}(\mathcal{U}_\varepsilon)/\mathrm{vol}(B_n(1)) \ge 1 - C\varepsilon^{-1}e^{-n\varepsilon^2/C}$,*

$$\sup_{\mu_0 \in \mathcal{U}_\varepsilon} \sup_{\eta \in \Xi_K} |\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)| \le \varepsilon, \quad \# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}. \tag{18}$$

Here with $\mathsf{SNR}_{\mu_0} = \|\mu_0\|^2/\sigma_\xi^2$, $\mathfrak{m}_\eta \equiv \mathfrak{m}(-\eta/\phi)$ and $\mathfrak{m}'_\eta \equiv \mathfrak{m}'(-\eta/\phi)$,

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

When $\Sigma = I_n$, we may take $\mathcal{U}_\varepsilon = B_n(1)$ and (18) holds with $\varepsilon = 0$.

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

Proposition 9. *Suppose $1/K \le \phi^{-1} \le K$ and $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ for some $K > 0$. Then there exists some $C = C(K) > 0$ such that for all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$, the derivative formulae*

$$\partial_\eta \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) = \sigma_\xi^2 \cdot \mathfrak{M}^{\#}(\eta) \cdot \big(\eta \cdot \mathsf{SNR}_{\mu_0} - 1\big), \quad \eta \ge 0$$

hold for some measurable functions $\big\{\mathfrak{M}^{\#} : \mathbb{R}_{\ge 0} \to [1/C, C]\big\}$.

Consequently, for all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$, $\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\cdot)$ attains its global minimum at the same $\eta_* \equiv \mathsf{SNR}_{\mu_0}^{-1} \in \Xi_K$, and $1/C \le |\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta_*)| / \{\|\mu_0\|^2(\eta - \eta_*)^2\} \le C$.

A more general version of the above proposition with precise formulae for $\mathfrak{M}^{\#}$ can be found in Proposition 53. As the maps $\eta \mapsto \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ are almost quadratic with the same global minimizer $\eta_* = \mathsf{SNR}_{\mu_0}^{-1}$ for all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$, in view of (17) and Theorem 8, it is natural to expect that for 'most' signal μ_0's and all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$,

$$R^{\#}_{(\Sigma,\mu_0)}(\eta_*) \approx \min_{\eta \in \Xi_L} R^{\#}_{(\Sigma,\mu_0)}(\eta) \text{ with high probability.} \tag{19}$$

A rigorous formulation of (19) is given in Theorem 54, which, along with its proof, is provided in Section F.4.

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:

$$\begin{cases} \widehat{\tau}_\eta \equiv \Big\{ \frac{1}{m} \operatorname{tr}\big(\frac{1}{m} XX^\top + \frac{\eta}{\phi} I_m \big)^{-1} \Big\}^{-1} = \big\{ \operatorname{tr}(XX^\top + \eta \cdot n I_m)^{-1} \big\}^{-1}, \\ \widehat{\gamma}_\eta \equiv \frac{\widehat{\tau}_\eta}{\sqrt{n}} \Big(\eta^{-1} \|Y - X\widehat{\mu}_\eta\| \mathbf{1}_{\phi^{-1}<1} + \|(XX^\top/n)^{-1} X\widehat{\mu}_\eta\| \mathbf{1}_{\phi^{-1}\geq 1} \Big). \end{cases} \tag{20}$$

It can be easily shown that

$$\sup_{\eta \in \Xi_K} |\widehat{\tau}_\eta - \tau_{\eta,*}|, \ \sup_{\eta \in \Xi_K} |\widehat{\gamma}_\eta - \gamma_{\eta,*}| \approx 0 \text{ with high probability.} \tag{21}$$

A rigorous statement of (21) is deferred to Theorem 57; its proof and the proofs for all other results in this section can be found in Section G. 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.

4.2 Validation of generalized cross-validation

Consider choosing η by minimizing the estimated effective noise $\widehat{\gamma}_\eta$ given in (20): for any $L > 0$,

$$\widehat{\eta}_L^{\mathsf{GCV}} \in \operatorname*{arg\,min}_{\eta \in \Xi_L} \widehat{\gamma}_\eta. \tag{22}$$

Here we recall Ξ_K from (12). The subscript on L in $\widehat{\eta}_L^{\mathsf{GCV}}$ will usually be suppressed for notational simplicity.

The proposal (22) is known in the literature as the *generalized cross validation* (Craven and Wahba, 1978/79; Golub et al., 1979), and is strongly tied to the so-called shortcut formula for leave-one-out cross validation that exists uniquely for Ridge regression, cf. (Hastie et al., 2022, Eqn. (46)). Here we take a different perspective on (22). From our developed theory, this tuning scheme is easily believed to "work" since

$$\widehat{\gamma}_\eta^2 \overset{\mathbb{P}}{\approx} \gamma_{\eta,*}^2 = \phi^{-1}\big(\sigma_\xi^2 + \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big) \overset{\mathbb{P}}{\approx} \phi^{-1}\big(\sigma_\xi^2 + R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big). \tag{23}$$

So minimization of $\eta \mapsto \widehat{\gamma}_\eta$ is approximately the same as that of $\eta \mapsto R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)$, and therefore simultaneously of $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}(\eta)$ for $\# \in \{\mathsf{est}, \mathsf{in}\}$ as per (19). We make precise the foregoing heuristics in the following theorem.

Theorem 10. *Suppose Assumption A holds, and the following hold some $K > 0$.*

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

- *Assumption B holds with either (i)* $\sigma_\xi^2 \in [1/K, K]$ *or (ii)* $\sigma_\xi^2 \in [0, K]$ *with* $\phi^{-1} \geq 1 + 1/K$.

Fix $\delta \in (0, 1/2]$, $L \geq K/\delta^2$ *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 for* $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$,

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

Remark 11. *Formally, the set* $\mathcal{U}_{\delta,\vartheta}$ *is defined as* $\mathcal{U}_{\delta,\vartheta} \equiv \mathcal{U}_\vartheta \setminus B_n(\delta)$, *where* $\mathcal{U}_\vartheta$ *is defined in Proposition 40. The cutoff* $\|\mu_0\| \geq \delta$ *excludes vanishing signals and ensures that* $\eta_* = \mathsf{SNR}_{\mu_0}^{-1}$ *is uniformly bounded, so that* $\eta_* \in \Xi_L$ *whenever* $L \geq K/\delta^2$.

Earlier low-dimension results for generalized cross validation in Ridge regression include Stone (1974, 1977); Craven and Wahba (1978/79); Li (1985, 1986, 1987); Dudoit and van der Laan (2005). In the proportional high-dimensional regime, Hastie et al. (2022); Patil et al. (2021) validate the optimality of $\widehat{\eta}^{\mathsf{GCV}}$ with respect to the prediction risk $R^{\mathsf{pred}}_{(\Sigma,\mu_0)}$ with increasing generality. In Theorem 10 above, we prove that the optimality of $\widehat{\eta}^{\mathsf{GCV}}$ holds *simultaneously* for all the three indicated risks. To the best of our knowledge, such optimality of $\widehat{\eta}^{\mathsf{GCV}}$ beyond the prediction risk has not been previously observed in the literature.

4.3 Validation of k-fold cross-validation

Next we consider the widely used k-fold cross-validation. 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 $\widehat{\mu}_\eta^{(\ell)} \equiv \arg\min_{\mu \in \mathbb{R}^n} \left\{\frac{1}{2n}\|Y^{(-\ell)} - X^{(-\ell)}\mu\|^2 + \frac{\eta}{2}\|\mu\|^2\right\}$ 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 $L > 0$,

$$\widehat{\eta}_L^{\mathsf{CV}} \in \arg\min_{\eta \in \Xi_L} \left\{\frac{1}{k}\sum_{\ell \in [k]} \frac{1}{m_\ell}\|Y^{(\ell)} - X^{(\ell)}\widehat{\mu}_\eta^{(\ell)}\|^2\right\} \equiv \arg\min_{\eta \in \Xi_L} R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta). \tag{24}$$

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 (22), we may therefore expect that $\widehat{\eta}^{\mathsf{CV}}$ in (24) simultaneously provides optimal prediction, estimation and in-sample risks for 'most' signal μ_0's. This is the content of the following theorem.

Theorem 12. *Suppose the same conditions as in Theorem 10 and* $\max_{\ell\in[k]} m_\ell/n \le 1/(2K)$ *hold for some* $K > 0$. *Fix* $\delta \in (0, 1/2]$, $L \ge K/\delta^2$ *and a small enough* $\vartheta \in (0, 1/50)$. *Further assume* $\min_{\ell\in[k]} m_\ell \ge \log^{2/\delta} m$. *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* $\mathrm{vol}(\mathcal{U}_{\delta,\vartheta})/\mathrm{vol}(B_n(1) \setminus B_n(\delta)) \ge 1 - Ce^{-n^\vartheta/C}$, *such that for* $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$,

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

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 (Miolane and Montanari, 2021, 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. (James et al., 2021, 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 $\mathcal{O}(1/k)$ in the error bound, which is known to be of the optimal order in Ridge regression (cf. Liu and Dobriban (2019)).
- For a large number of k, $R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta)$ has large fluctuations; this corresponds to the term $\mathcal{O}\big(k^{-1}\sum_{\ell\in[k]} m_\ell^{-(1-\delta)/2}\big) = \mathcal{O}\big((k/m)^{(1-\delta)/2}\big)$ in the equal-sized case. By a central limit heuristic (cf. Kissel and Lei (2022); Austern and Zhou (2025)), we also expect this term to be of a near optimal order.

For the choice of k, it is instructive to consider the common equal-sized–folds case $m_\ell = m/k$. In this case, the error bound in Theorem 12 suggests that the optimal theoretical value of k is $k \sim m^{1/3}$ (when δ is small). In our numerical experiments, choosing $k = 5$ already yields cross-validation performance that is close to the theoretically optimal behavior (cf. Figure 1), whereas the theoretically prescribed optimal value of k offers limited practical gain.

4.4 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. Bellec and Zhang (2023). Here the debiasing scheme for $\widehat{\mu}_\eta$ can be

readily read off from the distributional characterizations in Theorems 3 and 4. Assuming known covariance Σ, let the debiased Ridge(less) estimator be defined as

$$\widehat{\mu}_\eta^{\mathsf{dR}} \equiv (\Sigma + \tau_{\eta,*} I)\Sigma^{-1}\widehat{\mu}_\eta. \tag{25}$$

Note that $\tau_{\eta,*}$ and $\widehat{\tau}_\eta$ is interchangeable in the above display due to known Σ. Using Theorems 3 and 4, we expect that $\widehat{\mu}_\eta^{\mathsf{dR}} \overset{d}{\approx} \mu_0 + \gamma_{\eta,*}\Sigma^{-1/2}g/\sqrt{n}$. This motivates the following confidence intervals for $\{\mu_{0,j}\}$:

$$\mathrm{CI}_j(\eta) \equiv \Big[\widehat{\mu}_{\eta,j}^{\mathsf{dR}} \pm \widehat{\gamma}_\eta \cdot (\Sigma^{-1})_{jj}^{1/2} \cdot \frac{z_{\alpha/2}}{\sqrt{n}}\Big], \quad j \in [n]. \tag{26}$$

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 (22), 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 13. *Suppose the same conditions as in Theorem 10 (resp. Theorem 12) 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*

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

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

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

A 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. As the effective noise $\phi\gamma_{\eta,*}^2(0) = \bar{R}_{(\Sigma,\mu_0)}^{\mathsf{pred}}(\eta)$, Theorem 8 and Proposition 9 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.

5. Proof outlines

5.1 Technical tools

The main technical tool we use for the proof of Theorem 3 is the following version of convex Gaussian min-max theorem, taken from (Miolane and Montanari, 2021, Corollary G.1).

Theorem 14 (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 $t \in \mathbb{R}$, $\mathbb{P}\left(\Phi^p(G) \le t\right) \le 2\,\mathbb{P}\left(\Phi^a(g,h) \le t\right)$.*
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 $t \in \mathbb{R}$, $\mathbb{P}\left(\Phi^p(G) \ge t\right) \le 2\,\mathbb{P}\left(\Phi^a(g,h) \ge t\right)$.*

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 Thrampoulidis et al. (2015, 2018); Salehi et al. (2019); Loureiro et al. (2021); Deng et al. (2022); Celentano et al. (2023); Han (2023); Liang and Sur (2022); Wang et al. (2022); Zhang et al. (2022); Montanari et al. (2025) for some selected references.

5.2 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

$$F(w) \equiv F_{(\Sigma,\mu_0)}(w) = \frac{1}{2}\|\mu_0 + \Sigma^{-1/2} w\|^2, \tag{27}$$

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

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

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

$$h_{\eta;Z}(w,v) \equiv \frac{1}{\sqrt{n}}\langle v, Zw-\xi\rangle + F(w) - \frac{\eta\|v\|^2}{2},$$
$$\ell_\eta(w,v) \equiv \frac{1}{\sqrt{n}}\Big(-\|v\|\langle g,w\rangle + \|w\|\langle h,v\rangle - \langle v,\xi\rangle\Big) + F(w) - \frac{\eta\|v\|^2}{2}, \tag{28}$$

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

$$H_{\eta;Z}(w;L_v) \equiv \max_{v\in B_n(L_v)} h_{\eta;Z}(w,v) \equiv \max_{v\in B_n(L_v)} \left\{ \frac{\langle v, Zw-\xi\rangle}{\sqrt{n}} + F(w) - \frac{\eta\|v\|^2}{2}\right\}, \tag{29}$$
$$L_\eta(w;L_v) \equiv \max_{v\in B_n(L_v)} \ell_\eta(w,v) = \max_{\beta\in[0,L_v]} \left\{ \frac{\beta}{\sqrt{n}}\Big(\big\|\|w\|h-\xi\big\| - \langle g,w\rangle\Big) + F(w) - \frac{\eta\beta^2}{2}\right\}.$$

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

$$\sigma_m^2 \equiv \frac{\|\xi\|^2}{\|h\|^2}, \quad \sigma_\pm^2(L_w) \equiv \left(\sigma_m^2 \pm 2L_w\frac{|\langle h,\xi\rangle|}{\|h\|^2}\right)_+. \tag{30}$$

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

$$\mathsf{D}_{\eta,\pm}(\beta,\gamma) \equiv \frac{\beta}{2}\left(\gamma\big(\phi e_h^2 - e_g^2\big) + \frac{\sigma_\pm^2}{\gamma}\right) - \frac{\eta\beta^2}{2} + \mathsf{e}_F\left(\frac{\gamma}{\sqrt{n}}g;\frac{\gamma}{\beta}\right),$$
$$\overline{\mathsf{D}}_\eta(\beta,\gamma) \equiv \frac{\beta}{2}\left(\gamma\big(\phi-1\big) + \frac{\sigma_\xi^2}{\gamma}\right) - \frac{\eta\beta^2}{2} + \mathbb{E}\,\mathsf{e}_F\left(\frac{\gamma}{\sqrt{n}}g;\frac{\gamma}{\beta}\right). \tag{31}$$

Here recall e_F is the Moreau envelope of F in (27). Note that $\mathsf{D}_{\eta,\pm}$ depends on the choice of L_w, but for notational convenience we drop this dependence here.

5.3 Proof outline for Theorem 3 for $\eta = 0$

We shall outline below the main steps for the proof of Theorem 3 for $\eta = 0$ in the regime $\phi^{-1} > 1$ under a stronger condition $\|\Sigma^{-1}\|_{\rm op} \lesssim 1$. The high level strategy of the proof shares conceptual similarities to Miolane and Montanari (2021); Celentano et al. (2023), 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.),

$$\min_{w\in B_n(L_w)} H_0(w;L_v) = \min_{w\in\mathbb{R}^n} H_0(w). \tag{32}$$

A formal statement of the above localization can be found in Proposition 27. 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 14.

(***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 (31). In particular, we show in Theorem 28 that for $L_w, L_v \asymp 1$ chosen large enough, w.h.p.,

$$\min_{w \in B_n(L_w)} L_0(w; L_v) \approx \max_{\beta>0} \min_{\gamma>0} \overline{\mathsf{D}}_0(\beta, \gamma). \tag{33}$$

The proof of (33) 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 29 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 30 that for localized minimax problems, we may replace $\mathsf{D}_{0,\pm}$ by $\overline{\mathsf{D}}_0$: 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 31 that (de)localization holds for the (deterministic) max-min optimization problem with $\overline{\mathsf{D}}_0$:

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

Combining the above Steps 2.1-2.3 yields (33). An important step to prove the (de)localization claims above is to derive apriori estimates for the solutions of the fixed point equation (13) and its sample version, to be defined in (55). These estimates will be detailed in Section B.

(***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 $w_{0,*} \equiv \Sigma^{1/2}\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{0,*}; \tau_{0,*}) - \mu_0\big)$ 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 $D_{0;\varepsilon}(\mathsf{g}) \equiv \big\{w \in \mathbb{R}^n : |\mathsf{g}(w) - \mathbb{E}\,\mathsf{g}(w_{0,*})| \ge \varepsilon\big\}$ be the 'exceptional set'. We show in Theorem 32 that again for $L_w, L_v \asymp 1$ chosen large enough, w.h.p.,

$$\min_{w \in D_{0;\varepsilon}(\mathsf{g}) \cap B_n(L_w)} L_0(w; L_v) \ge \max_{\beta>0} \min_{\gamma>0} \overline{\mathsf{D}}_0(\beta, \gamma) + \Omega_\varepsilon(1). \tag{34}$$

The main challenge in proving (34) 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 33, we will use surrogate, strongly convex functions $L_{0,\pm}(\cdot; L_v)$, formally defined in (79), to provide a sufficiently tight bracket for $L_0(\cdot; L_v)$ over large enough compact sets.

- (***Step 3.2***). In Proposition 34, 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 35, 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 (34) 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 (33) in Step 2 and (34) in Step 3 to their counterparts with primal cost function H_0. For the global cost optimum, with the help of the localization in (32), by choosing $L_w, L_v \asymp 1$, we have w.h.p.,

$$\min_{w\in\mathbb{R}^n} H_0(w) \overset{(32)}{=} \min_{w\in B_n(L_w)} H_0(w; L_v) \overset{\mathbb{P}}{\approx} \min_{w\in B_n(L_w)} L_0(w; L_v) \overset{(33)}{\approx} \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_0(\beta,\gamma).$$

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

$$\begin{aligned}\min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_n(L_w)} H_0(w) &\geq \min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_n(L_w)} H_0(w; L_v) \\ &\overset{\mathbb{P}}{\geq} \min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_n(L_w)} L_0(w; L_v) \quad \overset{(34)}{\geq} \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_0(\beta,\gamma) + \Omega_\varepsilon(1).\end{aligned}$$

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 C.

5.4 Proof outline for Theorem 4 for $\eta = 0$

The main tool we will use to prove the universality Theorem 4 is the following set of comparison inequalities developed in Han and Shen (2023): 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 $\|\cdot\|_\infty$),

$$\begin{aligned}\Big|\mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_w}\max_{v\in\mathcal{S}_v} h_{\eta;Z}(w,v)\Big) - \mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_w}\max_{v\in\mathcal{S}_v} h_{\eta;G}(w,v)\Big)\Big| &\leq \mathsf{r}_n(L_n), \\ \Big|\mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_w} H_{\eta;Z}(w)\Big) - \mathbb{E}\,\mathsf{T}\Big(\min_{w\in\mathcal{S}_w} H_{\eta;G}(w)\Big)\Big| &\leq \mathsf{r}_n(L_n). \qquad (35)\end{aligned}$$

Here $\mathsf{r}_n(L_n) \to 0$ for $L_n = n^{\vartheta}$ with sufficiently small $\vartheta > 0$. The readers are referred to Theorems 38 and 39 for a precise statement of (35).

An important technical subtlety here is that while the first inequality in (35) 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 (35) lead to universality.

(***Step 1**: Universality of the global cost optimum*). In this step, we shall use the first inequality in (35) to establish the universality of the global Gordon cost:

$$\min_{w\in\mathbb{R}^n} H_{0;Z}(w) = \min_{w\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} h_{0;Z}(w,v) \overset{\mathbb{P}}{\approx} \min_{w\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} h_{0;G}(w,v). \tag{36}$$

See Theorem 43 for a formal statement of (36).

The crux to establish (36) via the first inequality of (35) 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. (118)), are delocalized. We prove such delocalization properties in Proposition 40 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 (35) to establish the universality of the Gordon cost over exceptional sets $D_{0;\varepsilon}(\mathsf{g})$. In particular, we show in Theorem 44 that with $L_n = Cn^{\vartheta}$ for sufficiently small $\vartheta > 0$ and a large enough $C_0 > 0$, w.h.p.,

$$\min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0, L_n/\sqrt{n})} H_{0;Z}(w) \ge \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_0(\beta,\gamma) + \Omega_\varepsilon(1). \tag{37}$$

Here $B_{(2,\infty)}(C_0, L_n/\sqrt{n}) = B_n(C_0) \cap L_\infty(L_n/\sqrt{n})$. A technical difficulty to apply the second inequality of (35) 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 (35) cannot be applied. This singularity issue will be resolved in two steps:

- (***Step 2.1***). First, we use the second inequality of (35) to show that, (37) is valid for a version with small enough $\eta > 0$:

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

  See (130) 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 45) and $\max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma)$ (cf. Eqn. (76)) with respect to η, for a small enough $\eta > 0$, we have the following series of inequalities:

  $$\min_{w\in D_{0;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0, \frac{L_n}{\sqrt{n}})} H_{0;Z}(w)$$

$$
\begin{aligned}
&\geq \min_{w \in D_{0;\varepsilon}(\mathsf{g}) \cap B_{(2,\infty)}(C_0, \frac{L_\eta}{\sqrt{n}})} H_{\eta;Z}(w) \quad \text{(by definition of } H_{\eta;Z}) \\
&\geq \min_{w \in D_{\eta;\varepsilon_\eta}(\mathsf{g}) \cap B_{(2,\infty)}(C_0, \frac{L_\eta}{\sqrt{n}})} H_{\eta;Z}(w) \quad (\varepsilon_\eta \approx \varepsilon \text{ by Lemma 45}) \\
&\overset{\mathbb{P}}{\geq} \max_{\beta>0} \min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + \Omega_\varepsilon(1) \quad \text{(by Step 2.1 above)} \\
&\geq \max_{\beta>0} \min_{\gamma>0} \overline{\mathsf{D}}_0(\beta,\gamma) - \mathcal{O}(\eta) + \Omega_\varepsilon(1) \quad \text{(by Eqn. (76))}.
\end{aligned}
$$

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 (37).

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

Acknowledgments

The research of Q. Han is partially supported by NSF grant DMS-2143468. Both authors would like to thank the referees for their helpful comments and suggestions that significantly improved the quality of the paper.

References

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

Morgane Austern and Wenda Zhou. Asymptotics of cross-validation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 61(4), 2025.

Zhidong Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010.

Zhigang Bao, Qiyang Han, and Xiaocong Xu. A leave-one-out approach to approximate message passing. *Ann. Appl. Probab.*, 35(4):2716–2766, 2025.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA*, 117(48):30063–30070, 2020.

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

Mohsen Bayati and Andrea Montanari. The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory*, 58(4):1997–2017, 2012.

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(32):15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM J. Math. Data Sci.*, 2(4):1167–1180, 2020.

Pierre C. Bellec and Cun-Hui Zhang. Debiasing convex regularized estimators and interval estimation in linear models. *Ann. Statist.*, 51(2):391–436, 2023.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford, 2013.

Zhiqi Bu, Jason M. Klusowski, Cynthia Rush, and Weijie J. Su. Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Trans. Inform. Theory*, 67(1):506–537, 2021.

Michael Celentano, Andrea Montanari, and Yuting Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *Ann. Statist.*, 51(5):2194–2220, 2023.

Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6974–6983, 2021.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *Ann. Statist.*, 52(6): 2879–2912, 2024.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

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(4):377–403, 1978/79.

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

Alexis Derumigny and Johannes Schmidt-Hieber. On lower bounds for the bias-variance trade-off. *Ann. Statist.*, 51(4):1510–1533, 2023.

Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: ridge regression and classification. *Ann. Statist.*, 46(1):247–279, 2018.

David Donoho and Andrea Montanari. High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probab. Theory Related Fields*, 166(3-4):935–969, 2016.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Sandrine Dudoit and Mark J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.*, 2(2):131–154, 2005.

Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, 99(467):619–642, 2004. With comments and a rejoinder by the author.

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

Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields*, 170(1-2):95–175, 2018.

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.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Qiyang Han. Noisy linear inverse problems under convex constraints: Exact risk asymptotics in high dimensions. *Ann. Statist.*, 51(4):1611–1638, 2023.

Qiyang Han and Yandi Shen. Universality of regularized regression estimators in high dimensions. *Ann. Statist.*, 51(4):1799–1823, 2023.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.*, 50(2):949–986, 2022.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Found. Comput. Math.*, 14(3):569–600, 2014.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

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. Second edition [of 3100153].

Nicholas Kissel and Jing Lei. On high-dimensional gaussian comparisons for cross-validation. *arXiv preprint arXiv:2211.04958*, 2022.

Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probab. Theory Related Fields*, 169(1-2):257–352, 2017.

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:Paper No. 169, 16, 2020.

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*, 34:20657–20668, 2021.

Ker-Chau Li. From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.*, 13(4):1352–1377, 1985.

Ker-Chau Li. Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.*, 14(3):1101–1112, 1986.

Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987.

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

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and minimum-ℓ1-norm interpolated classifiers. *Ann. Statist.*, 50(3):1669–1695, 2022.

Sifan Liu and Edgar Dobriban. Ridge regression: Structure, cross-validation, and sketching. *arXiv preprint arXiv:1910.02373*, 2019.

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

Léo Miolane and Andrea Montanari. The distribution of the Lasso: uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.*, 49(4):2313–2335, 2021.

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. *Ann. Statist.*, 53(2):822–853, 2025.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3178–3186. PMLR, 2021.

Pratik Patil, Jin-Hong Du, and Ryan J Tibshirani. Optimal ridge regularization for out-of-distribution prediction. *arXiv preprint arXiv:2404.01233*, 2024.

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pages 3889–3897. PMLR, 2021.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.*, 62(12):1707–1739, 2009.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974.

M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35, 1977.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci.*, 116(29):14516–14525, 2019.

Shange Tang, Jiayun Wu, Jianqing Fan, and Chi Jin. Benign overfitting in out-of-distribution generalization of linear models. *arXiv preprint arXiv:2412.14474*, 2024.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR, 2015.

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

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:Paper No. [123], 76, 2023.

Antonia M Tulino, Sergio Verdú, et al. Random matrix theory and wireless communications. *Foundations and Trends® in Communications and Information Theory*, 1(1):1–182, 2004.

Aad van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

Shuaiwen Wang, Haolei Weng, and Arian Maleki. Does SLOPE outperform bridge regression? *Inf. Inference*, 11(1):1–54, 2022.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Xianyang Zhang, Huijuan Zhou, and Hanxuan Ye. A modern theory for high-dimensional Cox regression models. *arXiv preprint arXiv:2204.01161*, 2022.

Lijia Zhou, Frederic Koehler, Danica J Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *ACM/JMS Journal of Data Science*, 1(2):1–51, 2024.

Appendix A. Proof preliminaries

A.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 15. *For any $(\gamma,\tau)\in(0,\infty)^2$,*

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

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

$$\Sigma^{1/2}\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)-\mu_0\big) = (\Sigma+\tau I)^{-1}\Sigma^{1/2}\big(-\tau\mu_0+\gamma\Sigma^{1/2}g_n\big). \tag{38}$$

The claims follow from direct calculations. ∎

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

Lemma 16. *It holds that*

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

Furthermore,

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

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

The derivative formula below for e_F will be useful.

Lemma 17. *It holds that*

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

Proof See e.g., (Thrampoulidis et al., 2018, Lemmas B.5 and D.1). ∎

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

Proposition 18. *There exists some universal constant $C > 0$ such that*

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

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

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 $\nabla_g y^{\mathsf{seq}}(\gamma) = \frac{\gamma}{\sqrt{n}} I$,

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

This means

$$\begin{aligned}\|\nabla_g \mathsf{e}_F(\gamma g_n;\tau)\|^2 &\le 2\gamma^2\cdot n^{-1}\Big\{\|(\Sigma+\tau I)^{-1}\|_{\rm op}^2\|\Sigma^{1/2}\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) - y^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma)\|^2\\ &\qquad + \|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\|_{\rm op}^2\|\nabla_g\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau)\|^2\Big\}\\ &\le 4\gamma^2\cdot n^{-1}\Big(\tau\|(\Sigma+\tau I)^{-1}\|_{\rm op}^2 + \|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\|_{\rm op}^2\Big)\cdot \mathsf{e}_F(\gamma g_n;\tau). \quad (39)\end{aligned}$$

From here we may conclude by setting $H(g) \equiv \mathsf{e}_F\big(\gamma g_n;\tau\big)$ and $\Gamma^2 \equiv 4\gamma^2 n^{-1}\big(\tau\|(\Sigma+\tau I)^{-1}\|_{\rm op}^2 + \|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\|_{\rm op}^2\big)$ in Proposition 59. ∎

A.2 Some high probability events

Let

$$e_h^2 = \|h\|^2/m, \quad e_g^2 \equiv \|g\|^2/n. \quad (40)$$

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

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

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

Proof Using the definition of $\sigma_\pm^2(L_w)$ in (30), 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 20. *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., (Rudelson and Vershynin, 2009, Theorem 1.1). ■

Lemma 21. *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. ■

A.3 Some connections of the fixed point equation (13) to RMT

The second equation of (13) has a natural connection to 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 $\mathfrak{m}_n(z) \equiv m^{-1}\operatorname{tr}\big(\check{\Sigma} - zI_m\big)^{-1}$ and $\mathfrak{m}(z)$ be the Stieltjes transforms of the empirical spectral distribution and the asymptotic eigenvalue density (cf. (Knowles and Yin, 2017, Definition 2.3)) of $\check{\Sigma}$, respectively. It is well-known that $\mathfrak{m}(z)$ can be determined uniquely via the fixed point equation

$$z = -\frac{1}{\mathfrak{m}(z)} + \frac{1}{\phi}\cdot\frac{1}{n}\operatorname{tr}\Big(\big(I_n + \Sigma\mathfrak{m}(z)\big)^{-1}\Sigma\Big). \tag{41}$$

See, e.g., (Knowles and Yin, 2017, 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 (13), and the Stieltjes transform $\mathfrak{m}$. This connection will prove important in some of the results ahead.

Proposition 22. *For any* $\eta > 0$ *and* $\eta = 0$ *with* $\phi^{-1} > 1$,

$$n^{-1}\mathrm{tr}\big((\Sigma + \tau_{\eta,*}I_n)^{-1}\Sigma\big) = \phi - \eta\cdot\mathfrak{m}(-\eta/\phi). \tag{42}$$

Proof By comparing (41) and the second equation of (13), 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 (42) appears somewhat purely algebraic, it actually admits a natural statistical interpretation. Suppose ξ is also Gaussian. We may then compute

$$\mathrm{df}(\widehat{\mu}_\eta) = \sum_{j=1}^n \frac{\mathrm{Cov}^X\big((X\widehat{\mu}_\eta)_j, Y_j\big)}{\sigma_\xi^2} = \operatorname{tr}\big((\widehat{\Sigma} + z_\eta I_n)^{-1}\widehat{\Sigma}\big) = n\big(\phi - \eta\cdot\mathfrak{m}_n(-z_\eta)\big). \tag{43}$$

Now comparing the above display with (42), 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":

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

Appendix B. Properties of the fixed point equations

B.1 The fixed point equation (13)

Proposition 23. *The following hold.*

1. *The fixed point equation (13) 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:*

$$\frac{1-\phi+\sqrt{\big(1-\phi\big)^2+4\mathcal{H}_\Sigma\eta}}{2\mathcal{H}_\Sigma} \le \tau_{\eta,*} \le \inf_{k\in[0:\min\{m-1,n\}]}\bigg\{\frac{\sum_{j>k}\lambda_j}{m-k}+\frac{n}{m-k}\cdot\eta\bigg\},$$
$$\frac{\sigma_\xi^2}{\phi} \le \gamma_{\eta,*}^2 \le \frac{\sigma_\xi^2+\|\Sigma\|_{\mathrm{op}}\|\mu_0\|^2}{\phi}\bigg(1+\frac{\|\Sigma\|_{\mathrm{op}}}{\tau_{\eta,*}}\bigg).$$

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

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

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

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

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 (13) as

$$\phi = \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_* I)^{-1}\Sigma\big) + \frac{\eta}{\tau_*} = \frac{1}{n}\sum_{j=1}^n \frac{\lambda_j}{\lambda_j+\tau_*} + \frac{\eta}{\tau_*} \equiv \mathsf{f}(\tau_*). \tag{44}$$

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 15, the equation $\phi\gamma_*^2 = \sigma_\xi^2 + \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_*;\tau_*)$ reads

$$\phi = \frac{1}{\gamma_*^2}\big(\sigma_\xi^2 + \tau_*^2\|(\Sigma+\tau_* I)^{-1}\Sigma^{1/2}\mu_0\|^2\big) + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_* I)^{-2}\Sigma^2\big). \tag{45}$$

As $n^{-1}\operatorname{tr}\big((\Sigma+\tau_*I)^{-2}\Sigma^2\big) < n^{-1}\operatorname{tr}\big((\Sigma+\tau_*I)^{-1}\Sigma\big) \le \phi$ by (44) and the fact $\tau_* > 0$, the above equation admits a unique solution $\gamma_* \in (0,\infty)$, analytically given by

$$\gamma_*^2 = \frac{\sigma_\xi^2 + \tau_*^2\|(\Sigma+\tau_*I)^{-1}\Sigma^{1/2}\mu_0\|^2}{\phi - \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_*I)^{-2}\Sigma^2\big)} = \frac{\sigma_\xi^2 + \tau_*^2\|(\Sigma+\tau_*I)^{-1}\Sigma^{1/2}\mu_0\|^2}{\frac{\eta}{\tau_*} + \frac{\tau_*}{n}\operatorname{tr}\big((\Sigma+\tau_*I)^{-2}\Sigma\big)}. \tag{46}$$

(2). For the upper bound for τ_*, using the equation (44), we have

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

Solving for τ_* yields the desired upper bound. For the lower bound for τ_*, note that (44) leads to

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

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

On the other hand, the lower bound $\gamma_*^2 \ge \sigma_\xi^2/\phi$ is trivial by (46). For the upper bound for γ_*, using that

$$\phi - \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_*I)^{-2}\Sigma^2\big) \ge \phi - \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_*I)^{-1}\Sigma\big)\cdot\max_{j\in[n]}\frac{\lambda_j}{\lambda_j+\tau_*} \ge \phi\cdot\frac{\tau_*}{\|\Sigma\|_{\mathrm{op}}+\tau_*}, \tag{47}$$

and the first identity in (46), we have

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

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 (44) 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

$$\tau_*'(\eta) = \frac{\tau_*(\eta)}{\eta + \tau_*^2(\eta)\cdot T_{-2,1}(\eta)} \equiv \frac{\tau_*(\eta)}{G_0(\eta)}. \tag{48}$$

Further taking derivative with respect to η on both sides of the above display (48), we have

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

$$= \frac{2\tau_*^2(\eta)\tau_*'(\eta)}{G_0^2(\eta)}\Big(\tau_*(\eta)T_{-3,1}(\eta) - T_{-2,1}(\eta)\Big) = -\frac{2\tau_*^2(\eta)\tau_*'(\eta)}{G_0^2(\eta)}T_{-3,2}(\eta). \tag{49}$$

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

$$1 \lesssim \inf_{\eta\in\Xi_K}(-1)^{q+1}\tau_*^{(q)}(\eta) \le \sup_{\eta\in\Xi_K}(-1)^{q+1}\tau_*^{(q)}(\eta) \lesssim 1. \tag{50}$$

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

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

Then

$$\gamma_*'(\eta) = \frac{G_1'(\eta)G_2(\eta) - G_1(\eta)G_2'(\eta)}{2\gamma_*(\eta)G_2^2(\eta)}. \tag{51}$$

We shall now prove bounds for G_1, G_1', G_2, G_2'. First, using (47), 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 $\eta \in \Xi_K$,

$$G_1(\eta), G_2(\eta) \asymp 1. \tag{52}$$

The derivatives G_1', G_2' are

$$\begin{aligned} G_1'(\eta) &= 2\tau_*(\eta)\tau_*'(\eta)\|(\Sigma + \tau_*(\eta)I)^{-1}\Sigma^{1/2}\mu_0\|^2 \\ &\qquad - 2\tau_*^2(\eta)\|(\Sigma + \tau_*(\eta)I)^{-3/2}\Sigma^{1/2}\mu_0\|^2 \cdot \tau_*'(\eta), \\ G_2'(\eta) &= 2 \cdot n^{-1}\operatorname{tr}\big((\Sigma + \tau_*(\eta)I)^{-3}\Sigma^2\big) \cdot \tau_*'(\eta). \end{aligned}$$

Using the apriori estimates on $\tau_*(\eta)$ and (50), it now follows that

$$\sup_{\eta\in\Xi_K}\big\{|G_1'(\eta)| \vee |G_2'(\eta)|\big\} \lesssim 1. \tag{53}$$

Combining (51)-(53) and using apriori estimates on $\gamma_*(\eta)$, we arrive at

$$\sup_{\eta\in\Xi_K}|\gamma_*'(\eta)| \lesssim 1. \tag{54}$$

The claim follows by collecting (50) and (54). ■

B.2 Sample version of (13)

Let the sample version of (13) be defined by

$$\begin{cases} \phi e_h^2\gamma^2 = \sigma_\pm^2(L_w) + \mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau), \\ \big(\phi e_h^2 - \frac{\eta}{\tau}\big)\cdot\gamma^2 = \mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau). \end{cases} \tag{55}$$

Here recall that e_h^2 is defined in (40), and $\sigma_\pm^2(L_w)$ is defined in (30).

Proposition 24. *$1/K \le \phi^{-1}, \sigma_\xi^2 \le K$ 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*

$$\mathscr{E}_{\Delta,\Xi}(M) \equiv \Big\{ \max_{\ell=1,2} \sup_{\tau \ge 0} |\Delta_\ell(\tau)| \vee \max_{\ell=1,2} \sup_{\tau \ge 0} n^{-1/2} |\Xi_\ell(\tau) - \mathbb{E}\,\Xi_\ell(\tau)| \le M \Big\}$$

with Δ_ℓ, Ξ_ℓ defined in Lemmas 25 and 26 ahead, the following hold.

1. *All solutions $(\gamma_{n,\eta,\pm}, \tau_{n,\eta,\pm})$ to the system of equations in (55) 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,*

$$\sup_{\eta \in \Xi_K} \big\{ |\tau_{n,\eta,\pm} - \tau_{\eta,*}| \vee |\gamma_{n,\eta,\pm} - \gamma_{\eta,*}| \big\} \le C_0 \cdot \big(M/\sqrt{n} + \delta \big).$$

We need two concentration lemmas before the proof of Proposition 24.

Lemma 25. *Let $\Delta_\ell(\tau) \equiv -\ell \cdot \tau \big\langle (\Sigma + \tau I)^{-\ell} \Sigma^{\ell - 1/2} \mu_0, g \big\rangle$ 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 $t \ge C \log(en)$,*

$$\mathbb{P}\Big(\max_{\ell=1,2} \sup_{\tau \ge 0} |\Delta_\ell(\tau)| \ge C\sqrt{t} \Big) \le e^{-t}.$$

Lemma 26. *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 $t \ge C \log(en)$,*

$$\mathbb{P}\Big(\max_{\ell=1,2} \sup_{\tau \ge 0} \big| \Xi_\ell(\tau) - \mathbb{E}\,\Xi_\ell(\tau) \big| \ge C(\sqrt{nt} + t) \Big) \le e^{-t}.$$

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

Proof [Proof of Proposition 24] All the constants in $\lesssim$, $\gtrsim$, $\asymp$ and $\mathcal{O}$ 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 (38), any solution (γ_n, τ_n) to the equations in (55) satisfies

$$\begin{cases} \phi e_h^2 - \frac{\eta}{\tau_n} + \frac{\Delta_1(\tau_n)}{\sqrt{n}\gamma_n} = \frac{1}{n} \operatorname{tr}\big((\Sigma + \tau_n I)^{-1} \Sigma \big) + \frac{1}{n} (\mathrm{id} - \mathbb{E}) \Xi_1(\tau_n), \\ \phi e_h^2 + \frac{\Delta_2(\tau_n)}{\sqrt{n}\gamma_n} = \frac{1}{\gamma_n^2} \big(\sigma_\pm^2 + \tau_n^2 \|(\Sigma + \tau_n I)^{-1} \Sigma^{1/2} \mu_0\|^2 \big) \\ \qquad\qquad + \frac{1}{n} \operatorname{tr}\big((\Sigma + \tau_n I)^{-2} \Sigma^2 \big) + \frac{1}{n} (\mathrm{id} - \mathbb{E}) \Xi_2(\tau_n). \end{cases} \tag{56}$$

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 19, the second equation in (56) becomes

$$\phi + \mathcal{O}\big(M \big(1 \vee \gamma_n^{-1} \big) / \sqrt{n} + \delta \big)$$

$$= \frac{1}{\gamma_n^2}\Big(\sigma_\pm^2 + \tau_n^2\|(\Sigma+\tau_n I)^{-1}\Sigma^{1/2}\mu_0\|^2\Big) + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-2}\Sigma^2\big) \gtrsim \frac{1}{\gamma_n^2}.$$

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 (56) reduce to

$$\begin{cases} \phi - \frac{\eta}{\tau_n} + \mathcal{O}(\varepsilon_n) = \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-1}\Sigma\big), \\ \phi + \mathcal{O}(\varepsilon_n) = \frac{1}{\gamma_n^2}\big(\sigma_\xi^2 + \tau_n^2\|(\Sigma+\tau_n I)^{-1}\Sigma^{1/2}\mu_0\|^2\big) + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-2}\Sigma^2\big). \end{cases} \tag{57}$$

The above equations match (13) up to the small perturbation $\mathcal{O}(\varepsilon_n) = \mathcal{O}(\varepsilon_n(M,\delta))$ 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 23.

(2). Now we shall prove the claimed error bounds. By using (44) and the first equation of (57), we have

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

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

$$\sup_{\eta \in \Xi_K} |\tau_n - \tau_*| \lesssim \varepsilon_n. \tag{58}$$

On the other hand, using (45) and the second equation of (57), we have

$$\begin{aligned} &\frac{1}{\gamma_n^2}\Big(\sigma_\xi^2 + \tau_n^2\|(\Sigma+\tau_n I)^{-1}\Sigma^{1/2}\mu_0\|^2\Big) + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_n I)^{-2}\Sigma^2\big) \\ &= \frac{1}{\gamma_*^2}\Big(\sigma_\xi^2 + \tau_*^2\|(\Sigma+\tau_* I)^{-1}\Sigma^{1/2}\mu_0\|^2\Big) + \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_* I)^{-2}\Sigma^2\big) + \mathcal{O}(\varepsilon_n). \end{aligned} \tag{59}$$

Using the error bound in (58) and apriori estimates for τ_n, τ_*, and the fact that $\mathcal{H}_\Sigma \lesssim 1$, by an easy derivative estimate we have

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

Now plugging these estimates into (59), 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

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

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

$$\sup_{\eta\in\Xi_K} |\gamma_n - \gamma_*| \lesssim \varepsilon_n. \tag{60}$$

The claimed error bounds follow by combining (58) and (60). ■

B.3 Proofs of Lemmas 25 and 26

Proof [Proof of Lemma 25] 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 $\sup_{\tau\geq Kn}|\Delta_1(\tau)| \leq \big|\sum_{j=1}^n \lambda_j^{1/2}\mu_{0,j}g_j\big| + Ce_g \cdot n^{-1/2}$, a standard concentration for the first term shows for $t \geq 1$, with probability $1-e^{-t}$,

$$\sup_{\tau\geq Kn} |\Delta_1(\tau)| \leq C_0\sqrt{t}. \tag{61}$$

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 $\varepsilon \equiv \sqrt{t}/\big\{\sqrt{n}(\sqrt{\log n}+\sqrt{t})\big\}$, we conclude that for $t \geq C_2\log(en)$, with probability $1-e^{-t}$,

$$\sup_{\tau\in[0,Kn]} |\Delta_1(\tau)| \leq C_2\sqrt{t}. \tag{62}$$

The claim follows by combining (61) and (62). ■

Proof [Proof of Lemma 26] We focus on the case $\ell = 1$ and will follow a similar idea used in the proof of Lemma 25 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 $1-e^{-t}$,

$$\sup_{\tau>Kn}\Big(|\Xi_1(\tau)| \vee \mathbb{E}|\Xi_1(\tau)|\Big) \lesssim 1 + t/n. \tag{63}$$

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 59 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 25 above. ■

Appendix C. Gaussian designs: Proof of Theorem 3

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

C.1 Localization of the primal problem

Proposition 27. *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*

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

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

$$\min_{w\in\mathbb{R}^n} H_\eta(w) = \min_{w\in\mathbb{R}^n}\max_{v\in\mathbb{R}^m}\left\{\frac{1}{\sqrt{n}}\langle v, Gw-\xi\rangle + F(w) - \frac{\eta\|v\|^2}{2}\right\}, \tag{64}$$

any saddle point (w_*, v_*) of (64) must satisfy $\nabla F(w_*) = -\frac{1}{\sqrt{n}}G^\top v_*$ and $\frac{1}{\sqrt{n}}(Gw_*-\xi) = \eta v_*$, or equivalently,

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

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

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

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,

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

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

C.2 Characterization of the Gordon cost optimum

Theorem 28. *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 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 $C' \log(en) \leq t \leq n/C'$, $\eta \in \Xi_K$ and $\xi \in \mathscr{E}_{1,\xi}(\sqrt{t/n})$,

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

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 28, 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 29. *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 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 24) with $\delta \in (0, 1/C^{100})$ and $M \leq \sqrt{n}/C$, we have for any $\eta \in \Xi_K$,*

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

and the following localization holds:

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

Proof We write $g_n \equiv g/\sqrt{n}$ in the proof.

(**Step 1**). Fix any $L_w, L_v > 0$. We may compute

$$\begin{aligned}
&\min_{w\in B_n(L_w)} L_\eta(w; L_v) \qquad (65)\\
&= \min_{w\in B_n(L_w)} \max_{\beta\in[0,L_v]} \left\{ \frac{\beta}{\sqrt{n}}\Big(\big\| \|w\| h - \xi \big\| - \langle g, w\rangle \Big) + F(w) - \frac{\eta\beta^2}{2} \right\}\\
&= \max_{\beta\in[0,L_v]} \min_{\gamma>0} \left\{ \frac{\beta\gamma\|h\|^2}{2n} - \frac{\eta\beta^2}{2} + \min_{w\in B_n(L_w)} \left(\frac{\beta}{2\gamma} \frac{\|\|w\|h-\xi\|^2}{\|h\|^2} - \langle w, \beta g_n\rangle + F(w) \right) \right\}.
\end{aligned}$$

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

$$\|w\|^2 + \sigma_-^2 \le \|\|w\|h - \xi\|^2/\|h\|^2 \le \|w\|^2 + \sigma_+^2, \qquad (66)$$

on the event $\mathscr{E}_1(\delta)$, we may further bound (65) as follows:

$$\begin{aligned}
&\pm \min_{w\in B_n(L_w)} L_\eta(w; L_v) \le \pm \max_{\beta\in[0,L_v]} \min_{\gamma>0}\\
&\quad \left\{ \frac{\beta\gamma\|h\|^2}{2n} - \frac{\eta\beta^2}{2} + \min_{w\in B_n(L_w)} \left(\frac{\beta}{2\gamma}\big(\|w\|^2 + \sigma_\pm^2(L_w)\big) - \langle w, \beta g_n\rangle + F(w) \right) \right\}. \qquad (67)
\end{aligned}$$

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 (67):

$$\begin{aligned}
&\max_{\beta>0} \min_{\gamma>0, w\in\mathbb{R}^n} \left\{ \frac{\beta\gamma\|h\|^2}{2n} - \frac{\eta\beta^2}{2} + \left(\frac{\beta}{2\gamma}\big(\|w\|^2 + \sigma_\pm^2\big) - \langle w, \beta g_n\rangle + F(w) \right) \right\}\\
&= \max_{\beta>0} \min_{\gamma>0} \left\{ \frac{\beta}{2} \left(\gamma\big(\phi e_h^2 - e_g^2\big) + \frac{\sigma_\pm^2}{\gamma} \right) - \frac{\eta\beta^2}{2} + \mathsf{e}_F(\gamma g_n; \gamma/\beta) \right\}. \qquad (68)
\end{aligned}$$

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

$$\begin{cases}
0 = \frac{1}{2}\big(\gamma_{n,\pm}(\phi e_h^2 - e_g^2) + \frac{\sigma_\pm^2}{\gamma_{n,\pm}}\big) - \eta\beta_{n,\pm} + \partial_\beta \mathsf{e}_F\big(\gamma_{n,\pm} g_n; \gamma_{n,\pm}/\beta_{n,\pm}\big),\\
0 = \frac{\beta_{n,\pm}}{2}\big((\phi e_h^2 - e_g^2) - \frac{\sigma_\pm^2}{\gamma_{n,\pm}^2}\big) + \partial_\gamma \mathsf{e}_F\big(\gamma_{n,\pm} g_n; \gamma_{n,\pm}/\beta_{n,\pm}\big),\\
w_{n,\pm} = \mathsf{prox}_F\big(\gamma_{n,\pm} g_n; \gamma_{n,\pm}/\beta_{n,\pm}\big).
\end{cases} \qquad (69)$$

Using the derivative formula in Lemma 17 and the form of prox_F in Lemma 16, we may compute

$$\begin{cases}
\partial_\beta \mathsf{e}_F(\gamma g_n; \gamma/\beta) = \frac{1}{2\gamma}\Big(\mathsf{err}_{(\Sigma,\mu_0)}(\gamma; \gamma/\beta) - 2\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma; \gamma/\beta) + \gamma^2 e_g^2 \Big),\\
\partial_\gamma \mathsf{e}_F(\gamma g_n; \gamma/\beta) = \frac{\beta}{2\gamma^2}\Big(\gamma^2 e_g^2 - \mathsf{err}_{(\Sigma,\mu_0)}(\gamma; \gamma/\beta) \Big).
\end{cases} \qquad (70)$$

Plugging (70) into (69), the first-order optimality condition for $(\beta_{n,\pm}, \gamma_{n,\pm})$ in the minimax program (68) is given by

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

Equivalently,

$$\begin{cases} \phi e_h^2 \gamma_{n,\pm}^2 = \sigma_\pm^2 + \mathsf{err}_{(\Sigma,\mu_0)}(\gamma_{n,\pm}; \gamma_{n,\pm}/\beta_{n,\pm}), \\ \big(\phi e_h^2 - \frac{\eta}{\gamma_{n,\pm}/\beta_{n,\pm}}\big)\gamma_{n,\pm}^2 = \mathsf{dof}_{(\Sigma,\mu_0)}(\gamma_{n,\pm}; \gamma_{n,\pm}/\beta_{n,\pm}). \end{cases} \tag{71}$$

Using the apriori estimates in Proposition 24, 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,

$$\gamma_{n,\pm} \asymp_K 1, \quad \beta_{n,\pm} \asymp_K 1. \tag{72}$$

Using the last equation of (69), we have

$$\|w_{n,\pm}\| = \Big\|\Sigma^{1/2}\Big(\Sigma + \frac{\gamma_{n,\pm}}{\beta_{n,\pm}} I\Big)^{-1}\Big(- \frac{\gamma_{n,\pm}}{\beta_{n,\pm}}\mu_0 + \gamma_{n,\pm}\Sigma^{1/2} g_n\Big)\Big\| \lesssim_K 1. \tag{73}$$

In view of (72)-(73), by choosing $L_w, L_v \in [C, C^2]$ for large enough $C > 0$, the constraints in the optimization in (67) can be dropped for free. ∎

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 30. *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 $t \ge C' \log(en)$,*

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

Proof In the proof, we write $g_n \equiv g/\sqrt{n}$. All the constants in $\lesssim, \gtrsim, \asymp$ and $\mathcal{O}$ below may depend on K.
(**Step 1**). We first prove the following: On the event $\mathscr{E}_1(\delta)$, for any $C_0 > 1$,

$$\sup_{\gamma,\tau \in [1/C_0, C_0]^2} |\partial_\# \mathsf{e}_F\big(\gamma g_n; \tau\big)| \vee |\partial_\# \,\mathbb{E}\, \mathsf{e}_F\big(\gamma g_n; \tau\big)| \lesssim 1. \tag{74}$$

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

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

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

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 (74).

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

$$\mathbb{P}\Big(\sup_{(\gamma,\tau)\in[C_0^{-1},C_0]^2} \big|(\mathrm{id}-\mathbb{E})\mathsf{e}_F(\gamma g_n;\tau)\big| \geq C_1\big(\sqrt{t/n}+t/n\big), \mathscr{E}_1(\delta)\Big) \leq C_1 e^{-t/C_1}. \tag{75}$$

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 $|\mathcal{S}(\varepsilon)| \leq C_0/\varepsilon + 1$. So on the event $\mathscr{E}_1(\delta)$, using the estimate in (74) and a union bound via the pointwise concentration inequality in Proposition 18, for $t \geq 1$, 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 16 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 18. The claim (75) follows by choosing $\varepsilon \equiv \sqrt{t/n} + t/n$ and some calculations.

(**Step 3**). By (75), 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

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

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 31. *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,

$$\big|\max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) - \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_0(\beta,\gamma)\big| \leq C\eta. \tag{76}$$

Proof The proof is essentially a deterministic version of Step 2 in the proof of Proposition 29. We give some details below. We write $g_n \equiv g/\sqrt{n}$. First, using similar calculations as that of (70),

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

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 (71), is given by

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

Finally using the apriori estimates in Proposition 23, we obtain a deterministic analogue of (72) 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 [Proof of Theorem 28] By Propositions 29, 30 and 31, there exist $C, C' > 0$ such that for any $\delta \in (0, 1/C^{100})$, $M \le \sqrt{n}/C$, $t \ge C'\log(en)$, $\xi \in \mathscr{E}_{1,\xi}(\delta)$ and $\eta \in \Xi_K$,

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

The claim now follows from the concentration estimates in Lemmas 21, 25 and 26, by choosing $M \equiv \sqrt{n}/C$ and $\delta \equiv C(\sqrt{t/n} + t/n)$, which is valid in the regime $t \le n/C_0$ for large C_0. ■

C.3 Locating the global minimizer of the Gordon objective

With $(\gamma_{\eta,*}, \tau_{\eta,*})$ denoting the unique solution to the system of equations (13), let

$$w_{\eta,*} \equiv \mathsf{prox}_F\big(\gamma_{\eta,*} g/\sqrt{n}; \tau_{\eta,*}\big) = \Sigma^{1/2}\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0\big). \tag{77}$$

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

$$D_{\eta;\varepsilon}(\mathsf{g}) \equiv \big\{w \in \mathbb{R}^n : |\mathsf{g}(w) - \mathbb{E}\,\mathsf{g}(w_{\eta,*})| \ge \varepsilon\big\}. \tag{78}$$

Theorem 32. *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 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* $L_w, L_v\in[C, C^2]$*,* $C'\log(en)\le t\le n/C'$*,* $\xi\in\mathscr{E}_{1,\xi}(\sqrt{t/n})$ *and* $\eta\in\Xi_K$*,*

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

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

$$L_{\eta,\pm}(w;L_v)\equiv\max_{\beta\in[0,L_v]}\left\{\frac{\beta}{\sqrt{n}}\Big(\|h\|\sqrt{\|w\|^2+\sigma_\pm^2(L_w)}-\langle g,w\rangle\Big)-\frac{\eta\beta^2}{2}+F(w)\right\}. \tag{79}$$

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

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

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

Proof The first claim (1) follows by the definition of $\sigma_\pm^2(L_w)$ in (30) and the simple inequality (66). For (2), note that

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

as desired. ∎

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

Proposition 34. *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 24) 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

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

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 (68) rests in its range constraint on β. As proven in (72), all solutions $\beta_{n,\pm}$ to the unconstrained minimax problem (68) 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 (68), whose minimizers $w_{n,\pm}$ admit the apriori estimate (73) (with minor modifications that change $\|\cdot\|$ to the stronger estimate in $\|\cdot\|_{\Sigma^{-1}}$).

Next, for the error bound, using the last equation in (69) and the definition of $w_{\eta,*}$ in (77), along with the estimates in Proposition 24, 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 32, but will turn out to be useful in the proof of Theorem 3 ahead.

Proposition 35. *Suppose the conditions in Theorem 32 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 $\eta\in\Xi_K$,*

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

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 34, it follows that

$$\begin{aligned}
\mathbb{P}^\xi\Big(E_1\equiv\Big\{&\|w_{n,\eta,\pm}\|_{\Sigma^{-1}}\vee\|w_{n,\eta,\pm}\|\le C_0/2,\\
&L_{\eta,\pm}(w_{n,\eta,\pm};L_v)=(68)\Big\}\Big)\ge 1-C_0e^{-n/C_0}.
\end{aligned}\tag{80}$$

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

$$\mathbb{P}^{\xi}\left(E_2(t) \equiv \left\{\|w_{n,\eta,\pm} - w_{\eta,*}\|_{\Sigma^{-1}} \leq C_0(t/n)^{1/4}\right\}\right) \geq 1 - C_0 e^{-t/C_0}. \tag{81}$$

On E_1, we may characterize the value of $L_{\eta,\pm}(w_{n,\eta,\pm}; L_v)$ by applying Propositions 29-31: for $C_0' \log(en) \leq t \leq n/C_0'$,

$$\mathbb{P}^{\xi}\left(E_3(t) \equiv \left\{\left|L_{\eta,\pm}(w_{n,\eta,\pm}; L_v) - \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_{\eta}(\beta,\gamma)\right| \leq C_0\sqrt{t/n}\right\}\right) \geq 1 - C_0 e^{-t/C_0}. \tag{82}$$

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 $E_3(t)$,

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

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

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

Consequently, by enlarging $C_0 > 0$ if necessary, using Lemma 33-(1), on $E_1 \cap E_3(t)$

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

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 33-(2),

$$\sup_{w\in B_n(C_0)} \left|L_{\eta}(w; L_v) - L_{\eta,\pm}(w; L_v)\right| \leq C_1\sqrt{t/n}. \tag{83}$$

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

$$\left|\min_{w\in\mathbb{R}^n} L_{\eta}(w; L_v) - \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_{\eta}(\beta,\gamma)\right| \leq C_2\sqrt{t/n}. \tag{84}$$

On this event, combining (83)-(84) with (82), 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 (81) that controls the $\mathbb{P}^\xi$-probability of $\|w_{n,\eta,+}-w_{\eta,*}\|_{\Sigma^{-1}}\leq C_0(t/n)^{1/4}$. ∎

Proof [Proof of Theorem 32] 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 34 and Propositions 29-31, 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 $1-C_1e^{-t/C_1}$, 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 $E_0(t)\cap E_1(t)$,

$$\begin{aligned}\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_n(L_w)}L_\eta(w;L_v)&\geq\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_n(L_w)}L_{\eta,-}(w;L_v)\\ &\geq L_{\eta,-}(w_{\eta,n,-};L_v)+\frac{1}{2}\big(\varepsilon-(C_0+C_1)(t/n)^{1/4}\big)_+^2\\ &\geq\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)+\frac{1}{2}\big(\varepsilon-(C_0+C_1)(t/n)^{1/4}\big)_+^2-C_1\sqrt{t/n}.\end{aligned}$$

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. ∎

C.4 Proof of Theorem 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 27 and the concentration estimate in Lemma 20, there exists some $C_0 = C_0(K) > 0$ such that on an event E_0 with $\mathbb{P}^\xi(E_0) \geq 1 - C_0 e^{-n/C_0}$,

$$\min_{w\in\mathbb{R}^n} H_\eta(w) = \min_{w\in\mathbb{R}^n} H_\eta(w; L_0) = \min_{w\in B_n(L_0)} H_\eta(w) = \min_{w\in B_n(L_0)} H_\eta(w; L_0). \tag{85}$$

where

$$L_0 \equiv C_0\Big\{1 + \Big(\|\Sigma^{-1}\|_{\mathrm{op}} \mathbf{1}^{-1}_{\phi^{-1}\geq 1+1/K} \wedge \eta^{-1}\Big)\Big\}. \tag{86}$$

Now we shall apply the convex(-side) Gaussian min-max theorem to obtain an upper bound for the right hand side of (85). Recall the definition of $h_\eta = h_{\eta;G}$ and ℓ_η in (28). Using Theorem 14-(2), for any $z \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_\eta(w) \geq z\Big) &\leq \mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} H_\eta(w; L_0) \geq z\Big) + \mathbb{P}^\xi(E_0^c) \\
&= \mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} \max_{v\in B_m(L_0)} h_\eta(w, v) \geq z\Big) + \mathbb{P}^\xi(E_0^c) \\
&\leq 2\,\mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} \max_{v\in B_m(L_0)} \ell_\eta(w, v) \geq z\Big) + \mathbb{P}^\xi(E_0^c) \\
&= 2\,\mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} L_\eta(w; L_0) \geq z\Big) + \mathbb{P}^\xi(E_0^c).
\end{aligned} \tag{87}$$

By Proposition 35, 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 28 and 32), 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 28: with $z \equiv \bar{z}(t) = \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + \sqrt{t/n}$, for $C_1' \log(en) \leq t \leq n/C_1'$,

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\min_{w\in B_n(L_0)} L_\eta(w; L_0) \geq \bar{z}(t)\Big) \\
&\leq \mathbb{P}^\xi\Big(\min_{w\in B_n(C_1)} L_\eta(w; L_0) \geq \bar{z}(t)\Big) + \mathbb{P}^\xi(E_1^c) \leq C_1 e^{-t/C_1} + \mathbb{P}^\xi(E_1^c).
\end{aligned} \tag{88}$$

Combining (87)-(88), by enlarging C_1 if necessary, for $C_1' \log(en) \leq t \leq n/C_1'$, and $\eta \in \Xi_K$,

$$\mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_\eta(w) \geq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) + \sqrt{t/n}\Big) \leq C_1 e^{-t/C_1}. \tag{89}$$

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

$$\mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^n} H_\eta(w) \leq \max_{\beta>0}\min_{\gamma>0} \overline{\mathsf{D}}_\eta(\beta,\gamma) - \sqrt{t/n}\Big) \leq C_1 e^{-t/C_1}. \tag{90}$$

(**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 (78), with a suitable choice of ε. Let us take $C_2, C_2' > 0$ to be

the constants in Theorem 32, 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 14-(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 $C_2'\log(en) \le t \le n/C_2'$,

$$\mathbb{P}^\xi\left(\min_{w\in B_n(L_0)\cap D_{\eta;\varepsilon(t,n)}(\mathsf{g})} H_\eta(w) \le \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) + 2\sqrt{\frac{t}{n}}\right) \le 2C_2 e^{-t/C_2}. \tag{91}$$

(**Step 3**). Combining (91) and the localization in (85), 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 $\mathbb{P}^\xi(E_3(t)) \ge 1 - C_3 e^{-t/C_3}$,

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

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 $\varepsilon \in (0, 1/2]$,

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

(**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 $\widehat{\mu}_\eta = n^{-1}X^\top\big(XX^\top/n + \eta I\big)^{-1}Y$, for $\eta_1, \eta_2 \in [0, K]$,

$$\begin{aligned}\|\widehat{\mu}_{\eta_1} - \widehat{\mu}_{\eta_2}\| &\lesssim n^{-1}\|G\|_{\mathrm{op}}(\|G\|_{\mathrm{op}} + \|\xi\|)\cdot\|\big(XX^\top/n + \eta_1 I\big)^{-1} - \big(XX^\top/n + \eta_2 I\big)^{-1}\|_{\mathrm{op}}\\ &\lesssim \|\Sigma^{-1}\|_{\mathrm{op}}^2\cdot\Big(1 + \frac{\|G\|_{\mathrm{op}} + \|\xi\|}{\sqrt{n}}\Big)^2\cdot\|(GG^\top/n)^{-1}\|_{\mathrm{op}}^2\cdot|\eta_1 - \eta_2|.\end{aligned} \tag{92}$$

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 $\mathbb{P}^\xi(E_4) \ge 1 - C_4 e^{-n/C_4}$,

$$\|\widehat{\mu}_{\eta_1} - \widehat{\mu}_{\eta_2}\| \le C_4 n^2|\eta_1 - \eta_2|. \tag{93}$$

On the other hand, note that for $\eta_1, \eta_2 \in [0, K]$, using Proposition 23-(3),

$$\|\widehat{\mu}^{\mathsf{seq},*}_{\eta_1;(\Sigma,\mu_0)} - \widehat{\mu}^{\mathsf{seq},*}_{\eta_2;(\Sigma,\mu_0)}\| \lesssim (1 \vee e_g)\|\Sigma^{-1}\|^2_{\mathrm{op}}|\eta_1 - \eta_2|. \tag{94}$$

So we have

$$\big|\, \mathbb{E}\, \mathsf{g}_0\big(\widehat{\mu}^{\mathsf{seq},*}_{\eta_1;(\Sigma,\mu_0)}\big) - \mathbb{E}\, \mathsf{g}_0\big(\widehat{\mu}^{\mathsf{seq},*}_{\eta_2;(\Sigma,\mu_0)}\big)\big| \le C_4 n^2 |\eta_1 - \eta_2|. \tag{95}$$

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

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\sup_{\eta\in[0,K]} \big|\mathsf{g}_0(\widehat{\mu}_\eta) - \mathbb{E}\, \mathsf{g}_0\big(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}\big)\big| \ge 2\varepsilon\Big) \\
&\le \mathbb{P}^\xi\Big(\max_{\eta\in\Lambda_\varepsilon} \big|\mathsf{g}_0(\widehat{\mu}_\eta) - \mathbb{E}\, \mathsf{g}_0\big(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}\big)\big| \ge \varepsilon\Big) + \mathbb{P}(E_4^c) \\
&\le (1 + 2C_4Kn^2/\varepsilon)\cdot C_3 n e^{-n\varepsilon^4/C_3} + C_4 e^{-n/C_4} \le C\cdot\varepsilon^{-1}n^3 e^{-n\varepsilon^4/C}.
\end{aligned} \tag{96}$$

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 $\mathbb{P}(\xi \in \mathscr{E}_{1,\xi}(\sqrt{t/n})) \ge 1 - Ce^{-t/C}$.

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 $\widehat{\mu}_\eta = n^{-1}\big(X^\top X/n + \eta I\big)^{-1}X^\top Y$. As $\eta \ge 1/K$, the spectrum of the middle inverse matrix is bounded by $1/\eta \le K$, so we may replicate the above calculations in (93) and (95) to reach a similar estimate as in (96). □

C.5 Proof of Theorem 3 for $\widehat{r}_{\eta;G}$

Recall the cost function $h_\eta = h_{\eta;G}, \ell_\eta$ defined in (28). It is easy to see that

$$\widehat{v}_\eta \equiv \arg\max_{v\in\mathbb{R}^m}\min_{w\in\mathbb{R}^n} h_\eta(w, v) = \frac{1}{\sqrt{n}\eta}(G\widehat{w}_\eta - \xi) = -\frac{\widehat{r}_\eta}{\eta}. \tag{97}$$

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

$$v_{\eta,*} \equiv \frac{1}{\phi\tau_{\eta,*}}\left(\sqrt{\phi\gamma^2_{\eta,*} - \sigma^2_\xi}\cdot\frac{h}{\sqrt{n}} - \frac{\xi}{\sqrt{n}}\right) \tag{98}$$

in the Gordon problem.

Proposition 36. *Suppose the following hold for some $K > 0$.*

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

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

$$\mathbb{P}^\xi\Bigg(\begin{array}{l}\textit{The map } v \mapsto \ell_\eta(w_{\eta,*}; v) \textit{ is } \eta\textit{-strongly concave with unique maximizer } v_{\eta,n} \\ \textit{satisfying } \|v_{\eta,n}\| \le C \textit{ and } \|v_{\eta,n} - v_{\eta,*}\| \le C\sqrt{t/n}. \\ \textit{Furthermore, } \big|\max_v \ell_\eta(w_{\eta,*}, v) - \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)\big| \le C\sqrt{t/n}.\end{array}\Bigg) \ge 1 - Ce^{-t/C}.$$

We need the following before the proof of Proposition 36.

Lemma 37. *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* $K > 0$. *Recall* $w_{\eta,*}$ *defined in (77). Then there exist constants* $C, C' > 0$ *depending on* K *such that for* $C' \log(en) \le t \le n/C'$, $\eta \in \Xi_K$ *and* $\xi \in \mathscr{E}_{1,\xi}(\sqrt{t/n})$,

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

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

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

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 $\|\nabla H_1(g)\|^2 = 4\|(\Sigma + \tau_{\eta,*}I)^{-1}\Sigma g\|^2 \le 4H_1(g)$ and $\mathbb{E}\, H_1(g) \le n$, so Proposition 59 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 $\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma;\tau) = (\Sigma + \tau I)^{-1}\Sigma^{1/2}\big(\Sigma^{1/2}\mu_0 + \gamma g/\sqrt{n}\big)$, 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 [Proof of Proposition 36] 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 37, 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 $E_0(t)$,

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

Note that $n^{-1}\mathrm{tr}\big((\Sigma + \tau_{\eta,*}I)^{-1}\Sigma\big) \gtrsim \mathcal{H}_{\Sigma}^{-1} \gtrsim 1$, 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:

$$v_{\eta,n} = \frac{1}{\sqrt{n}\eta}\bigg(1 - \frac{\langle g, w_{\eta,*}\rangle}{\|\,\|w_{\eta,*}\|h-\xi\|}\bigg)_+ \cdot \big(\|w_{\eta,*}\|h - \xi\big). \tag{99}$$

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 (99). 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 (98) into the following form

$$v_{\eta,*} = \frac{1}{\sqrt{n}\eta}\bigg(1 - \frac{\mathbb{E}\langle g, w_{\eta,*}\rangle}{\mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|\cdot h-\xi\|^2}\bigg) \cdot \big(\mathbb{E}^{1/2}\,\|w_{\eta,*}\|^2 \cdot h - \xi\big). \tag{100}$$

This can be seen by observing

$$\begin{cases} \mathbb{E}\,\|w_{\eta,*}\|^2 = \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) = \phi\gamma_{\eta,*}^2 - \sigma_\xi^2, \\ \mathbb{E}\langle g, w_{\eta,*}\rangle = \frac{\sqrt{n}}{\gamma_{\eta,*}} \cdot \mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) = \sqrt{n}\gamma_{\eta,*} \cdot \big(\phi - \frac{\eta}{\tau_{\eta,*}}\big), \\ \mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|\cdot h-\xi\|^2 = \sqrt{m}\big(\mathbb{E}\,\|w_{\eta,*}\|^2 + \sigma_\xi^2\big)^{1/2} = \sqrt{m\phi}\gamma_{\eta,*}, \end{cases} \tag{101}$$

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 (99)-(100), we may use Lemma 37 to estimate

$$\begin{aligned}\|v_{\eta,n}-v_{\eta,*}\| &\le \frac{1}{\sqrt{n}\eta}\big|\|w_{\eta,*}\|-\mathbb{E}^{1/2}\|w_{\eta,*}\|^2\big|\cdot\|h\| \\ &\quad+\frac{1}{\sqrt{n}\eta}\bigg|\frac{\langle g,w_{\eta,*}\rangle}{\|\,\|w_{\eta,*}\|h-\xi\|}-\frac{\mathbb{E}\langle g,w_{\eta,*}\rangle}{\mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|\cdot h-\xi\|^2}\bigg|\|\mathbb{E}^{1/2}\|w_{\eta,*}\|^2\cdot h-\xi\| \\ &\equiv V_1+V_2. \end{aligned}\tag{102}$$

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

$$V_1\lesssim\frac{\|h\|}{\sqrt{n}}\cdot\frac{|\,\|w_{\eta,*}\|^2-\mathbb{E}\,\|w_{\eta,*}\|^2|}{\mathbb{E}^{1/2}\,\|w_{\eta,*}\|^2}\lesssim\sqrt{t/n}.\tag{103}$$

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

$$\begin{aligned}V_2&\lesssim\|\,\|w_{\eta,*}\|h-\xi\|^{-1}\cdot\big|\langle g,w_{\eta,*}\rangle-\mathbb{E}\langle g,w_{\eta,*}\rangle\big| \\ &\qquad+\mathbb{E}\langle g,w_{\eta,*}\rangle\cdot\big|\,\|\,\|w_{\eta,*}\|h-\xi\|^{-1}-\mathbb{E}^{-1/2}\|\,\|w_{\eta,*}\|\cdot h-\xi\|^2\big| \\ &\lesssim n^{-1/2}\big|\langle g,w_{\eta,*}\rangle-\mathbb{E}\langle g,w_{\eta,*}\rangle\big|+n^{-1/2}\big|\,\|\,\|w_{\eta,*}\|h-\xi\|-\mathbb{E}^{1/2}\|\,\|w_{\eta,*}\|h-\xi\|^2\big| \\ &\lesssim n^{-1/2}\big|\langle g,w_{\eta,*}\rangle-\mathbb{E}\langle g,w_{\eta,*}\rangle\big|+n^{-1}\big|\,\|\,\|w_{\eta,*}\|h-\xi\|^2-\mathbb{E}\|\,\|w_{\eta,*}\|h-\xi\|^2\big| \\ &\lesssim\sqrt{t/n}.\end{aligned}\tag{104}$$

The desired estimate for $\|v_{\eta,n}-v_{\eta,*}\|$ follows from (102)-(104).
(**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

$$\begin{aligned}&\max_{v\in\mathbb{R}^m}\ell_\eta(w_{\eta,*},v)\\ &\equiv\max_{v\in\mathbb{R}^m}\bigg\{\frac{1}{\sqrt{n}}\Big(-\|v\|\langle g,w_{\eta,*}\rangle+\|w_{\eta,*}\|\langle h,v\rangle-\langle v,\xi\rangle\Big)+F(w_{\eta,*})-\frac{\eta\|v\|^2}{2}\bigg\}\\ &=\frac{1}{2n\eta}\big(\|\,\|w_{\eta,*}\|h-\xi\|-\langle g,w_{\eta,*}\rangle\big)_+^2+F(w_{\eta,*}).\end{aligned}\tag{105}$$

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

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

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:

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

$$= \frac{\beta_{\eta,*}}{\gamma_{\eta,*}}\Big(\phi\gamma_{\eta,*}^2 - \mathbb{E}\,\mathsf{dof}^*_{\eta;(\Sigma,\mu_0)}\Big) - \frac{\eta\beta_{\eta,*}^2}{2} + \mathbb{E}\,F(w_{\eta,*}).$$

Further using the second and third equations in (101), it now follows that

$$\begin{aligned}\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) &= \frac{\beta_{\eta,*}}{\sqrt{n}}\Big(\mathbb{E}^{1/2}\,\|\,\|w_{\eta,*}\|h-\xi\|^2 - \mathbb{E}\langle g, w_{\eta,*}\rangle\Big) - \frac{\eta\beta_{\eta,*}^2}{2} + \mathbb{E}\,F(w_{\eta,*})\\ &= \frac{1}{2n\eta}\Big(\mathbb{E}^{1/2}\,\|\,\|w_{\eta,*}\|h-\xi\|^2 - \mathbb{E}\langle g, w_{\eta,*}\rangle\Big)^2 + \mathbb{E}\,F(w_{\eta,*}). \qquad (106)\end{aligned}$$

Now combining (105) and (106), 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 [Proof of Theorem 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

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

(**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 $C_1'\log(en)\le t\le n/C_1'$,

$$\mathbb{P}^\xi\Big(E_1(t)^c \equiv \Big\{\max_{v\in D_{\eta;C_1(t/n)^{1/4}}(\mathsf{h})}\ell_\eta(w_{\eta,*},v)\ge\overline{\mathscr{D}}_\eta - C_1^{-1}\sqrt{t/n}\Big\}\Big)\le C_1e^{-t/C_1}. \qquad (107)$$

To this end, first note that by the Lipschitz property of h, the Gaussian concentration and Proposition 36, there exist some $C_0, C_0'>0$ depending on K such that for $C_0'\log(en)\le t\le n/C_0'$, on an event $E_{1,0}(t)$ with probability at least $1-C_0e^{-t/C_0}$, we have uniformly in $v\in\mathsf{D}_{\eta;\varepsilon}(\mathsf{h})$,

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

and all the properties in Proposition 36 hold. In other word, on $E_{1,0}(t)$ with the prescribed range of t,

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

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

$$\max_{v\in D_{\eta;\varepsilon}(\mathsf{h})}\ell_\eta(w_{\eta,*},v) \le \ell_\eta(w_{\eta,*},v_{\eta,n}) - \frac{\eta}{2}\inf_{v\in\mathsf{D}_{\eta;\varepsilon}(\mathsf{h})}\|v - v_{\eta,n}\|^2$$

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

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 $E_{1,0}(t)$,

$$\max_{v\in D_{\eta;\varepsilon_{\eta;v}(t,n)}(\mathsf{h})} \ell_\eta(w_{\eta,*},v) \leq \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) - C_1\sqrt{t/n}. \tag{108}$$

Adjusting constants proves the claim in (107).

(**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 $C_2'\log(en) \leq t \leq n/C_2'$,

$$\mathbb{P}^\xi\Big(E_2(t)^c \equiv \Big\{\max_{v\in D_{\eta;C_2(t/n)^{1/4}}(\mathsf{h})\cap B_m(L_v)}\min_{w\in\mathbb{R}^n} h_\eta(w,v) \geq \overline{\mathscr{D}}_\eta - C_2^{-1}\sqrt{t/n}\Big\}\Big) \leq C_2 e^{-t/C_2}. \tag{109}$$

To see this, first note by Proposition 35, there exists some $C_2 = C_2(K) > 0$ such that on an event $E_{2,0}$ with $\mathbb{P}^\xi(E_{2,0}) \geq 1 - C_2e^{-n/C_2}$, $\|w_{\eta,*}\| \leq C_2$. 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 14-(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 (107) 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 3. In particular, using (85), (86) and (90), there exist $C_3, C_3', C_3'' > 0$ depending on K, such that for $C_3'\log(en) \leq t \leq n/C_3'$,

$$\mathbb{P}^\xi\Big(E_{3,0}(t)^c \equiv \Big\{\max_{v\in B_m(C_3'')}\min_{w\in\mathbb{R}^n} h_\eta(w,v) \leq \overline{\mathscr{D}}_\eta - C_3^{-1}\sqrt{t/n}\Big\}\Big) \leq C_3e^{-t/C_3}, \tag{110}$$

and

$$\mathbb{P}^\xi\Big(E_{3,1}^c \equiv \Big\{\max_{v\in B_m(C_3'')}\min_{w\in\mathbb{R}^n} h_\eta(w,v) = \max_{v\in\mathbb{R}^m}\min_{w\in\mathbb{R}^n} h_\eta(w,v)\Big\}\Big) \leq C_3e^{-n/C_3}. \tag{111}$$

(**Step 4**). By choosing without loss of generality $C_3 > C_2$, on the event $E_2(t) \cap E_{3,0}(t) \cap E_{3,1}$, (109)-(111) yield that for any $C' \log(en) \le t \le n/C'$,

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

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 $C_4' \log(en) \le t \le n/C_4'$ and $1/K \le \eta \le K$,

$$\mathbb{P}^\xi \Big(|\mathsf{h}(\widehat{v}_\eta) - \mathbb{E}^\xi \mathsf{h}(v_{\eta,*})| \ge C_4 (t/n)^{1/4} \Big) \le C_4 e^{-t/C_4}. \tag{112}$$

(**Step 5**). In this final step, we shall prove uniform version of the estimate (112). For $\eta_1, \eta_2 \in [1/K, K]$, using the definition of $\widehat{v}_\eta$ in (97),

$$\begin{aligned}
&|\mathsf{h}(\widehat{v}_{\eta_1}) - \mathsf{h}(\widehat{v}_{\eta_2})| \le \|\widehat{v}_{\eta_1} - \widehat{v}_{\eta_2}\| \\
&\le n^{-1/2} \big\| \eta_1^{-1} G \widehat{w}_{\eta_1} - \eta_2^{-1} G \widehat{w}_{\eta_2} \big\| + (\|\xi\|/\sqrt{n}) \cdot |\eta_1^{-1} - \eta_2^{-1}| \\
&\le \frac{\|G\widehat{w}_{\eta_1}\| + \|\xi\|}{\sqrt{n}} \cdot |\eta_1^{-1} - \eta_2^{-1}| + \frac{1}{\sqrt{n}\eta_2} \big\| G(\widehat{w}_{\eta_1} - \widehat{w}_{\eta_2}) \big\| \\
&\lesssim \Big(1 + \|\widehat{\mu}_{\eta_1}\| \frac{\|G\|_{\mathrm{op}}}{\sqrt{n}} \Big) \cdot |\eta_1 - \eta_2| + \frac{\|G\|_{\mathrm{op}}}{\sqrt{n}} \cdot \|\widehat{\mu}_{\eta_1} - \widehat{\mu}_{\eta_2}\|.
\end{aligned}$$

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

$$|\mathsf{h}(\widehat{v}_{\eta_1}) - \mathsf{h}(\widehat{v}_{\eta_2})| \lesssim \big(1 + \|G\|_{\mathrm{op}}/\sqrt{n}\big)^3 \cdot \big(|\eta_1 - \eta_2| \vee \|\widehat{\mu}_{\eta_1} - \widehat{\mu}_{\eta_2}\| \big). \tag{113}$$

In view of (93), there exists some $C_5 > 0$ depending on K, such that on an event $E_{5,1}$ with $\mathbb{P}^\xi(E_{5,1}) \ge 1 - C_5 e^{-n/C_5}$,

$$|\mathsf{h}(\widehat{v}_{\eta_1}) - \mathsf{h}(\widehat{v}_{\eta_2})| \le C_5 n^2 |\eta_1 - \eta_2|. \tag{114}$$

On the other hand, using the definition of $v_{\eta,*}$ in (98), Proposition 23-(3) and the fact that $\phi\gamma_{\eta,*}^2 - \sigma_\xi^2 = \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma_{\eta,*}; \tau_{\eta,*}) \ge \mathrm{tr}\,\big((\Sigma + \tau_{\eta,*} I)^{-2}\Sigma^2\big) \gtrsim 1$, we have

$$\begin{aligned}
&\big| \mathbb{E}^\xi \mathsf{h}(v_{\eta_1,*}) - \mathbb{E}^\xi \mathsf{h}(v_{\eta_2,*}) \big| \le \mathbb{E}^{1/2,\xi} \|v_{\eta_1,*} - v_{\eta_2,*}\|^2 \\
&\lesssim \big| \tau_{\eta_1,*}^{-1} \sqrt{\phi\gamma_{\eta_1,*}^2 - \sigma_\xi^2} - \tau_{\eta_2,*}^{-1} \sqrt{\phi\gamma_{\eta_2,*}^2 - \sigma_\xi^2} \big| + \big| \tau_{\eta_1,*}^{-1} - \tau_{\eta_2,*}^{-1} \big| \\
&\lesssim \big| \gamma_{\eta_1,*}^2 - \gamma_{\eta_2,*}^2 \big| + \big| \tau_{\eta_1,*}^{-1} - \tau_{\eta_2,*}^{-1} \big| \le C_5 |\eta_1 - \eta_2|.
\end{aligned} \tag{115}$$

Now we may mimic the proof in (96) to conclude that, by possibly enlarging $C_5 > 0$, for any $\varepsilon \in (0, 1/2]$ and $\xi \in \mathscr{E}_{1,\xi}(\varepsilon^2/C_5)$,

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

as desired. ∎

Appendix D. Universality: Proof of Theorem 4

D.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 (Han and Shen, 2023, Theorem 2.3).

Theorem 38. *Suppose $1/K\le\phi^{-1}\le K$ 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 $L_n\ge1$, and any $\mathsf{T}\in C^3(\mathbb{R})$, we have

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

Here $K_{\mathsf{T}}\equiv1+\max_{\ell\in[0:3]}\|\mathsf{T}^{(\ell)}\|_\infty$, and $\mathsf{r}_{\mathsf{f}}(L_n)$ is defined by

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

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 $Q:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$

$$X(u,w;A)\equiv u^\top Aw+Q(u,w).\tag{116}$$

The following theorem is proved in (Han and Shen, 2023, Theorem 2.5).

Theorem 39. *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 $L_u,L_w\ge1$, and any $\mathsf{T}\in C^3(\mathbb{R})$, we have*

$$\begin{aligned}&\Big|\mathbb{E}\,\mathsf{T}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w}X(u,w;A)\Big)-\mathbb{E}\,\mathsf{T}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w}X(u,w;B)\Big)\Big|\\&\le C_0\cdot K_{\mathsf{T}}\cdot\inf_{\delta\in(0,1)}\Big\{M_1L\delta+\mathscr{N}_Q(L,\delta)+\log_+^{2/3}(L/\delta)\cdot(m+n)^{2/3}M_3^{1/3}L^2\Big\}.\end{aligned}$$

Here $K_{\mathsf{T}}\equiv1+\max_{\ell\in[0:3]}\|\mathsf{T}^{(\ell)}\|_\infty$, $L\equiv L_u+L_w$, $M_\ell\equiv\sum_{i\in[m],j\in[n]}\big(\mathbb{E}|A_{ij}|^\ell+\mathbb{E}|B_{ij}|^\ell\big)$, 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.

D.2 Delocalization

Recall that $\widehat{\mu}_\eta$ defined in (3) can be rewritten as

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

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

$$\widehat{\mu}_\eta = n^{-1}\big(X^\top X/n + \eta I_n\big)^{-1} X^\top Y, \quad \widehat{v}_\eta = -(\sqrt{n}\eta)^{-1}(Y - X\widehat{\mu}_\eta). \tag{117}$$

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,

$$\widehat{\mu}_\eta = n^{-1}X^\top\big(XX^\top/n + \eta I_m\big)^{-1} Y, \quad \widehat{v}_\eta = -n^{-1/2}\big(XX^\top/n + \eta I_m\big)^{-1} Y. \tag{118}$$

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

Proposition 40. *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 $\min\{\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)), \mathbb{P}(\xi \in \mathcal{E}_\vartheta)\} \ge 1 - Ce^{-n^{2\vartheta}/C}$, such that

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

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

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

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

Remark 41. *Proposition 40 formalizes the delocalization required by our comparison argument: uniformly over $\eta \in \Xi_K$, $\widehat{w}_\eta = \Sigma^{1/2}(\widehat{\mu}_\eta - \mu_0)$ (and likewise $\widehat{v}_\eta$) is small in ℓ_∞. The set $\mathcal{U}_\vartheta$ encodes this coordinatewise control via the sequence model proxy, thereby ruling out highly localized signals. The event $\mathcal{E}_\vartheta$ imposes mild noise concentration in sup-norm and empirical variance; for i.i.d. sub-gaussian coordinates, this event holds with overwhelming probability.*

Remark 42. *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 $\|\mathsf{P}\|_{\mathrm{op}} \vee \|\mathsf{q}\| \le 1$, with a (slightly) different construction of $\mathcal{U}_\vartheta$.*

Proof [Proof of Proposition 40] All the constants in $\lesssim, \gtrsim, \asymp$ below may depend on K.
(1). Let us consider delocalization for $\widehat{w}_\eta$. Using (118), for any $s \in [n]$,

$$\begin{aligned}\langle e_s, \widehat{w}_\eta\rangle &= n^{-1}\langle \Sigma^{1/2}e_s, X^\top(\phi\check{\Sigma}+\eta I_m)^{-1}X\mu_0\rangle - \langle \Sigma^{1/2}e_s, \mu_0\rangle \\ &\quad + n^{-1}\langle \Sigma^{1/2}e_s, X^\top(\phi\check{\Sigma}+\eta I_m)^{-1}\xi\rangle \equiv A_{1;s}+A_{2;s}. \end{aligned} \tag{119}$$

We first handle $A_{1;s}$. Let ρ be the asymptotic eigenvalue density of $\check{\Sigma} = XX^\top/m$ and fix $c>0$. By (Knowles and Yin, 2017, Theorem 3.16-(i), Remark 3.17 and Lemma 4.4-(i)), for any small $\vartheta>0$ and large $D>0$,

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

holds for all $z \in [-1/c, 1/c]\times(0, 1/c]$. With $\kappa \equiv \kappa(z) \equiv \mathrm{dist}(\Re z, \mathrm{supp}\ \rho) \ge n^{-2/3+c}$, by further using the simple relation $\Im m(z)/\Im z = \int \frac{\rho(\mathrm{d}x)}{(\Re z - x)^2+\Im^2 z} \le \kappa^{-2}$, 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 $\phi^{-1} \ge 1+1/K$, according to (Bai and Silverstein, 2010, 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 $\kappa \ge \kappa_0 \equiv (\eta/\phi)\vee C_0^{-1}\mathbf{1}_{\phi^{-1}\ge 1+1/K}$. Therefore, on an event $E_{1,0;s}(b)$ with $\mathbb{P}^\xi(E_{1,0;s}(b)) \ge 1-Cn^{-D}$,

$$\begin{aligned}&\big|m^{-1}\langle \Sigma^{1/2}e_s, X^\top(\check{\Sigma}-z(0)I_m)^{-1}X\mu_0\rangle \\ &\quad - \langle \Sigma^{1/2}e_s, \mathfrak{m}(z(0))\Sigma(I_n+\mathfrak{m}(z(0))\Sigma)^{-1}\mu_0\rangle\big| \le (I)+(II)+\kappa_0^{-1}n^{-1/2+\vartheta},\end{aligned} \tag{120}$$

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

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

Now by using the concentration result in (Rudelson and Vershynin, 2009, Theorem 1.1), on an event $E_{1,1;s}$ with $\mathbb{P}^\xi(E_{1,1;s}) \ge 1-e^{-n/C}$, we have $(I)\le Cb$.

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

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

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 $w_{\eta,*} = \Sigma^{1/2}\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*}) - \mu_0\big)$ defined in (77), we then have

$$\sup_{\mu_0\in B_n(1)} \mathbb{P}^{\xi}\Big(\max_{s\in[n]}|A_{1;s} - \big\langle e_s, \mathbb{E}\, w_{\eta,*}\big\rangle| \geq Cn^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{121}$$

The term $A_{2;s}$ can be handled similarly, now reading off the $(1,2)$ element in (Knowles and Yin, 2017, Eqn. (3.10)), which shows that for any $\xi \in \mathbb{R}^m$,

$$\mathbb{P}^{\xi}\Big(\max_{s\in[n]}|A_{2;s}| \geq C(\|\xi\|/\sqrt{m})\cdot n^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{122}$$

Combining (119), (121) and (122), we have

$$\sup_{\mu_0\in B_n(1),\xi\in\mathcal{E}_\vartheta} \mathbb{P}^{\xi}\Big(\|\widehat{w}_\eta\|_\infty \geq \|\mathbb{E}\, w_{\eta,*}\|_\infty + Cn^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{123}$$

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 23-(3) and a standard Gaussian tail bound, $\mathbb{P}_{\mu_0}\big(\mathcal{U}_\vartheta \equiv \big\{\sup_{\eta\in\Xi_K}\|\mathbb{E}\, w_{\eta,*}\|_\infty \geq C_1 n^{-1/2+\vartheta}\big\}\big) \leq Ce^{-n^{2\vartheta}/C}$. 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 (118), 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 (Knowles and Yin, 2017, Eqn. (3.10)), which shows that

$$\sup_{\mu_0\in B_n(1)} \mathbb{P}^{\xi}\Big(\max_{t\in[m]}|B_{1;t}| \geq Cn^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{124}$$

The term $B_{2;t}$ relies on the local law described by the $(2,2)$ element in (Knowles and Yin, 2017, Eqn. (3.10)): for any $\xi \in \mathbb{R}^m$,

$$\mathbb{P}^{\xi}\Big(\max_{t\in[m]}|B_{2;t} - \phi^{-1}\mathfrak{m}(-\eta/\phi)\xi_t| \geq C(\|\xi\|/\sqrt{m})\cdot n^{-1/2+\vartheta}\Big) \leq Cn^{-D}. \tag{125}$$

Consequently, combining (124)-(125), 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. ∎

D.3 Universality of the global cost optimum

Theorem 43. *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 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 $\rho_0 \le 1/C$, $\eta \in \Xi_K$ and $\xi \in \mathcal{E}_\vartheta$,

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

Here $\mathcal{U}_\vartheta$ is specified as in Proposition 40.

Proof Fix $\vartheta > 0$, $\mu_0 \in \mathcal{U}_\vartheta$ and $\xi \in \mathcal{E}_\vartheta$ as specified in Proposition 40. Let $L_n \equiv C_0 n^\vartheta$. By the same proposition, with $\mathbb{P}^\xi$-probability at least $1 - C_0 n^{-100}$,

$$\begin{aligned}
\min_{w \in \mathbb{R}^n} H_{\eta;Z}(w) &= \min_{\|w\|_\infty \le L_n/\sqrt{n}} \max_{\|v\|_\infty \le L_n/\sqrt{n}} \Big\{ \frac{1}{\sqrt{n}} \langle v, Zw \rangle - \frac{1}{\sqrt{n}} \langle v, \xi \rangle - \frac{\eta}{2} \|v\|^2 + F(w) \Big\} \\
&= \min_{\|\widetilde{w}\|_\infty \le L_n} \max_{\|\widetilde{v}\|_\infty \le L_n} \Big\{ \frac{1}{n^{3/2}} \langle \widetilde{v}, Z\widetilde{w} \rangle - \frac{1}{n} \langle \widetilde{v}, \xi \rangle - \frac{\eta}{2n} \|\widetilde{v}\|^2 + F(\widetilde{w}/\sqrt{n}) \Big\}, \qquad (126)
\end{aligned}$$

and

$$\min_{w \in \mathbb{R}^n} H_{\eta;G}(w) = \min_{\|\widetilde{w}\|_\infty \le L_n} \max_{\|\widetilde{v}\|_\infty \le L_n} \Big\{ \frac{1}{n^{3/2}} \langle \widetilde{v}, G\widetilde{w} \rangle - \frac{1}{n} \langle \widetilde{v}, \xi \rangle - \frac{\eta}{2n} \|\widetilde{v}\|^2 + F(\widetilde{w}/\sqrt{n}) \Big\}. \quad (127)$$

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

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

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 39, we have for any $\mathsf{T} \in C^3(\mathbb{R})$,

$$\begin{aligned}
&\Big| \mathbb{E}^\xi \mathsf{T} \Big(\min_{\|\widetilde{w}\|_\infty \le L_n} \max_{\|\widetilde{v}\|_\infty \le L_n} X_Q(\widetilde{v}, \widetilde{w}; Z) \Big) - \mathbb{E}^\xi \mathsf{T} \Big(\min_{\|\widetilde{w}\|_\infty \le L_n} \max_{\|\widetilde{v}\|_\infty \le L_n} X_Q(\widetilde{v}, \widetilde{w}; G) \Big) \Big| \\
&\lesssim_K K_{\mathsf{T}} \cdot \inf_{\delta \in (0,1)} \Big\{ \sqrt{n} L_n \delta + L_n \delta + \log_+^{2/3}(L_n/\delta) \cdot n^{-1/6} L_n^2 \Big\} \le C_1 \cdot K_{\mathsf{T}} \cdot n^{-1/6+3\vartheta}. \qquad (128)
\end{aligned}$$

Replicating the last paragraph of proof of (Han and Shen, 2023, 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 (126)-(127), 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 (89) (in Step 1 of the final proof of Theorem 3), for $\rho_0 \in (C_3 n^{-1/2+\vartheta}, 1/C_3)$, we take $z \equiv z_\eta \equiv \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)$ and $t \equiv \rho_0^2 n/C_3$ therein, so that for $\xi \in \mathcal{E}_\vartheta \subset \mathscr{E}_{1,\xi}(\rho_0/C_3^{1/2})$,

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

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

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

The first term above can be assimilated into the second one, and $\rho_0 \ge C_3 n^{-1/2+\vartheta}$ can be dropped. The lower bound follow similarly by utilizing (90). ∎

D.4 Universality of the cost over exceptional sets

Theorem 44. *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 40.

Proof Fix $\varepsilon, \vartheta > 0$, $\mu_0 \in \mathcal{U}_\vartheta$ and $\xi \in \mathcal{E}_\vartheta$ as specified in Proposition 40. We define a renormalized version of $D_{\varepsilon;\eta}(\mathsf{g})$ as

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

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 $Z_n \equiv Z/\sqrt{n}$,

$$\begin{aligned}&\mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) \le z + \rho_0\Big) \qquad (129)\\ &= \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} \Big\{F(w) + \frac{1}{2n\eta}\|Zw - \xi\|^2\Big\} \le z + \rho_0\Big)\end{aligned}$$

$$= \mathbb{P}^{\xi}\Big(\min_{\widetilde{w}\in \widetilde{D}_{\eta;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(\sqrt{n}C_0, L_n)} \Big\{ \eta F(\widetilde{w}/\sqrt{n}) + \frac{1}{2n}\|Z_n\widetilde{w} - \xi\|^2 \Big\} \le \eta(z+\rho_0) \Big).$$

Now we may apply Theorem 38. 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

$$\mathscr{M}_{\mathsf{f}}(L,\delta) \equiv \sup_{\|\widetilde{w}\|_\infty \vee \|\widetilde{w}'\|_\infty \le L, \|\widetilde{w}-\widetilde{w}'\|_\infty \le \delta} |\mathsf{f}(\widetilde{w}) - \mathsf{f}(\widetilde{w}')| \lesssim_K (1\vee L)\delta,$$

Consequently, an application of Theorem 38 leads to

$$\begin{aligned}
&\text{RHS of (129)} - C_1\big(1 \vee (\eta\rho_0)^{-3}\big) L_n^2 n^{-1/6} \log^{2/3}(L_n n) \\
&\le \mathbb{P}^{\xi}\Big(\min_{\widetilde{w}\in \widetilde{D}_{\eta;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(\sqrt{n}C_0, L_n)} \Big\{ \eta F(\widetilde{w}/\sqrt{n}) + \frac{1}{2n}\|G_n\widetilde{w} - \xi\|^2 \Big\} \le \eta(z+3\rho_0) \Big) \\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0, \frac{L_n}{\sqrt{n}})} H_{\eta;G}(w) \le z + 3\rho_0 \Big) \\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta;\varepsilon}(\mathsf{g})\cap B_n(C_0)} H_{\eta;G}(w) \le z + 3\rho_0 \Big).
\end{aligned}$$

Here in the last inequality we simply drop the L_∞ constraint. Now for $C_2 n^{-1/2+\vartheta} \le \rho_0 \le 1/C_2$, 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 32, where C_3 is the constant therein, we have

$$\begin{aligned}
&\mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta;C_4\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C, \frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w) \le \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma) + \rho_0 \Big) \\
&\le C\Big\{ e^{-\rho_0^2 n/C_3} + (\eta\rho_0)^{-3}\cdot n^{-1/6+3\vartheta} \Big\} \le C_4\cdot(\eta\rho_0)^{-3}\cdot n^{-1/6+3\vartheta}. \qquad (130)
\end{aligned}$$

The constraints $\rho_0 \ge C_2 n^{-1/2+\vartheta}$ 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 $\phi^{-1} \ge 1 + 1/K$. 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 (76) in Proposition 31, $|z_{\eta_1} - z_{\eta_0}| \le C_5\eta$. By enlarging C_5 if necessary we assume that C_5 exceeds the constant in Lemma 45. Using Lemma 45, 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 $\rho_0 \le 1$),

$$\begin{aligned}
&\mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta_0;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0, \frac{L_n}{\sqrt{n}})} H_{\eta_0;Z}(w) \le z_{\eta_0} + \rho_0 \Big) \\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta_0;\varepsilon}(\mathsf{g})\cap B_{(2,\infty)}(C_0, \frac{L_n}{\sqrt{n}})} H_{\eta_1;Z}(w) \le z_{\eta_0} + \rho_0 \Big) \quad (\text{since } H_{\eta_1;Z} \le H_{\eta_0;Z}) \\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta_1;(\varepsilon-C_5\eta)_+}(\mathsf{g})\cap B_{(2,\infty)}(C_0, \frac{L_n}{\sqrt{n}})} H_{\eta_1;Z}(w) \le z_{\eta_0} + \rho_0 \Big) \quad (\text{by Lemma 45}) \\
&\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta_1;(\varepsilon-C_5\eta)_+}(\mathsf{g})\cap B_{(2,\infty)}(C_0, \frac{L_n}{\sqrt{n}})} H_{\eta_1;Z}(w) \le z_{\eta_1} + C_5\eta + \rho_0 \Big)
\end{aligned}$$

$$\le \mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta;C_4\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta_1;Z}(w) \le \max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_{\eta_1}(\beta,\gamma)+C\rho_0\Big)$$
$$\le C\cdot(\eta_0+\rho_0)^{-3}\rho_0^{-3}\cdot n^{-1/6+3\vartheta} \le C\cdot\rho_0^{-6}n^{-1/6+3\vartheta}.$$

The proof is complete by adjusting constants. ■

Lemma 45. *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 (77), we have

$$|\mathbb{E}\,\mathsf{g}(w_{\eta_1,*})-\mathbb{E}\,\mathsf{g}(w_{\eta_0,*})| \le \mathbb{E}\|w_{\eta_1,*}-w_{\eta_0,*}\|_{\Sigma^{-1}}$$
$$= \mathbb{E}\left\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta_1,*};\tau_{\eta_1,*})-\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta_0,*};\tau_{\eta_0,*})\right\| \le C\cdot(\eta_1-\eta_0).$$

Here the last inequality follows from the calculations in (94). So for any $w\in D_{\eta_0;\varepsilon}(\mathsf{g})$,

$$\varepsilon \le |\mathsf{g}(w)-\mathbb{E}\,\mathsf{g}(w_{\eta_0,*})| \le |\mathsf{g}(w)-\mathbb{E}\,\mathsf{g}(w_{\eta_1,*})|+C(\eta_1-\eta_0).$$

This means $w\in D_{\eta_1;(\varepsilon-C(\eta_1-\eta_0))_+}(\mathsf{g})$, as desired. ■

D.5 Proof of the universality Theorem 4 for $\widehat{\mu}_{\eta;Z}$

Fix $\vartheta>0$, $\mu_0\in\mathcal{U}_\vartheta$ and $\xi\in\mathcal{E}_\vartheta$. Let $L_n\equiv C_0n^{\vartheta}$, 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 40 and Theorem 44. By Proposition 40 and a simple ℓ_2 estimate, $\mathbb{P}^{\xi}(E_0^c)\le C_0n^{-100}$. We further let $z_\eta\equiv\max_{\beta>0}\min_{\gamma>0}\overline{\mathsf{D}}_\eta(\beta,\gamma)$ for $\eta\ge0$.

Let $\mathsf{g}:\mathbb{R}^n\to\mathbb{R}$ be 1-Lipschitz with respect to $\|\cdot\|_{\Sigma^{-1}}$. Then for $\rho_0\le1/C_0$ and $\eta\in\Xi_K$, we have

$$\mathbb{P}^{\xi}\left(\widehat{w}_{\eta;Z}\in D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g})\right)$$
$$\le \mathbb{P}^{\xi}\left(\widehat{w}_{\eta;Z}\in D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C_0,L_n/\sqrt{n})\right)+\mathbb{P}^{\xi}(E_0^c)$$
$$\le \mathbb{P}^{\xi}\Big(\min_{w\in B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w)\ge z_\eta+\rho_0\Big)$$
$$+\mathbb{P}^{\xi}\Big(\min_{w\in D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w)\le z_\eta+2\rho_0\Big)+C_0n^{-100}.$$

Here in the last inequality we used the simple fact that

$$\Big\{\min_{w\in B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w)<z_\eta+\rho_0\Big\}\cap\Big\{\min_{w\in D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C_0,\frac{L_n}{\sqrt{n}})} H_{\eta;Z}(w)>z_\eta+2\rho_0\Big\}$$
$$\subset\left\{\widehat{w}_{\eta;Z}\notin D_{\eta;C_0\rho_0^{1/2}}(\mathsf{g})\cap B_{(2,\infty)}(C_0,L_n/\sqrt{n})\right\}.$$

Invoking Theorems 43 and 44, 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 $\|\cdot\|$,

$$\mathbb{P}^\xi\Big(\big|\mathsf{g}_0(\widehat{\mu}_{\eta;Z}) - \mathbb{E}\,\mathsf{g}_0\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})\big)\big| \ge \rho_0\Big) \le C_0 \cdot \rho_0^{-12} n^{-1/6+3\vartheta}.$$

Now we may follow Step 4 in the proof of Theorem 3 to strengthen the above statement to a uniform one in η; we only sketch the differences below. Using (92) with G therein replaced by Z, and the assumption $\|\Sigma^{-1}\|_{\mathrm{op}} \le K$, we arrive at a modified form of (93): on an event E_1 with $\mathbb{P}^\xi(E_1) \ge 1 - C_1 e^{-n/C_1}$, for any $\eta_1, \eta_2 \in \Xi_K$,

$$\|\widehat{\mu}_{\eta_1;Z} - \widehat{\mu}_{\eta_2;Z}\| \le C_1|\eta_1 - \eta_2|. \tag{131}$$

Using (94) with $\|\Sigma^{-1}\|_{\mathrm{op}} \le K$, we arrive at a modified form of (95): for any $\eta_1, \eta_2 \in \Xi_K$,

$$\big|\,\mathbb{E}\,\mathsf{g}_0\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta_1,*};\tau_{\eta_1,*})\big) - \mathbb{E}\,\mathsf{g}_0\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta_2,*};\tau_{\eta_2,*})\big)\big| \le C_1|\eta_1 - \eta_2|.$$

Now using a standard discretization and a union bound, we have

$$\mathbb{P}^\xi\Big(\sup_{\eta\in\Xi_K}\big|\mathsf{g}_0(\widehat{\mu}_{\eta;Z}) - \mathbb{E}\,\mathsf{g}_0\big(\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*};\tau_{\eta,*})\big)\big| \ge \rho_0\Big) \le C_2 \cdot \rho_0^{-13} n^{-1/6+3\vartheta}.$$

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 40. □

D.6 Proof of the universality Theorem 4 for $\widehat{r}_{\eta;Z}$

Proposition 46. *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 40.

Proof Fix $\varepsilon, \vartheta > 0$, $\mu_0 \in \mathcal{U}_\vartheta$ and $\xi \in \mathcal{E}_\vartheta$ as specified in Proposition 40. We define a renormalized version of $D_{\varepsilon;\eta}(\mathsf{h})$ as

$$\widetilde{D}_{\varepsilon;\eta}(\mathsf{h}) \equiv \big\{\widetilde{r} \in \mathbb{R}^m : |\mathsf{h}(\widetilde{r}/\sqrt{n}) - \mathbb{E}^\xi\,\mathsf{h}(\widetilde{r}_{\eta,*}/\sqrt{n})| \ge \varepsilon\big\},$$

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 43. 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 39 and a similar calculation as in (128), with $s_n(\rho_0) \equiv \rho_0^{-3} n^{-1/6+3\vartheta}$,

$$\begin{aligned}
&\mathbb{P}^{\xi} \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} h_{\eta;Z}(w,v) \ge z - \rho_0 \Big) \\
&= \mathbb{P}^{\xi} \Big(\max_{\widetilde{v} \in \widetilde{D}_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n)} \min_{\widetilde{w} \in L_\infty(L_n)} \left\{ \frac{1}{n^{3/2}} \langle \widetilde{v}, Z\widetilde{w} \rangle + Q(\widetilde{v}, \widetilde{w}) \right\} \ge z - \rho_0 \Big) \\
&\le \mathbb{P}^{\xi} \Big(\max_{\widetilde{v} \in \widetilde{D}_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n)} \min_{\widetilde{w} \in L_\infty(L_n)} \left\{ \frac{1}{n^{3/2}} \langle \widetilde{v}, G\widetilde{w} \rangle + Q(\widetilde{v}, \widetilde{w}) \right\} \ge z - 3\rho_0 \Big) + C s_n(\rho_0) \\
&= \mathbb{P}^{\xi} \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} h_{\eta;G}(w,v) \ge z - 3\rho_0 \Big) + C_1 s_n(\rho_0).
\end{aligned}$$

Using the convex Gaussian min-max theorem (cf. Theorem 14),

$$\begin{aligned}
&\mathbb{P}^{\xi} \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} h_{\eta;Z}(w,v) \ge z - \rho_0 \Big) \\
&\le 2\,\mathbb{P} \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} \ell_{\eta}(w,v) \ge z - 3\rho_0 \Big) + C_1 s_n(\rho_0). \qquad (132)
\end{aligned}$$

On the other hand, using the definition of $w_{\eta,*}$ in (77), and the fact that for any $\mu_0 \in \mathcal{U}_\vartheta$, $\|\mathbb{E}\, w_{\eta,*}\|_\infty \le L_n/\sqrt{n}$, we have $\mathbb{P}\left(\|w_{\eta,*}\|_\infty \ge L_n/\sqrt{n}\right) \le C e^{-n^{2\vartheta}/C}$. Combined with (132), we have

$$\begin{aligned}
&\mathbb{P}^{\xi} \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \min_{w \in L_\infty(L_n/\sqrt{n})} h_{\eta;Z}(w,v) \ge z - \rho_0 \Big) \\
&\le 2\,\mathbb{P} \Big(\max_{v \in D_{\eta;\varepsilon}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})} \ell_{\eta}(w_{\eta,*}, v) \ge z - 3\rho_0 \Big) + C_2 s_n(\rho_0).
\end{aligned}$$

In view of (108), 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 [Proof of Theorem 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 40. 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 40, on an event E_0 with $\mathbb{P}^\xi(E_0) \geq 1 - C_0 n^{-100}$, we have $\|\widehat{w}_{\eta;Z}\|_\infty \vee \|\widehat{r}_{\eta;Z}\|_\infty \leq L_n/\sqrt{n}$ with $L_n = C_0 n^\vartheta$. Using Theorem 43, for $\rho_0 \leq 1/C$, and $\eta \in \Xi_K$, by possibly adjusting $C_0 > 0$,

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\max_{v\in L_\infty(L_n/\sqrt{n})} \min_{w\in\mathbb{R}^m} h_{\eta;Z}(w,v) \leq z_\eta - \rho_0/2\Big) \\
&\leq \mathbb{P}^\xi\Big(\min_{w\in\mathbb{R}^m} H_{\eta;Z}(w) \leq z_\eta - \rho_0/2\Big) + \mathbb{P}^\xi(E_0^c) \leq C_0\rho_0^{-3}\cdot n^{-1/6+3\vartheta}. \qquad (133)
\end{aligned}$$

Let us take $C_1 > 0$ to be the constant in Proposition 46. By noting that

$$\begin{aligned}
&\Big\{\max_{v\in L_\infty(L_n/\sqrt{n})} \min_{w\in\mathbb{R}^m} h_{\eta;Z}(w,v) > z_\eta - \rho_0/2\Big\} \\
&\quad \cap \Big\{\max_{v\in D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h})\cap L_\infty(L_n/\sqrt{n})} \min_{w\in\mathbb{R}^n} h_{\eta;Z}(w,v) < z_\eta - \rho_0\Big\} \\
&\subset \big\{\widehat{v}_{\eta;Z} \notin D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})\big\},
\end{aligned}$$

it follows from (133) and Proposition 46 that

$$\begin{aligned}
&\mathbb{P}^\xi\Big(\widehat{v}_{\eta;Z} \in D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h})\Big) \\
&\leq \mathbb{P}^\xi\Big(\widehat{v}_{\eta;Z} \in D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h}) \cap L_\infty(L_n/\sqrt{n})\Big) + \mathbb{P}^\xi\Big(\widehat{v}_{\eta;Z} \notin L_\infty(L_n/\sqrt{n})\Big) \\
&\leq \mathbb{P}^\xi\Big(\max_{v\in L_\infty(L_n/\sqrt{n})} \min_{w\in\mathbb{R}^m} h_{\eta;Z}(w,v) \leq z_\eta - \rho_0/2\Big) \\
&\quad + \mathbb{P}^\xi\Big(\max_{v\in D_{\eta;C_1\rho_0^{1/2}}(\mathsf{h})\cap L_\infty(L_n/\sqrt{n})} \min_{w\in\mathbb{R}^n} h_{\eta;Z}(w,v) \geq z_\eta - \rho_0\Big) + \mathbb{P}^\xi(E_0^c) \\
&\leq C\rho_0^{-3}\cdot n^{-1/6+3\vartheta}.
\end{aligned}$$

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 3 for $\widehat{r}_{\eta;G}$. By (113) (where G therein is replaced by Z) and (131), on an event E_1 with $\mathbb{P}^\xi(E_1) \geq 1 - Ce^{-n/C}$, for any $\eta_1, \eta_2 \in [1/K, K]$,

$$\|\widehat{r}_{\eta_1;Z} - \widehat{r}_{\eta_2;Z}\| \leq C|\eta_1 - \eta_2|.$$

On the other hand, (115) remains valid, so we may proceed with an ε-net argument over $[1/K, K]$ to conclude. ∎

Appendix E. Proof of Theorem 6

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 47. *Suppose the conditions in Theorem 6 hold for some $K > 0$. Fix $q \in [1, \infty)$. 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 $\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*}; \tau_{\eta,*}) - \mu_0\|_q = \mathbb{E}\left(\sum_{j=1}^n |a_j + b_j g_j|^q\right)^{1/q}$ 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 $c_j \in \mathbb{R}$,

$$\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*}; \tau_{\eta,*}) - \mu_0\|_q \gtrsim \mathbb{E}\left(\sum_{j=1}^n |c_j + g_j|^q\right)^{1/q}.$$

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 $\mathbb{E}\, Z^2 \lesssim \mathbb{E}\left(\sum_{j=1}^n (|g_j|^q + \mathbb{E}|g_j|^q)\right)^2 \lesssim (n^{1-q/2})^2$, 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 48. *Suppose the conditions in Theorem 6 hold for some $K > 0$. Fix $q \in [1, \infty)$. 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*

$$\sup_{\mu_0\in\mathcal{U}_\vartheta} n^{\frac{1}{2}-\frac{1}{q}} \big|\, \mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_{\eta,*}; \tau_{\eta,*}) - \mu_0\|_q - n^{-\frac{1}{2}} \|\mathrm{diag}\big(\Gamma_{\eta;(\Sigma,\|\mu_0\|)}\big)\|^{\frac{1}{2}}_{q/2} M_q \big| \leq Cn^{-\vartheta}.$$

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 52 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 $\eta \in \Xi_K$,

$$\begin{aligned}
&\big|\, \mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_\eta; \tau_\eta) - \mu_0\|_q - \mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\widetilde{\gamma}_\eta(\|\mu_0\|); \tau_\eta) - \mu_0\|_q \big| \\
&\lesssim |\gamma_\eta - \widetilde{\gamma}_\eta(\|\mu_0\|)| \cdot n^{-1/2}\, \mathbb{E}\|(\Sigma + \tau_\eta I)^{-1}\Sigma^{1/2} g\|_q \\
&\leq n^{-\frac{1}{2}-\vartheta} \cdot n^{-\frac{1}{2}+\frac{1}{q\wedge 2}} \cdot \mathbb{E}\left\{ \|(\Sigma + \tau_\eta I)^{-1}\Sigma^{1/2} g\|^{\frac{2}{q\vee 2}} \cdot \|(\Sigma + \tau_\eta I)^{-1}\Sigma^{1/2} g\|_\infty^{1-\frac{2}{q\vee 2}} \right\} \\
&\lesssim n^{-\frac{1}{2}+\frac{1}{q}-\vartheta} \big(\log n\big)^{\frac{1}{2}-\frac{1}{q\vee 2}}. \qquad (134)
\end{aligned}$$

For $g' \in \mathbb{R}^n$, let $\mathsf{f}(g') \equiv n^{-1/2}\|(\Sigma + \tau_\eta I)^{-1} g'\|_q$, and

$$\mathsf{F}_{\|\mu_0\|}(g') \equiv n^{-1/2} \big\|(\Sigma + \tau_\eta I)^{-1}\big(-\tau_\eta \|\mu_0\| g' + \widetilde{\gamma}_\eta(\|\mu_0\|)\Sigma^{1/2} g\big)\big\|_q,$$

$$\mathsf{F}_{\|\mu_0\|,0}(g') \equiv \bigg\|(\Sigma+\tau_\eta I)^{-1}\bigg(-\tau_\eta\|\mu_0\|\frac{g'}{\|g'\|}+\widetilde{\gamma}_\eta(\|\mu_0\|)\Sigma^{1/2}\frac{g}{\sqrt{n}}\bigg)\bigg\|_q.$$

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 $g' \in \mathcal{G}_{\vartheta,\|\mu_0\|}$,

$$\begin{aligned}\max\Big\{&\big|\|g'\|-\sqrt{n}\big|, n^{1-\frac{1}{q\wedge 2}}\big|\mathsf{F}_{\|\mu_0\|}(g') - \mathbb{E}_{g_0}\mathsf{F}_{\|\mu_0\|}(g_0)\big|,\\ &n^{1-\frac{1}{q\wedge 2}}\big|\mathsf{f}(g') - \mathbb{E}\,\mathsf{f}(g_0)\big|\Big\} \leq n^{\vartheta}. \end{aligned} \tag{135}$$

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 $g' \in \mathcal{G}_{\vartheta,\|\mu_0\|}$,

$$\begin{aligned}&\big|\mathsf{F}_{\|\mu_0\|}(g') - \mathsf{F}_{\|\mu_0\|,0}(g')\big| \lesssim |\mathsf{f}(g')|\cdot|1-\sqrt{n}/\|g'\||\\ &\lesssim \big(\mathbb{E}\,\mathsf{f}(g_0) + n^{-1+\frac{1}{q\wedge 2}+\vartheta}\big)\cdot n^{-\frac{1}{2}+\vartheta} \lesssim n^{-1+\frac{1}{q}+\vartheta}(\log n)^{\frac{1}{2}-\frac{1}{q\vee 2}}. \end{aligned} \tag{136}$$

Combining (135)-(136), for ϑ small enough,

$$\sup_{g'\in\mathcal{G}_{\vartheta,\|\mu_0\|}} n^{1-\frac{1}{q\wedge 2}}\big|\mathsf{F}_{\|\mu_0\|,0}(g') - \mathbb{E}_{g_0}\mathsf{F}_{\|\mu_0\|}(g_0)\big| \lesssim n^{\vartheta}. \tag{137}$$

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

$$\mathcal{V}_\vartheta \equiv \big\{\mu_0 = U_0 g' : U_0 \in [0,1], g' \in \partial\mathcal{G}_{\vartheta,U_0}\big\} \subset B_n(1),$$

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 (137),

$$n^{1-\frac{1}{q\wedge 2}}\sup_{\mu_0\in\mathcal{V}_\vartheta}\big|\,\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\widetilde{\gamma}_\eta(\|\mu_0\|);\tau_\eta) - \mu_0\|_q - \mathbb{E}_{g_0}\mathsf{F}_{\|\mu_0\|}(g_0)\big| \lesssim n^{\vartheta}. \tag{138}$$

On the other hand, using the definition of $\Gamma_{\eta;(\Sigma,\|\mu_0\|)}$ in (16), we may compute

$$\mathbb{E}_{g_0}\mathsf{F}_{\|\mu_0\|}(g_0) = \mathbb{E}\,\big\|\Gamma^{1/2}_{\eta;(\Sigma,\|\mu_0\|)}g/\sqrt{n}\big\|_q. \tag{139}$$

Combining (134), (138) and (139), for ϑ chosen small enough,

$$\begin{aligned}&\sup_{\mu_0\in\mathcal{U}_\vartheta\cap\mathcal{V}_\vartheta} n^{1/2-1/q}\big|\,\mathbb{E}\|\widehat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}(\gamma_\eta;\tau_\eta) - \mu_0\|_q - \mathbb{E}\,\big\|\Gamma^{1/2}_{\eta;(\Sigma,\|\mu_0\|)}g/\sqrt{n}\big\|_q\big|\\ &\lesssim n^{\frac{1}{2}-\frac{1}{q}}\cdot\Big(n^{-\frac{1}{2}+\frac{1}{q}-\vartheta}\big(\log n\big)^{\frac{1}{2}-\frac{1}{q\vee 2}} + n^{-1+\frac{1}{q\wedge 2}+\vartheta}\Big)\end{aligned}$$

$$= n^{-\vartheta}\big(\log n\big)^{\frac{1}{2}-\frac{1}{q\vee 2}} + n^{-\frac{1}{2}-\frac{1}{q}+\frac{1}{q\wedge 2}+\vartheta} \lesssim n^{-\vartheta/2}.$$

The claim follows from Lemma 60. ■

Proof [Proof of Theorem 6] 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 $1 \le q \le 2$. 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 3 and 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 47 and 48 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 40 and its proofs below (123), 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

$$\sup_{\mu_0\in\mathcal{U}_{\vartheta_1}} \mathbb{P}\bigg(\sup_{\eta\in\Xi_K} \Big\{ \|\widehat{\mu}_\eta - \mu_0\|_\infty \vee \|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_\infty \Big\} \ge \frac{L_n}{\sqrt{n}} \bigg) \le Cn^{-2D}, \tag{140}$$

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

$$\mathsf{g}_q(x) \equiv \bigg[\bigg(\frac{L_n}{\sqrt{n}} \bigg)^{\frac{2}{q}-1} \bigg\| (x-\mu_0) \wedge \bigg(\frac{L_n}{\sqrt{n}} \bigg)^{\frac{2}{q}-1} \vee \bigg\{ -\bigg(\frac{L_n}{\sqrt{n}} \bigg)^{\frac{2}{q}-1} \bigg\} \bigg\|_q \bigg]^{\frac{q}{2}},$$

which verifies that g_q is 1-Lipschitz with respect to $\|\cdot\|$. Using (140),

$$\inf_{\mu_0\in\mathcal{U}_{\vartheta_1}} \mathbb{P}\Big(\mathsf{g}_q(\widehat{\mu}_\eta) = n^{(1-\frac{q}{2})\vartheta_1} \cdot \big\{ n^{\frac{1}{2}-\frac{1}{q}} \|\widehat{\mu}_\eta - \mu_0\|_q \big\}^{\frac{q}{2}}, \forall \eta \in \Xi_K \Big) \ge 1 - Cn^{-D}, \tag{141}$$

and with $E_{\mu_0} \equiv \big\{ \sup_{\eta\in\Xi_K} \|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_\infty \le L_n/\sqrt{n} \big\}$,

$$\begin{aligned}
&\sup_{\mu_0\in\mathcal{U}_{\vartheta_1}} \sup_{\eta\in\Xi_K} \Big| \mathbb{E}\,\mathsf{g}_q\big(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}\big) - n^{(1-\frac{q}{2})\vartheta_1}\, \mathbb{E}\,\big\{ n^{\frac{1}{2}-\frac{1}{q}} \|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_q \big\}^{\frac{q}{2}} \Big| \\
&= n^{(1-\frac{q}{2})\vartheta_1} \sup_{\mu_0\in\mathcal{U}_{\vartheta_1}} \sup_{\eta\in\Xi_K} \mathbb{E}\,\big\{ n^{\frac{1}{2}-\frac{1}{q}} \|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_q \big\}^{\frac{q}{2}} \mathbf{1}_{E^c_{\mu_0}} \\
&\quad + \sup_{\mu_0\in\mathcal{U}_{\vartheta_1}} \sup_{\eta\in\Xi_K} \mathbb{E}\,\mathsf{g}_q\big(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}\big) \mathbf{1}_{E^c_{\mu_0}} \lesssim n^{-D}.
\end{aligned} \tag{142}$$

As the map $g \mapsto \|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_q = \|(\Sigma+\tau_{\eta,*}I)^{-1}\big(-\tau_{\eta,*}\mu_0 + \gamma_{\eta,*}\Sigma^{1/2}g/\sqrt{n}\big)\|_q$ 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 $\mu_0 \in \mathcal{U}_{\vartheta_1}, \eta \in \Xi_K$,

$$\left| \mathbb{E}\left\{ n^{\frac{1}{2}-\frac{1}{q}} \|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_q \right\}^{\frac{q}{2}} - \left\{ n^{\frac{1}{2}-\frac{1}{q}}\, \mathbb{E}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_q \right\}^{\frac{q}{2}} \right| \lesssim n^{-\frac{1}{q}+\vartheta_1}. \tag{143}$$

Combining (142)-(143), we have uniformly in $\mu_0 \in \mathcal{U}_{\vartheta_1}, \eta \in \Xi_K$,

$$\left| \mathbb{E}\, \mathsf{g}_q\big(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)}\big) - n^{(1-\frac{q}{2})\vartheta_1} \left\{ n^{\frac{1}{2}-\frac{1}{q}}\, \mathbb{E}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_q \right\}^{q/2} \right| \le C n^{(2-\frac{q}{2})\vartheta_1 - \frac{1}{q}}. \tag{144}$$

Combining (141) and (144) 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

$$\sup_{\mu_0 \in \mathcal{U}_{\vartheta_2}} \mathbb{P}\left(n^{\frac{1}{2}-\frac{1}{q}} \sup_{\eta \in \Xi_K} \left| \|\widehat{\mu}_\eta - \mu_0\|_q - \mathbb{E}\|\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\|_q \right| \ge n^{-\vartheta_2} \right) \le C n^{-1/7}.$$

The ratio formulation follows again from Lemmas 47 and 48. ■

Appendix F. Proofs for Section 3.2

F.1 A rigorous version of (17) and its proof

The following theorem follows easily from Theorems 3 and 4.

Theorem 49. *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 50.

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 55.*
2. *The closest non-asymptotic results on exact risk characterizations related to our Theorem 49, appear to be those presented in (i) (Hastie et al., 2022, 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) (Cheng and Montanari, 2024, 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 Hastie et al. (2022); Cheng and Montanari (2024) 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 49 above complements the results in Hastie et al. (2022); Cheng and Montanari (2024) by providing uniform control in η when $\phi^{-1} > 1$ (under a set of different conditions).

Proof [Proof of Theorem 49] For ϑ chosen small enough, we fix $\mu_0 \in \mathcal{U}_\vartheta$, where $\mathcal{U}_\vartheta$ is specified in Theorem 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 4-(1) with $\mathsf{g}(x) = \|\Sigma^{1/2}(x-\mu_0)\|$, on an event E_0 with $\mathbb{P}(E_0^c) \le C_0\varepsilon^{-c_0}n^{-1/6.5}$,

$$\sup_{\eta\in\Xi_K} \big|\sqrt{R(\eta)} - \mathbb{E}\|\Sigma^{1/2}\big(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\big)\|\big| \le \varepsilon.$$

By Gaussian-Poincaré inequality, $0 \le \bar{R}(\eta) - \big(\mathbb{E}\|\Sigma^{1/2}\big(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0\big)\|\big)^2 = \mathrm{Var}\big(\|\Sigma^{1/2}(\widehat{\mu}^{\mathsf{seq},*}_{\eta;(\Sigma,\mu_0)} - \mu_0)\|\big) \lesssim n^{-1}$. As $\bar{R}(\eta) \asymp 1$ uniformly in $\eta \in \Xi_K$, on E_0,

$$\sup_{\eta\in\Xi_K} \big|R^{1/2}(\eta) - \bar{R}^{1/2}(\eta)\big| \le \varepsilon + C_0' n^{-1}. \tag{145}$$

On the other hand, using both the standard form $\widehat{\mu}_\eta = n^{-1}\big(X^\top X/n + \eta I_n\big)^{-1}X^\top Y$ and the alternative form $\widehat{\mu}_\eta = n^{-1}X^\top\big(XX^\top/n + \eta I_m\big)^{-1}Y$, we have

$$\sup_{\eta\in\Xi_K} \|\widehat{\mu}_\eta\| \lesssim \Big(\|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}\mathbf{1}^{-1}_{\phi^{-1}\ge 1+1/K} \wedge 1\Big)\cdot\Big(1 + \frac{\|Z\|_{\mathrm{op}} + \|\xi\|}{\sqrt{n}}\Big)^2. \tag{146}$$

Consequently, on an event E_1 with $\mathbb{P}(E_1^c) \le C_1 e^{-n/C_1}$,

$$\sup_{\eta\in\Xi_K} \|\widehat{\mu}_\eta\| \le C_1. \tag{147}$$

Finally, using (145) and (147), 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 4 (the concrete form of $\mathcal{E}_\vartheta$ is given in Proposition 40). Using the same Theorem 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, $0 \le \mathbb{E}^\xi\|r_{\eta,*}\|^2 - \big(\mathbb{E}^\xi\|r_{\eta,*}\|\big)^2 = \mathrm{Var}^\xi\big(\|r_{\eta,*}\|\big) \lesssim 1/n$. Combined with the fact that $\mathbb{E}^\xi\|r_{\eta,*}\|^2 = (\eta\gamma_{\eta,*}/\tau_{\eta,*})^2 + \mathcal{O}(|\|\xi\|^2/m - \sigma_\xi^2|)$, for $\eta \in [1/K, K]$, using the stability estimate in Proposition 23-(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 $\varepsilon \in (Cn^{-1/2+\vartheta}, 1/C]$,

$$\mathbb{P}^\xi\Big(\sup_{\eta\in[1/K,K]}\big|\|\widehat{r}_\eta\| - \eta\gamma_{\eta,*}/\tau_{\eta,*}\big| \ge \varepsilon\Big) \le C\varepsilon^{-c_0}\cdot n^{-1/6.5}.$$

Now taking expectation over ξ, for the same range of ε,

$$\mathbb{P}\Big(\sup_{\eta\in[1/K,K]}\big|\|\widehat{r}_\eta\| - \eta\gamma_{\eta,*}/\tau_{\eta,*}\big| \ge \varepsilon\Big) \le C\varepsilon^{-c_0}\cdot n^{-1/6.5}. \tag{148}$$

On the other hand, using (146),

$$\sup_{\eta\in[1/K,K]}\|\widehat{r}_\eta\| \lesssim \Big(\|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}\mathbf{1}_{\phi^{-1}\ge 1+1/K}^{-1} \wedge 1\Big)\cdot\Big(1 + \frac{\|Z\|_{\mathrm{op}} + \|\xi\|}{\sqrt{n}}\Big)^3.$$

Consequently, on an event E_3 with $\mathbb{P}(E_3^c) \le C_3 e^{-n/C_3}$, $\sup_{\eta\in[1/K,K]}\|\widehat{r}_\eta\| \le C_3$, and therefore

$$\sup_{\eta\in[1/K,K]}\big|\|\widehat{r}_\eta\|^2 - \big(\eta\gamma_{\eta,*}/\tau_{\eta,*}\big)^2\big| \le C_3\cdot\sup_{\eta\in[1/K,K]}\big|\|\widehat{r}_\eta\| - \eta\gamma_{\eta,*}/\tau_{\eta,*}\big|.$$

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

$$\begin{aligned}\mathbb{E}^\xi\|r_{\eta,*} - \xi/\sqrt{n}\|^2 &= \phi\cdot\bigg\{\Big(\frac{\eta}{\phi\tau_{\eta,*}}\Big)^2(\phi\gamma_{\eta,*}^2 - \sigma_\xi^2) + \frac{\|\xi\|^2}{m}\cdot\Big(\frac{\eta}{\phi\tau_{\eta,*}} - 1\Big)^2\bigg\}\\ &= \Big(\frac{\eta\gamma_{\eta,*}}{\tau_{\eta,*}}\Big)^2 + \phi\sigma_\xi^2\cdot\bigg[\Big(\frac{\eta}{\phi\tau_{\eta,*}} - 1\Big)^2 - \Big(\frac{\eta}{\phi\tau_{\eta,*}}\Big)^2\bigg] + \mathcal{O}\big(|\|\xi\|^2/m - \sigma_\xi^2|\big).\end{aligned}$$

The proof is complete. ∎

F.2 Proof of Theorem 8

Lemma 51. *Suppose $1/K \le \phi^{-1} \le K$, and $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ 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 $q \in \{0, 1/2\}$,*

$$\mathbb{P}\Big(\sup_{\eta\in\Xi_K}\big|\|(\Sigma + \tau_{\eta,*}I)^{-1}\Sigma^q g/\|g\|\|^2 - n^{-1}\,\mathrm{tr}\,\big((\Sigma + \tau_{\eta,*}I)^{-2}\Sigma^{2q}\big)\big| > \varepsilon\Big) \le C\varepsilon^{-1}e^{-n\varepsilon^2/C}.$$

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

- $\|A_\eta\|_F^2 = \operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-4}\Sigma^2\big) \lesssim n(1\wedge\tau_{\eta,*})^{-4} \asymp n$, and
- $\|A_\eta\|_F^2 \gtrsim \operatorname{tr}(\Sigma^2)\cdot(1\vee\tau_{\eta,*})^{-4} \gtrsim n$,

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 $\varepsilon\in(0,1)$,

$$\begin{aligned}
&\mathbb{P}\Big(\big|\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^{1/2}g/\|g\|\|^2 - n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big)\big| > \varepsilon\Big)\\
&\le \mathbb{P}\Big(\big|n^{-1}\big(\|A_\eta^{1/2}g\|^2 - \mathbb{E}\|A_\eta^{1/2}g\|^2\big)\big| > \varepsilon/4\Big) + \mathbb{P}\big(|e_g^{-2}-1|>\varepsilon/2\big) + \mathbb{P}(e_g^2\le 1/2)\\
&\le C_1 e^{-n\varepsilon^2/C_1}.
\end{aligned}$$

On the other hand, for any $\eta_1,\eta_2\in\Xi_K$, using Proposition 23-(3),

$$\begin{aligned}
\big|\|(\Sigma+\tau_{\eta_1,*}I)^{-1}\Sigma^{1/2}g/\|g\|\|^2 - \|(\Sigma+\tau_{\eta_2,*}I)^{-1}\Sigma^{1/2}g/\|g\|\|^2\big| &\lesssim |\eta_1-\eta_2|,\\
n^{-1}\big|\operatorname{tr}\big((\Sigma+\tau_{\eta_1,*}I)^{-2}\Sigma\big) - \operatorname{tr}\big((\Sigma+\tau_{\eta_2,*}I)^{-2}\Sigma\big)\big| &\lesssim |\eta_1-\eta_2|,
\end{aligned}$$

so we may conclude by a standard discretization and union bound argument. ■

Proposition 52. *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*

$$\begin{aligned}
\frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big) &= \frac{\phi\mathfrak{m}_\eta^2}{\mathfrak{m}'_\eta}\big(\mathfrak{m}_\eta - (\eta/\phi)\mathfrak{m}'_\eta\big),\\
\frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\big) &= \frac{\phi\mathfrak{m}_\eta^2}{\mathfrak{m}'_\eta}\big((\phi^{-1}-1)\mathfrak{m}'_\eta + 2(\eta/\phi)\cdot\mathfrak{m}_\eta\mathfrak{m}'_\eta - \mathfrak{m}_\eta^2\big).
\end{aligned}$$

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

$$\sup_{\mu_0\in\mathcal{U}_\varepsilon}\sup_{\eta\in\Xi_K}\left|\gamma_{\eta,*}^2 - \frac{\sigma_\xi^2\mathfrak{m}'_\eta + \|\mu_0\|^2\big(\phi\mathfrak{m}_\eta - \eta\mathfrak{m}'_\eta\big)}{\phi\mathfrak{m}_\eta^2}\right| \le \varepsilon.$$

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 (42) with respect to η yields that

$$-n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big)\cdot\partial_\eta\tau_{\eta,*} = -\big(\mathfrak{m}_\eta - (\eta/\phi)\mathfrak{m}'_\eta\big).$$

Now using $\partial_\eta\tau_{\eta,*} = \mathfrak{m}'_\eta/(\phi\mathfrak{m}_\eta^2)$ to obtain the formula for $n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big)$.

Next, using that $\phi - \frac{\eta}{\tau_{\eta,*}} = n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-1}\Sigma\big) = 1 - \tau_{\eta,*}\cdot n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-1}\big)$, we may solve

$$n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-1}\big) = \mathfrak{m}_\eta\big(1-\phi+\eta\cdot\mathfrak{m}_\eta\big).$$

Differentiating with respect to η on both sides of the above display, we obtain

$$\begin{aligned}-n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\big)\cdot\partial_\eta\tau_{\eta,*} &= -\phi^{-1}\mathfrak{m}_\eta'\big(1-\phi+\eta\cdot\mathfrak{m}_\eta\big)+\mathfrak{m}_\eta\cdot\big(\mathfrak{m}_\eta-(\eta/\phi)\mathfrak{m}_\eta'\big)\\ &= -(\phi^{-1}-1)\mathfrak{m}_\eta' - 2(\eta/\phi)\cdot\mathfrak{m}_\eta\mathfrak{m}_\eta' + \mathfrak{m}_\eta^2,\end{aligned}$$

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

$$\mathcal{G}_\varepsilon \equiv \Big\{g\in\mathbb{R}^n : \sup_{\eta\in\Xi_K}\Big|\big\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^{1/2}\frac{g}{\|g\|}\big\|^2 - \frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big)\Big|\le\varepsilon\Big\}. \tag{149}$$

Let $\mathcal{U}_\varepsilon\equiv\{Ug/\|g\| : U\in[0,1], g\in\mathcal{G}_\varepsilon\}\subset B_n(1)$. Using Lemma 51, there exists some constant $C_0=C_0(K)>0$ such that $\mathrm{vol}(\mathcal{U}_\varepsilon)/\mathrm{vol}(B_n(1)) = \mathbb{P}_{\mu_0}(\mu_0\in\mathcal{U}_\varepsilon)\ge 1-C_0\varepsilon^{-1}e^{-n\varepsilon^2/C_0}$, and moreover,

$$\sup_{\mu_0\in\mathcal{U}_\varepsilon}\sup_{\eta\in\Xi_K}\Big|\|(\Sigma+\tau_{\eta,*}I)^{-1}\Sigma^{1/2}\mu_0\|^2 - \|\mu_0\|^2\cdot n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big)\Big|\le\varepsilon.$$

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 (46) for $\gamma_{\eta,*}^2$, and the fact that the denominator therein is of order 1 (depending on K), we have

$$\sup_{\mu_0\in\mathcal{U}_\varepsilon}\sup_{\eta\in\Xi_K}\left|\gamma_{\eta,*}^2 - \frac{\sigma_\xi^2+\|\mu_0\|^2\tau_{\eta,*}^2\cdot\frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big)}{\frac{\eta}{\tau_{\eta,*}}+\tau_{\eta,*}\cdot\frac{1}{n}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big)}\right|\le C_1\varepsilon.$$

Now using (2), the second term in the above display equals to

$$\frac{\sigma_\xi^2+\|\mu_0\|^2\cdot\frac{\phi}{\mathfrak{m}_\eta'}\big(\mathfrak{m}_\eta-\frac{\eta}{\phi}\mathfrak{m}_\eta'\big)}{\eta\mathfrak{m}_\eta+\frac{\phi\mathfrak{m}_\eta}{\mathfrak{m}_\eta'}\big(\mathfrak{m}_\eta-\frac{\eta}{\phi}\mathfrak{m}_\eta'\big)} = \frac{\sigma_\xi^2\mathfrak{m}_\eta'+\|\mu_0\|^2\big(\phi\mathfrak{m}_\eta-\eta\mathfrak{m}_\eta'\big)}{\phi\mathfrak{m}_\eta^2}.$$

The claim follows by adjusting constants. ∎

Proof [Proof of Theorem 8] As $\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) = \phi\gamma_{\eta,*}^2-\sigma_\xi^2$, directly invoking Proposition 52-(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

$$\bar{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) = \tau_{\eta,*}^2\|(\Sigma+\tau_{\eta,*}I)^{-1}\mu_0\|^2 + \gamma_{\eta,*}^2\cdot n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big). \tag{150}$$

Using a similar construction as in the proof of Proposition 52 via the help of Lemma 51, 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 52-(3) and

$$\sup_{\mu_0\in\mathcal{U}_\varepsilon}\sup_{\eta\in\Xi_K}\Big|\|(\Sigma+\tau_{\eta,*}I)^{-1}\mu_0\|^2 - \|\mu_0\|^2\cdot n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\big)\Big|\le\varepsilon \tag{151}$$

hold. Combining (150)-(151), we may set

$$\begin{aligned}\mathscr{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) &\equiv \tau_{\eta,*}^2\|\mu_0\|^2\cdot n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\big)\\ &\quad+(\phi\mathfrak{m}_\eta^2)^{-1}\Big(\sigma_\xi^2\mathfrak{m}_\eta'+\|\mu_0\|^2\big(\phi\mathfrak{m}_\eta-\eta\mathfrak{m}_\eta'\big)\Big)\cdot n^{-1}\operatorname{tr}\big((\Sigma+\tau_{\eta,*}I)^{-2}\Sigma\big)\\ &\equiv R_{2,1}+R_{2,2}.\end{aligned}$$

By Proposition 52-(2), we may compute $R_{2,1}, R_{2,2}$ separately:

$$\begin{aligned}R_{2,1} &= \|\mu_0\|^2\cdot\frac{\phi}{\mathfrak{m}_\eta'}\cdot\Big((\phi^{-1}-1)\mathfrak{m}_\eta'+2(\eta/\phi)\cdot\mathfrak{m}_\eta\mathfrak{m}_\eta'-\mathfrak{m}_\eta^2\Big)\\ &= \|\mu_0\|^2(1-\phi)+\Big\{2\|\mu_0\|^2\eta\mathfrak{m}_\eta-\|\mu_0\|^2\phi\cdot\frac{\mathfrak{m}_\eta^2}{\mathfrak{m}_\eta'}\Big\},\\ R_{2,2} &= \frac{1}{\mathfrak{m}_\eta'}\Big(\sigma_\xi^2\mathfrak{m}_\eta'+\|\mu_0\|^2\big(\phi\mathfrak{m}_\eta-\eta\mathfrak{m}_\eta'\big)\Big)\cdot\big(\mathfrak{m}_\eta-(\eta/\phi)\mathfrak{m}_\eta'\big)\\ &= \sigma_\xi^2\big(\mathfrak{m}_\eta-(\eta/\phi)\mathfrak{m}_\eta'\big)+\phi^{-1}\|\mu_0\|^2\eta^2\mathfrak{m}_\eta'-\Big\{2\|\mu_0\|^2\eta\mathfrak{m}_\eta-\|\mu_0\|^2\phi\cdot\frac{\mathfrak{m}_\eta^2}{\mathfrak{m}_\eta'}\Big\}.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathscr{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) &= \|\mu_0\|^2(1-\phi)+\sigma_\xi^2\big(\mathfrak{m}_\eta-(\eta/\phi)\mathfrak{m}_\eta'\big)+\phi^{-1}\|\mu_0\|^2\eta^2\mathfrak{m}_\eta'\\ &= \sigma_\xi^2\cdot\Big\{\mathsf{SNR}_{\mu_0}(1-\phi)+\mathfrak{m}_\eta+(\eta/\phi)\big(\eta\cdot\mathsf{SNR}_{\mu_0}-1\big)\mathfrak{m}_\eta'\Big\}.\end{aligned}$$

The claims for $\mathscr{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta)$ and $\mathscr{R}^{\mathsf{res}}_{(\Sigma,\mu_0)}(\eta)$ follow from Proposition 52-(3). ■

F.3 Proof of Proposition 9

We will prove the following version of Proposition 9, 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 9.

Proposition 53. *Recall* $\mathsf{SNR}_{\mu_0}=\|\mu_0\|^2/\sigma_\xi^2$. *Then for* $\#\in\{\mathsf{pred},\mathsf{est},\mathsf{in}\}$,

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

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* $\eta_* \equiv \mathsf{SNR}_{\mu_0}^{-1} \in \Xi_K$, *then* $1/C \le |\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta_*)|/\|\mu_0\|^2(\eta-\eta_*)^2 \le C$.

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

$$\mathfrak{m}_\eta = \tau_\eta^{-1}, \quad \mathfrak{m}'_\eta = \phi\tau'_\eta/\tau_\eta^2, \quad \mathfrak{m}''_\eta = -\phi^2\big(\tau''_\eta\tau_\eta - 2(\tau'_\eta)^2\big)/\tau_\eta^3. \tag{152}$$

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 $\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}$,

$$\begin{aligned}\partial_\eta \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) &= \sigma_\xi^2 \cdot \partial_\eta\Big\{\mathfrak{m}_\eta^{-2}\big(\phi\cdot\mathsf{SNR}_{\mu_0}\,\mathfrak{m}_\eta - \big(\eta\cdot\mathsf{SNR}_{\mu_0}-1\big)\mathfrak{m}'_\eta\big)\Big\}\\ &= \phi^{-1}\sigma_\xi^2\cdot\mathfrak{m}_\eta^{-3}\big(\mathfrak{m}_\eta\mathfrak{m}''_\eta - 2(\mathfrak{m}'_\eta)^2\big)\cdot\big(\eta\cdot\mathsf{SNR}_{\mu_0}-1\big).\end{aligned}$$

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 $\mathscr{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}$,

$$\begin{aligned}\partial_\eta \mathscr{R}^{\mathsf{est}}_{(\Sigma,\mu_0)}(\eta) &= \sigma_\xi^2\cdot\partial_\eta\big(\mathsf{SNR}_{\mu_0}(1-\phi) + \mathfrak{m}_\eta + (\eta/\phi)(\eta\cdot\mathsf{SNR}_{\mu_0}-1)\mathfrak{m}'_\eta\big)\\ &= \phi^{-1}\sigma_\xi^2\cdot\big(2\mathfrak{m}'_\eta - (\eta/\phi)\mathfrak{m}''_\eta\big)\cdot(\eta\cdot\mathsf{SNR}_{\mu_0}-1).\end{aligned} \tag{153}$$

To compute the second term in the above display, recall the identity for τ'_η, τ''_η in (48)-(49). Also recall $G_0(\eta) = \eta + \tau_\eta^2 T_{-2,1}(\eta) = \tau_\eta/\tau'_\eta$ defined in (48). Then

$$\begin{aligned}&2\mathfrak{m}'_\eta - \frac{\eta}{\phi}\mathfrak{m}''_\eta = \frac{\phi}{\tau_\eta}\Big\{\frac{2\tau'_\eta}{\tau_\eta}\Big(1-\frac{\eta\tau'_\eta}{\tau_\eta}\Big) + \frac{\eta\tau''_\eta}{\tau_\eta}\Big\} = \frac{\phi}{\tau_\eta}\Big\{\frac{2\tau'_\eta}{\tau_\eta}\frac{\tau_\eta^2 T_{-2,1}(\eta)}{G_0(\eta)} - \frac{2\eta\tau_\eta\tau'_\eta}{G_0^2(\eta)}T_{-3,2}(\eta)\Big\}\\ &= \frac{2\phi\tau'_\eta}{G_0(\eta)}\Big(T_{-2,1}(\eta) - \frac{\eta}{G_0(\eta)}T_{-3,2}(\eta)\Big) \overset{(*)}{=} \frac{2\phi\tau'_\eta}{G_0(\eta)}\Big\{\tau_\eta T_{-3,1}(\eta) + \Big(1-\frac{\eta}{G_0(\eta)}\Big)T_{-3,2}(\eta)\Big\}\\ &= 2\phi(\tau'_\eta)^2\big(T_{-3,1}(\eta) + \tau'_\eta T_{-2,1}(\eta)T_{-3,2}(\eta)\big).\end{aligned}$$

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 (153). 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 $\mathscr{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}$,

$$\partial_\eta\mathscr{R}^{\mathsf{in}}_{(\Sigma,\mu_0)}(\eta) = \sigma_\xi^2\cdot\partial_\eta\Big\{\phi^{-1}\eta^2\big(\phi\cdot\mathsf{SNR}_{\mu_0}\,\mathfrak{m}_\eta - (\eta\cdot\mathsf{SNR}_{\mu_0}-1)\mathfrak{m}'_\eta\big) + (\phi - 2\eta\mathfrak{m}_\eta)\Big\}$$

$$= \sigma_\xi^2 \cdot \big(2\mathfrak{m}_\eta - 4(\eta/\phi)\mathfrak{m}'_\eta + \phi^{-2}\eta^2\mathfrak{m}''_\eta\big) \cdot (\eta \cdot \mathsf{SNR}_{\mu_0} - 1). \tag{154}$$

The second term in the above display requires some non-trivial calculations:

$$\begin{aligned}
2\mathfrak{m}_\eta - \frac{4\eta}{\phi}\mathfrak{m}'_\eta + \frac{\eta^2}{\phi^2}\mathfrak{m}''_\eta &= \frac{1}{\tau_\eta}\left\{2 - 4\eta\frac{\tau'_\eta}{\tau_\eta} - \eta^2\frac{\tau''_\eta}{\tau_\eta} + 2\eta^2\left(\frac{\tau'_\eta}{\tau_\eta}\right)^2\right\} \\
&= \frac{1}{\tau_\eta}\left\{2 - \frac{4\eta}{G_0(\eta)} + \eta^2\frac{2\tau_\eta\tau'_\eta T_{-3,2}(\eta)}{G_0^2(\eta)} + \frac{2\eta^2}{G_0^2(\eta)}\right\} \\
&= \frac{2}{\tau_\eta G_0^2(\eta)}\big\{G_0^2(\eta) - 2\eta G_0(\eta) + \eta^2\tau_\eta\tau'_\eta T_{-3,2}(\eta) + \eta^2\big\}.
\end{aligned}$$

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 (154). Using ρ,

$$2\mathfrak{m}_\eta - \frac{4\eta}{\phi}\mathfrak{m}'_\eta + \frac{\eta^2}{\phi^2}\mathfrak{m}''_\eta = 2\int \frac{x^2}{(x+\eta/\phi)^3}\,\rho(\mathrm{d}x).$$

Finally, the claimed first two-sided bound on $\mathfrak{M}^{\#}$ follows from Proposition 23, and the second bound follows by using the fundamental theorem of calculus. ∎

F.4 A rigorous version of (19) and its proof

The theorem below presents a rigorous formulation of (19).

Theorem 54. *Suppose Assumptions A-B hold, and $\|\Sigma^{-1}\|_{\mathrm{op}} \vee \|\Sigma\|_{\mathrm{op}} \le K$ 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 $1/K \le \phi^{-1} \le K$ and $1/K \le \sigma_\xi^2 \le K$. Fix $\delta \in (0, 1/2]$ and $L \ge K/\delta^2$. 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 $\mathrm{vol}(\mathcal{U}_{\delta,\vartheta})/\mathrm{vol}(B_n(1) \setminus B_n(\delta)) \ge 1 - Ce^{-n^\vartheta/C}$, such that*

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

2. *(**Noiseless case**). Suppose $1 + 1/K \le \phi^{-1} \le K$ 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 $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \ge 1 - Ce^{-n^\vartheta/C}$, such that*

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

We need a few lemmas to prove Theorem 54.

The following lemma gives a technical extension of Theorem 49 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 55. *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 $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}, \mathsf{res}\}$,

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

Proof All the constants in $\lesssim, \gtrsim, \asymp$ below may possibly depend on K.
(**Part 1**). We shall first extend the claim of Theorem 49 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 $\sigma_\xi, \sigma_\xi' \in [0, K]$,

$$\|\widehat{\mu}_\eta(\sigma_\xi) - \widehat{\mu}_\eta(\sigma_\xi')\| \lesssim |\sigma_\xi - \sigma_\xi'| \cdot n^{-1} \|Z\|_{\mathrm{op}} \|\xi_0\| \cdot \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}. \tag{155}$$

Using the estimate (146), 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 56-(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 $\sigma_\xi \le \sigma_\xi'$,

$$\begin{aligned} &\mathbb{P}\Big(\sup_{\eta \in [0,K]} |R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma_\xi) - \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma_\xi)| \ge 2\varepsilon \Big) \\ &\le \mathbb{P}\Big(\sup_{\eta \in [0,K]} |R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma_\xi') - \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta, \sigma_\xi')| \ge \varepsilon \Big) + C_1 e^{-n/C_1}. \end{aligned} \tag{156}$$

The first term on the right hand side of the above display can be handled by the proven claim in Theorem 49, 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 49 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 49 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 (155), 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 (156) to conclude the extension of the claim of Theorem 49 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 (92) (with G replaced by Z) and (146), 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 (156) to extend the claim of Theorem 49 for $\# = \mathsf{in}$ to $\eta \in [0, K]$. The case for $\# = \mathsf{res}$ is similar so we omit repetitive details. ∎

Lemma 56. *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* $1+1/K \le \phi^{-1} \le K$ *and* $\|\Sigma\|_{\mathrm{op}} \vee \mathcal{H}_\Sigma \le K$ *for some* $K > 0$, *then there exists some* $C = C(K) > 0$ *such that* $|\gamma^2_{\eta,*}(\sigma_\xi) - \gamma^2_{\eta,*}(0)| \le C\sigma_\xi^2$.

Proof The claim (1) follows verbatim from the proof of Proposition 23-(1) by setting $\sigma_\xi^2 = 0$ therein. The claim (2) follows by using the formula (46). ■

Proof [Proof of Theorem 54] Let $\mathcal{U}_\vartheta \subset B_n(1)$ be as specified in Theorem 49 or 4. In view of its explicit form given in Proposition 40, with $\mathcal{U}_{\delta,\vartheta} \equiv \mathcal{U}_\vartheta \cap \big(B_n(1) \setminus B_n(\delta)\big)$, the volume estimates $\min\big\{\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)), \mathrm{vol}(\mathcal{U}_{\delta,\vartheta})/\mathrm{vol}(B_n(1) \setminus B_n(\delta))\big\} \ge 1 - Ce^{-n^\vartheta/C}$ hold.

On the other hand, using the construction around (149), 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 $\min\big\{\mathrm{vol}(\mathcal{V}_\varepsilon)/\mathrm{vol}(B_n(1)), \mathrm{vol}(\mathcal{V}_{\varepsilon,\delta})/\mathrm{vol}(B_n(1) \setminus B_n(\delta))\big\} \ge 1 - C\varepsilon^{-1}e^{-n\varepsilon^2/C}$, such that for $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$,

$$\sup_{\mu_0 \in \{\mathcal{V}_\varepsilon, \mathcal{V}_{\varepsilon,\delta}\}} \sup_{\eta \in \Xi_L} |\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)| \le \varepsilon. \tag{157}$$

Now let

$$\mathcal{W}_{\varepsilon,\vartheta} \equiv \mathcal{U}_\vartheta \cap \mathcal{V}_\varepsilon, \quad \mathcal{W}_{\varepsilon,\delta,\vartheta} \equiv \mathcal{U}_{\delta,\vartheta} \cap \mathcal{V}_{\varepsilon,\delta}. \tag{158}$$

Then we have the volume estimates $\min\big\{\mathrm{vol}(\mathcal{W}_{\varepsilon,\vartheta})/\mathrm{vol}(B_n(1)), \mathrm{vol}(\mathcal{W}_{\varepsilon,\delta,\vartheta})/\mathrm{vol}(B_n(1) \setminus B_n(\delta))\big\} \ge 1 - C\varepsilon^{-1}e^{-n\varepsilon^2/C} - Ce^{-n^\vartheta/C}$.

Moreover, by Proposition 53, provided $\eta_* \equiv \sigma_\xi^2/\|\mu_0\|^2 = \mathsf{SNR}_{\mu_0}^{-1} \in \Xi_L$,

$$\|\mu_0\|^2/C_0 \le \frac{|\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta_*)|}{(\eta - \eta_*)^2} \le C_0\|\mu_0\|^2 \tag{159}$$

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 (157) and (159), for any $\eta' \ge \eta_*$,

$$\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta') - \inf_{\eta \in \Xi_L} \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta) \ge \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta') - \inf_{\eta \in \Xi_L} \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - 2\varepsilon \ge \frac{\delta^2(\eta' - \eta_*)^2}{C_0} - 2\varepsilon.$$

Combined with a similar inequality for $\eta' \le \eta_*$, we conclude that for any $\mu_0 \in \mathcal{W}_{\varepsilon,\delta,\vartheta}$ and $\eta' \in \Xi_L$,

$$\big|\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta') - \inf_{\eta \in \Xi_L} \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)\big| \ge \frac{\delta^2(\eta' - \eta_*)^2}{C_0} - 2\varepsilon.$$

Now for $|\eta' - \eta_*| \ge \Delta$, choosing $\varepsilon \equiv \varepsilon_0 \equiv \delta^2\Delta^2/(4C_0)$, we have

$$\inf_{\mu_0 \in \mathcal{W}_{\varepsilon,\delta,\vartheta}} \inf_{\eta' \in \Xi_L : |\eta' - \eta_*| \ge \Delta} \big|\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta') - \inf_{\eta \in \Xi_L} \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)\big| \ge \frac{\delta^2\Delta^2}{2C_0}.$$

From here the claim follows from Theorem 49.

(**Noiseless case** $\sigma_\xi^2 = 0$). In this case, (159) implies that the map $\eta \mapsto \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ attains global minimum at $\eta = 0$. So together with (157), it implies that uniformly in $\mu_0 \in \mathcal{W}_{\varepsilon,\vartheta}$,

$$\big| \min_{\eta\in[0,K]} \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta) - \bar{R}^{\#}_{(\Sigma,\mu_0)}(0)\big| \le \varepsilon.$$

From here the claim follows from Lemma 55 that holds for $\sigma_\xi = 0$. ■

Appendix G. Proofs for Section 4

G.1 A rigorous version of (21) and its proof

Theorem 57. *Suppose Assumption A holds, and* $1/K \le \phi^{-1} \le K$, $\|\Sigma^{-1}\|_{\mathrm{op}} \vee \|\Sigma\|_{\mathrm{op}} \le K$ *hold for some* $K > 0$.

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

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

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* $1 + 1/K \le \phi^{-1} \le K$. *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* $\mathrm{vol}(\mathcal{U}_\vartheta)/\mathrm{vol}(B_n(1)) \ge 1 - Ce^{-n^\vartheta/C}$, *such that*

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

Proof [Proof of Theorem 57 for $\widehat{\tau}_\eta$] All the constants in $\lesssim, \gtrsim, \asymp$ may depend on K.

Let κ_0 be defined in the same way as in the proof of Proposition 40. Using a similar local law and continuity argument as in the proof of that proposition, on an event E_0 with $\mathbb{P}(E_0) \ge 1 - Cn^{-D}$,

$$\sup_{\eta\in\Xi_K} \big| m^{-1} \operatorname{tr}\big(\check{\Sigma} + (\eta/\phi)I\big)^{-1} - \mathfrak{m}\big(-\eta/\phi\big)\big| \lesssim \kappa_0^{-1} n^{-1/2+\varepsilon}.$$

So on $E_0 \cap \mathscr{E}(C_1)$, where $\mathscr{E}(C_1) \equiv \{\|Z\|_{\mathrm{op}}/\sqrt{n} \le C_1\}$ with $\mathbb{P}(\mathscr{E}(C_1)) \ge 1 - Ce^{-n/C}$, uniformly in $\eta \in \Xi_K$,

$$|\widehat{\tau}_\eta - \tau_{\eta,*}| \le \frac{\big|\frac{1}{m}\operatorname{tr}\big(\check{\Sigma} + \frac{\eta}{\phi}I\big)^{-1} - \mathfrak{m}\big(-\frac{\eta}{\phi}\big)\big|}{\frac{1}{m}\operatorname{tr}\big(\check{\Sigma} + \frac{\eta}{\phi}I\big)^{-1}\cdot \mathfrak{m}\big(-\frac{\eta}{\phi}\big)} \lesssim \Big\{ C_1^2 \mathbf{1}^{-1}_{\phi^{-1}\ge 1+1/K} \wedge \eta^{-1}\Big\} \cdot \kappa_0^{-1} n^{-1/2+\varepsilon}.$$

Here in the last inequality, we use the following estimate for $\mathfrak{m}(z)$: As $\mathfrak{m}$ is the Stieltjes transform of ρ (cf. (Knowles and Yin, 2017, Lemma 2.2)), $\mathfrak{m}(z) \ge 0$ for $z \le 0$, 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 [Proof of Theorem 57 for $\widehat{\gamma}_\eta$] All the constants in $\lesssim, \gtrsim, \asymp$ may depend on K.

Using Theorem 49, the stability of $\tau_{\eta,*}$ in Proposition 23, and the proven fact in (1) on $\widehat{\tau}_\eta$, it holds for $\varepsilon \in (0, 1/2]$ that

$$\mathbb{P}\Big(\sup_{\eta\in[1/K,K]} \big|\eta^{-1}\widehat{\tau}_\eta \|\widehat{r}_\eta(\sigma_\xi)\| - \gamma_{\eta,*}(\sigma_\xi)\big| \ge \varepsilon \Big) \le C_1 \varepsilon^{-c_0} n^{-1/6.5}. \tag{160}$$

Next we consider extension to $\eta \in [0, K]$ in the regime $\phi^{-1} \ge 1 + 1/K$. 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 (92) (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 $\phi^{-1} \ge 1 + 1/K$ 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 (155),

$$\begin{aligned}
&\sup_{\eta\in[0,K]} \big|\sqrt{n}\widehat{\tau}_\eta\big\|(XX^\top)^{-1}X\widehat{\mu}_\eta(\sigma_\xi)\big\| - \sqrt{n}\widehat{\tau}_\eta\big\|(XX^\top)^{-1}X\widehat{\mu}_\eta(\sigma'_\xi)\big\|\big| \\
&\lesssim \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}^2 \cdot (\|Z\|_{\mathrm{op}}/\sqrt{n}) \cdot \sup_{\eta\in[0,K]} \|\widehat{\mu}_\eta(\sigma_\xi) - \widehat{\mu}_\eta(\sigma'_\xi)\| \\
&\lesssim \Big(1 + \frac{\|Z\|_{\mathrm{op}} + \|\xi_0\|}{\sqrt{n}} + \|(ZZ^\top/n)^{-1}\|_{\mathrm{op}}\Big)^{c_0} \cdot |\sigma_\xi - \sigma'_\xi|.
\end{aligned}$$

The claimed bound follows. ■

G.2 Proof of Theorem 10

Recall we have $\gamma_{\eta,*}^2 = \phi^{-1}\big(\sigma_\xi^2 + \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big)$. For both the case $\sigma_\xi^2 \in [1/K, K]$ and $\sigma_\xi^2 \in [0, K]$ with $\phi^{-1} \ge 1 + 1/K$, we take $\mathcal{W}_{\varepsilon,\delta,\vartheta} \subset B_n(1) \setminus B_n(\delta)$ as constructed in (158) in the proof of Theorem 54, 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 8 and 57, on an event E_0 with $\mathbb{P}(E_0^c) \le Cn^{-1/7}$,

$$\sup_{\eta\in\Xi_L} \big|\widehat{\gamma}_\eta^2 - \phi^{-1}\big(\sigma_\xi^2 + \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big)\big| \le \varepsilon. \tag{161}$$

This in particular implies that on E_0, both the following inequalities hold:

$$\phi\widehat{\gamma}^2_{\widehat{\eta}^{\mathsf{GCV}}} - \sigma_\xi^2 - \phi\varepsilon \le \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) \le \phi\widehat{\gamma}^2_{\widehat{\eta}^{\mathsf{GCV}}} - \sigma_\xi^2 + \phi\varepsilon,$$

$$\phi \min_{\eta\in\Xi_L} \widehat{\gamma}_\eta^2 - \sigma_\xi^2 - \phi\varepsilon \le \min_{\eta\in\Xi_L} \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) \le \phi \min_{\eta\in\Xi_L} \widehat{\gamma}_\eta^2 - \sigma_\xi^2 + \phi\varepsilon. \tag{162}$$

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,

$$\big|\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \min_{\eta\in\Xi_L} \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big| \le 2\phi\varepsilon. \tag{163}$$

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 53, we have on the event E_0,

$$|\widehat{\eta}^{\mathsf{GCV}} - \eta_*| \le \frac{C}{\|\mu_0\|}\big|\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta_*)\big|^{1/2} \le C_1\varepsilon^{1/2}. \tag{164}$$

This means on E_0, for both $\# \in \{\mathsf{est}, \mathsf{in}\}$,

$$\big|\mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \min_{\eta\in\Xi_L} \mathscr{R}^{\#}_{(\Sigma,\mu_0)}(\eta)\big| \le C_2\varepsilon.$$

We may conclude from here by virtues of Theorems 49 and 8, together with Lemma 55. □

G.3 Proof of Theorem 12

Lemma 58. *Consider the following version of (13) with sample size $m - m_\ell$:*

$$\begin{cases} \frac{m-m_\ell}{n} \cdot \gamma^2 = \sigma_\xi^2 + \mathbb{E}\,\mathsf{err}_{(\Sigma,\mu_0)}(\gamma;\tau), \\ \big(\frac{m-m_\ell}{n} - \frac{\eta}{\tau}\big) \cdot \gamma^2 = \mathbb{E}\,\mathsf{dof}_{(\Sigma,\mu_0)}(\gamma;\tau). \end{cases} \tag{165}$$

1. *The fixed point equation (165) admits a unique solution $(\gamma^{(\ell)}_{\eta,*}, \tau^{(\ell)}_{\eta,*}) \in (0,\infty)^2$, 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 $\eta \in \Xi_K$, $1/C \le \gamma^{(\ell)}_{\eta,*}, \tau^{(\ell)}_{\eta,*} \le C$. Moreover,*

$$|\gamma^{(\ell)}_{\eta,*} - \gamma_{\eta,*}| \vee |\tau^{(\ell)}_{\eta,*} - \tau_{\eta,*}| \le \frac{Cm_\ell}{n}.$$

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 23-(3). Instead of considering (165), we shall consider the system of equations

$$\begin{cases} \phi - \alpha = \frac{1}{\gamma^2}\big(\sigma_\xi^2 + \tau^2\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\mu_0\|^2\big) + \frac{1}{n}\,\mathrm{tr}\,\big((\Sigma+\tau I)^{-2}\Sigma^2\big), \\ \phi - \alpha = \frac{1}{n}\,\mathrm{tr}\,\big((\Sigma+\tau I)^{-1}\Sigma\big) + \frac{\eta}{\tau}, \end{cases} \tag{166}$$

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 23-(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 (166) with respect to α, we obtain

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

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 (166), we obtain

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

Using similar calculations as in (52)-(53), 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 [Proof of Theorem 12] 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 61 first conditionally on $(Z^{(-\ell)},\xi^{(-\ell)})$ and then further taking expectation over $(Z^{(-\ell)},\xi^{(-\ell)})$, we have for $0<\varrho\le 1$,

$$\begin{aligned}\mathbb{P}\Big(E_{0,\ell}^c(\eta)\equiv\Big\{&\big|m_\ell^{-1}\|Y^{(\ell)}-X^{(\ell)}\widehat{\mu}_\eta^{(\ell)}\|^2-\big(\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)}-\mu_0)\|^2+\sigma_\xi^2\big)\big|\\ &\ge C_0\big(\sigma_\xi^2\vee\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)}-\mu_0)\|^2\big)m_\ell^{-(1-\varrho)/2}\Big\}\Big)\le C_0e^{-m_\ell^\varrho/C_0}.\end{aligned}$$

Here $C_0>0$ is a universal constant. Using similar arguments as in (147) (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 $\mathbb{P}(E_{1,\ell}^c)\le C_1e^{-m_\ell/C_1}$, $\sup_{\eta\in\Xi_L}\|\widehat{\mu}_\eta^{(\ell)}\|\le C_1$. This means that for any $\eta\in\Xi_L$, on the event $\cap_{\ell\in[k]}(E_{0,\ell}(\eta)\cap E_{1,\ell})$,

$$\max_{\ell\in[k]} m_\ell^{(1-\varrho)/2}\cdot\big|m_\ell^{-1}\|Y^{(\ell)}-X^{(\ell)}\widehat{\mu}_\eta^{(\ell)}\|^2-\big(\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)}-\mu_0)\|^2+\sigma_\xi^2\big)\big|\le C_1'. \tag{167}$$

On the other hand, using Theorem 49, we may find some $\mathcal{U}_{\vartheta;\ell}\subset B_n(1)$ with $\mathrm{vol}(\mathcal{U}_{\vartheta;\ell})/\mathrm{vol}(B_n(1))\ge 1-C_2e^{-n^\vartheta/C_2}$, such that for $\varepsilon\in(0,1/2]$, on an event $E_{2,\ell}(\varepsilon)$ with $\mathbb{P}(E_{2,\ell}^c(\varepsilon))\le C_2(ne^{-n\varepsilon^4/C_2}+\varepsilon^{-c_0}n^{-1/6.5}\mathbf{1}_{Z\neq G})$, for $\mu_0\in\mathcal{U}_{\vartheta;\ell}$,

$$\sup_{\eta\in\Xi_L}\left|\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)}-\mu_0)\|^2-\left\{\frac{m-m_\ell}{n}(\gamma_{\eta,*}^{(\ell)})^2-\sigma_\xi^2\right\}\right|\le\varepsilon.$$

Here $\gamma_{\eta,*}^{(\ell)}$ is taken from Lemma 58, 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 58, on the event $E_{2,\ell}(\varepsilon)$, we then have

$$\sup_{\eta\in\Xi_L}\big|\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)}-\mu_0)\|^2-\big\{\phi\gamma_{\eta,*}^2-\sigma_\xi^2\big\}\big|\le\varepsilon+\frac{C_2m_\ell}{n}.$$

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 $\cap_{\ell\in[0:k]} E_{2,\ell}(\varepsilon)$,

$$\sup_{\eta\in\Xi_L} \big|\|\Sigma^{1/2}(\widehat{\mu}_\eta^{(\ell)} - \mu_0)\|^2 - R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big| \le 2\varepsilon + \frac{C_2 m_\ell}{n}. \tag{168}$$

Combining (167)-(168), for $\mu_0 \in \mathcal{U}_\vartheta \equiv \cap_{\ell\in[0:k]}\mathcal{U}_{\vartheta;\ell}$, $\varepsilon \in (0, 1/2]$ and $\eta \in \Xi_L$,

$$\begin{aligned}
&\mathbb{P}\left(\big|R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta) - \big(R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) + \sigma_\xi^2\big)\big| \ge C_2' \cdot \left\{\frac{1}{k}\sum_{\ell\in[k]} \frac{1}{m_\ell^{(1-\varrho)/2}} + \frac{1}{k} + \varepsilon\right\}\right) \\
&\le C_2' \cdot \begin{cases} \sum_{\ell\in[k]} e^{-m_\ell^\varrho/C_0} + kne^{-n\varepsilon^4/C_2}, & Z = G; \\ \sum_{\ell\in[k]} e^{-m_\ell^\varrho/C_0} + \varepsilon^{-c_0}\cdot kn^{-1/6.5}, & \text{otherwise.} \end{cases}
\end{aligned} \tag{169}$$

Now we strengthen the estimate (169) into a uniform version. It is easy to verify that on an event $E_{3,\ell}$ with $\mathbb{P}(E_{3,\ell}^c) \le C_3 e^{-m_\ell/C_3}$, $\|Z^{(\ell)}\|_{\mathrm{op}} \le C_3(\sqrt{m_\ell} + \sqrt{n})$, $\|\xi^{(\ell)}\| \le C_3\sqrt{m_\ell}$, and for $\eta_1, \eta_2 \in \Xi_L$, $\|\widehat{\mu}_{\eta_1}^{(\ell)} - \widehat{\mu}_{\eta_2}^{(\ell)}\| \le C_3|\eta_1 - \eta_2|$. So on $\cap_{\ell\in[k]}(E_{1,\ell} \cap E_{3,\ell})$, for $\eta_1, \eta_2 \in \Xi_L$,

$$\begin{aligned}
\big|R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta_1) - R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta_2)\big| &\lesssim \frac{1}{k}\sum_{\ell\in[k]} \frac{1}{m_\ell}\big|\|Z^{(\ell)}\|_{\mathrm{op}}\|\widehat{\mu}_{\eta_1}^{(\ell)} - \widehat{\mu}_{\eta_2}^{(\ell)}\| \cdot \big(\|Z^{(\ell)}\|_{\mathrm{op}} + \|\xi^{(\ell)}\|\big)\big| \\
&\lesssim \frac{1}{k}\sum_{\ell\in[k]} \frac{m_\ell + n}{m_\ell}\cdot|\eta_1 - \eta_2| \le C_3' \cdot \frac{1}{k}\sum_{\ell\in[k]} \frac{n}{m_\ell}\cdot|\eta_1 - \eta_2|,
\end{aligned}$$

and

$$\big|R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta_1) - R^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta_2)\big| \le C_3'|\eta_1 - \eta_2|.$$

From here, using (i) (169) along with a discretization and union bound that strengthens (169) to a uniform control, and (ii) Theorem 49 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 $\varepsilon \in (0, 1/2]$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\eta\in\Xi_L}\big|R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta) - \big(\bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) + \sigma_\xi^2\big)\big| \ge C_3'' \cdot \left\{\frac{1}{k}\sum_{\ell\in[k]} \frac{1}{m_\ell^{(1-\varrho)/2}} + \frac{1}{k} + \varepsilon\right\} \equiv \varepsilon_{\{m_\ell\}}\right) \\
&\le \mathfrak{p}_0 \equiv \frac{C_3''}{\varepsilon k}\sum_{\ell\in[k]} \frac{n}{m_\ell}\cdot\begin{cases} \sum_{\ell\in[k]} e^{-m_\ell^\varrho/C_0} + kne^{-n\varepsilon^4/C_2}, & Z = G; \\ \sum_{\ell\in[k]} e^{-m_\ell^\varrho/C_0} + \varepsilon^{-c_0}\cdot kn^{-1/6.5}, & \text{otherwise.} \end{cases}
\end{aligned}$$

Now with the same $\mathcal{W}_{\varepsilon,\delta,\vartheta}$ as in (158) 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 $\mathbb{P}(E_4^c) \le \mathfrak{p}_0$,

$$\sup_{\eta\in\Xi_L}\big|R^{\mathsf{CV},k}_{(\Sigma,\mu_0)}(\eta) - \big(\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) + \sigma_\xi^2\big)\big| \le \varepsilon_{\{m_\ell\}}.$$

From here, using similar arguments as in (162)-(163), on the event E_4,

$$\big|\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{CV}}) - \min_{\eta\in\Xi_L}\mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta)\big| \le 2\varepsilon_{\{m_\ell\}}.$$

Similar to (164), on the event E_4, we have

$$|\widehat{\eta}^{\mathsf{CV}} - \eta_*| \le C_4 \cdot \varepsilon_{\{m_\ell\}}^{1/2}. \tag{170}$$

From here we may argue along the same lines as those following (164) in the proof of Theorem 10 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 $\min_{\ell\in[k]} m_\ell \ge \log^{2/\delta} m$, by taking $\varrho = \delta$, $\mathfrak{p}_0$ simplifies as indicated in the statement of the theorem for n large. ■

G.4 Proof of Theorem 13

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 (158) 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 57-(2), on an event E_0 with the probability indicated therein,

$$\max_{j\in[n]} \sup_{\eta\in\Xi_L} \left| |\mathrm{CI}_j(\eta)| - 2\gamma_{\eta,*} (\Sigma^{-1})_{jj}^{1/2} \frac{z_{\alpha/2}}{\sqrt{n}} \right| \le \frac{2\|\Sigma^{-1}\|_{\mathrm{op}}^{1/2} z_{\alpha/2}}{\sqrt{n}} \sup_{\eta\in\Xi_L} |\widehat{\gamma}_\eta - \gamma_{\eta,*}| \lesssim \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \varepsilon.$$

Consequently, on the event E_0, for any $\mu_0 \in \mathcal{W}_{\varepsilon,\delta,\vartheta}$,

$$\begin{aligned}
&\sqrt{n} z_{\alpha/2}^{-1} \cdot \max_{j\in[n]} \big| |\mathrm{CI}_j(\widehat{\eta}^{\mathsf{GCV}})| - \min_{\eta\in\Xi_L} |\mathrm{CI}_j(\eta)| \big| \\
&\lesssim |\gamma_{\widehat{\eta}^{\mathsf{GCV}},*} - \min_{\eta\in\Xi_L} \gamma_{\eta,*}| + \varepsilon \\
&\lesssim |\gamma^2_{\widehat{\eta}^{\mathsf{GCV}},*} - \min_{\eta\in\Xi_L} \gamma^2_{\eta,*}| + \varepsilon \quad \text{(using Proposition 23-(3))} \\
&\asymp \big| \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \min_{\eta\in\Xi_L} \bar{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) \big| + \varepsilon \quad \text{(using definition of } \gamma^2_{\eta,*}) \\
&\lesssim \big| \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\widehat{\eta}^{\mathsf{GCV}}) - \min_{\eta\in\Xi_L} \mathscr{R}^{\mathsf{pred}}_{(\Sigma,\mu_0)}(\eta) \big| + \varepsilon \quad \text{(using Theorem 8).}
\end{aligned}$$

As in the proof of Theorem 10, for $\sigma_\xi^2 \le K$, $\eta_* = \mathsf{SNR}_{\mu_0}^{-1} \in \Xi_L$, so by using Proposition 53-(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 (164), by adjusting constants, on an event E_1 with $\mathbb{P}(E_1^c) \le C_1 n^{-1/7}$, it holds that

$$\sqrt{n} z_{\alpha/2}^{-1} \cdot \max_{j\in[n]} \big| |\mathrm{CI}_j(\widehat{\eta}^{\mathsf{GCV}})| - \min_{\eta\in\Xi_L} |\mathrm{CI}_j(\eta)| \big| \vee |\gamma_{\widehat{\eta}^{\mathsf{GCV}},*} - \gamma_{\eta_*,*}| \vee |\widehat{\eta}^{\mathsf{GCV}} - \eta_*|^2 \le \varepsilon. \tag{171}$$

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 55 shows that for any 1-Lipschitz $\mathsf{g}:\mathbb{R}^n\to\mathbb{R}$, on an event $E_2(\mathsf{g})$ with $\mathbb{P}(E_2(\mathsf{g})^c)\le Cn^{-1/7}$,

$$\sup_{\eta\in\Xi_L}\big|\mathsf{g}(\widehat{\mu}_\eta^{\mathsf{dR}})-\mathbb{E}\,\mathsf{g}\big(\mu_0+\gamma_{\eta,*}\Sigma^{-1/2}g/\sqrt{n}\big)\big|\le\varepsilon. \tag{172}$$

On the other hand, using the Lipschitz continuity of $\eta\mapsto\tau_{\eta,*}$ in Proposition 23-(3),

$$\begin{aligned}\big|\mathsf{g}(\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}}}^{\mathsf{dR}})-\mathsf{g}(\widehat{\mu}_{\eta_*}^{\mathsf{dR}})\big|&\le\|\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}}}^{\mathsf{dR}}-\widehat{\mu}_{\eta_*}^{\mathsf{dR}}\|\lesssim|\tau_{\widehat{\eta}^{\mathsf{GCV}},*}-\tau_{\eta_*,*}|\sup_{\eta\in\Xi_L}\|\widehat{\mu}_\eta\|+\|\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}}}-\widehat{\mu}_{\eta_*}\|\\&\lesssim\|\widehat{\eta}^{\mathsf{GCV}}-\eta_*\|\sup_{\eta\in\Xi_L}\|\widehat{\mu}_\eta\|+\|\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}}}-\widehat{\mu}_{\eta_*}\|.\end{aligned}$$

So by enlarging C_1 if necessary, we may assume without loss of generality that on $E_1\cap E_2(\mathsf{g})$,

$$\big|\mathsf{g}(\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}}}^{\mathsf{dR}})-\mathsf{g}(\widehat{\mu}_{\eta_*}^{\mathsf{dR}})\big|\le C_1\varepsilon^{1/2}. \tag{173}$$

Now we shall make a good choice of g in (171). 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

$$\mathsf{g}(u)\equiv\frac{\Delta}{n}\sum_{j=1}^n\mathsf{g}_{0,\Delta}\bigg(\frac{u_j-\mu_{0,j}}{(\gamma_{\eta_*,*}+\varepsilon)(\Sigma^{-1})_{jj}^{1/2}z_{\alpha/2}/\sqrt{n}}\bigg). \tag{174}$$

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 (172) with g defined in (174) to obtain that on the event $E_1\cap E_2(\mathsf{g})$,

$$\begin{aligned}\mathscr{C}^{\mathsf{dR}}(\widehat{\eta}^{\mathsf{GCV}})&=\frac{1}{n}\sum_{j=1}^n\mathbf{1}\Big(\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}},j}^{\mathsf{dR}}\in\Big[\mu_{0,j}\pm\widehat{\gamma}_{\widehat{\eta}^{\mathsf{GCV}}}(\Sigma^{-1})_{jj}^{1/2}\frac{z_{\alpha/2}}{\sqrt{n}}\Big]\Big)\\&\le\frac{1}{n}\sum_{j=1}^n\mathbf{1}\Big(\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}},j}^{\mathsf{dR}}\in\Big[\mu_{0,j}\pm(\gamma_{\eta_*,*}+\varepsilon)(\Sigma^{-1})_{jj}^{1/2}\frac{z_{\alpha/2}}{\sqrt{n}}\Big]\Big)\\&\le\Delta^{-1}\cdot\mathsf{g}\big(\widehat{\mu}_{\widehat{\eta}^{\mathsf{GCV}}}^{\mathsf{dR}}\big)\quad(\text{using }\mathbf{1}_{[-1,1]}\le\mathsf{g}_{0,\Delta})\\&\le\Delta^{-1}\cdot\mathsf{g}\big(\widehat{\mu}_{\eta_*}^{\mathsf{dR}}\big)+\mathcal{O}(\varepsilon^{1/2}/\Delta)\quad(\text{by (173)})\\&\le\Delta^{-1}\cdot\mathbb{E}\,\mathsf{g}\big(\mu_0+\gamma_{\eta_*,*}\Sigma^{-1/2}g/\sqrt{n}\big)+\mathcal{O}(\varepsilon^{1/2}/\Delta).\end{aligned} \tag{175}$$

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

$$\begin{aligned}&\Delta^{-1}\cdot\mathbb{E}\,\mathsf{g}\big(\mu_0+\gamma_{\eta_*,*}\Sigma^{-1/2}g/\sqrt{n}\big)=\mathbb{E}\,\mathsf{g}_{0,\Delta}\bigg(\frac{\gamma_{\eta_*,*}}{\gamma_{\eta_*,*}+\varepsilon}\cdot\frac{g}{z_{\alpha/2}}\bigg)\\&\le\mathbb{P}\Big(\mathcal{N}(0,1)\in\Big[\pm z_{\alpha/2}\cdot\big(1+\varepsilon/\gamma_{\eta_*,*}\big)\cdot(1+\Delta)\Big]\Big)\le1-\alpha+\mathcal{O}(\varepsilon+\Delta).\end{aligned} \tag{176}$$

Combining the above two displays (175)-(176), on the event $E_1\cap E_2(\mathsf{g})$,

$$\mathscr{C}^{\mathsf{dR}}(\widehat{\eta}^{\mathsf{GCV}})\le1-\alpha+\mathcal{O}(\varepsilon+\Delta+\varepsilon^{1/2}/\Delta).$$

Finally choosing $\Delta=\varepsilon^{1/4}$ to conclude the upper control. The lower control can be proved similarly so we omit the details. □

Appendix H. Auxiliary results

Proposition 59. *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 $\|\nabla G(g)\|^2 = \lambda^2 \|\nabla H(g)\|^2 \leq \lambda^2\Gamma^2 \cdot H(g) = \lambda^2\Gamma^2 \cdot \big(Z + \mathbb{E}\, H(g)\big)$. By the Gaussian log-Sobolev inequality (see e.g., (Boucheron et al., 2013, Theorem 5.4), or (Giné and Nickl, 2016, Theorem 2.5.6)),

$$\mathrm{Ent}(e^{\lambda Z}) = \mathbb{E}[\lambda Z e^{\lambda Z}] - \mathbb{E}\, e^{\lambda Z} \log \mathbb{E}\, e^{\lambda Z} \leq \frac{1}{2}\mathbb{E}\left[\lambda^2\Gamma^2\big(Z + \mathbb{E}\, H(g)\big)e^{\lambda Z}\right].$$

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 $\big(\log m_Z(\lambda)/\lambda\big)' \leq \frac{\Gamma^2}{2}\big(\log m_Z(\lambda) + \lambda\, \mathbb{E}\, H(g)\big)'$. 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 $\log m_Z(\lambda) \leq \frac{\Gamma^2}{2}\big(\lambda \log m_Z(\lambda) + \lambda^2\, \mathbb{E}\, H(g)\big)$. Solving for $\log m_Z(\lambda)$ and using the standard method to convert to tail bound yield the claimed inequality. ■

Lemma 60. *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 $q \in [1, \infty)$, there exists some $C = C(K, q) > 0$ such that*

$$\left|\frac{\mathbb{E}\|\mathcal{N}(0, \Sigma)\|_q}{\|\mathrm{diag}(\Sigma)\|_{q/2}^{1/2} M_q} - 1\right| \leq Cn^{-\frac{1}{q\vee 2}}\sqrt{\log n}.$$

where $M_q \equiv \mathbb{E}^{1/q}|\mathcal{N}(0,1)|^q = 2^{1/2}\big\{\Gamma\big((q+1)/2\big)/\sqrt{\pi}\big\}^{1/q}$.

Proof Let $g \sim \mathcal{N}(0, I_n)$. We first prove that for some $C_0 > 1$,

$$n^{\frac{1}{q\vee 2}}/C_0 \leq \mathbb{E}\|\Sigma^{1/2} g\|_q \leq C_0 n^{\frac{1}{q}}. \tag{177}$$

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 (177).

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 (177),

$$\begin{aligned}\mathbb{E}\|\Sigma^{1/2}g\|_q^q &\le \mathbb{E}\|\Sigma^{1/2}g\|_q^q \mathbf{1}_{E(t)} + \mathbb{E}^{1/2}\|\Sigma^{1/2}g\|_q^{2q}\cdot \mathbb{P}^{1/2}(E(t)^c)\\ &\le \big(\mathbb{E}\|\Sigma^{1/2}g\|_q + C\sqrt{t}\big)^q + C\cdot n^{1/q}\,\mathbb{P}^{1/2}(E(t)^c)\\ &\le \big(\mathbb{E}\|\Sigma^{1/2}g\|_q\big)^q\cdot\Big\{\big(1+Cn^{-\frac{1}{q\vee 2}}\sqrt{t}\big)^q + C\cdot n^{\frac{1}{q}-\frac{1}{q\vee 2}}\,\mathbb{P}^{1/2}(E(t)^c)\Big\}.\end{aligned}$$

By choosing $t = C_1\log n$ for some sufficiently large $C_1 > 0$, we have

$$\frac{\mathbb{E}\|\mathcal{N}(0,\Sigma)\|_q}{\|\mathrm{diag}(\Sigma)\|_{q/2}^{1/2}M_q} = \frac{\mathbb{E}\|\Sigma^{1/2}g\|_q}{\mathbb{E}^{1/q}\|\Sigma^{1/2}g\|_q^q} \ge \big(1 - Cn^{-\frac{1}{q\vee 2}}\sqrt{\log n}\big)_+.$$

The upper bound follows similarly. ■

Lemma 61. *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 $0 < \varrho \le 1$, with probability at least $1 - Ce^{-m^\varrho/C}$,*

$$\big|m^{-1}\|Zb+\xi\|^2 - \big(\|b\|^2+\sigma_\xi^2\big)\big| \le C\cdot(\sigma_\xi^2\vee\|b\|^2)\cdot m^{-(1-\varrho)/2}.$$

Proof Let $Z_1,\dots,Z_m\in\mathbb{R}^n$ 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 $1 - Ce^{-m^\varrho/C}$,

- $\big|\|b\|^2\frac{1}{m}\sum_{i=1}^m\langle Z_i,\frac{b}{\|b\|}\rangle^2 - \|b\|^2\big| \le C\|b\|^2\cdot m^{-(1-\varrho)/2}$,
- $\big|\frac{2\sigma_\xi\|b\|}{m}\sum_{i=1}^n\frac{\xi_i}{\sigma_\xi}\langle Z_i,\frac{b}{\|b\|}\rangle\big| \le C\sigma_\xi\|b\|\cdot m^{-(1-\varrho)/2}$,
- $\big|\sigma_\xi^2\frac{\|\xi/\sigma_\xi\|^2}{m} - \sigma_\xi^2\big| \le C\sigma_\xi^2\cdot m^{-(1-\varrho)/2}$.

Collecting the bounds to conclude. ■

Appendix I. Simulation details for Figure 1 and additional simulations

I.1 Common numerical settings

We set $\Sigma = 1.99\cdot I_n + 0.01\cdot 1_n 1_n^\top$, with 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.

I.2 Simulation details for Figure 1

We investigate the efficacy of two cross validation schemes in Section 4, namely $\widehat{\eta}^{\mathsf{GCV}}$ in (22) and $\widehat{\eta}^{\mathsf{CV}}$ in (24). 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 10 and 12, we report in the left panel of Figure 1 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 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 10 and 12 and empirically confirmed here.

To empirically verify Theorem 13, we report in the middle and right panels of Figure 1 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 $\eta_* = \mathsf{SNR}^{-1}_{\mu_0}$. 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.

I.3 Validation of (19)

We next verify the optimal oracle regularization rule in (19) (see Theorem 54 for a rigorous formulation) by simulation. We use $m = 100$, $n = 200$, and a unit vector μ_0 chosen randomly (and then fixed) from the sphere $\partial B_n(1)$. For this setting, we plot both the theoretical risk curve $\eta \mapsto \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ and the empirical risk curve $\eta \mapsto R^{\#}_{(\Sigma,\mu_0)}(\eta)$ for all $\# \in \{\mathsf{pred}, \mathsf{est}, \mathsf{in}\}$. The left panel of Figure 2 reports the noisy case with noise level $\sigma_\xi^2 = 1$ and $\mathsf{SNR}^{-1}_{\mu_0} = 1$, while the middle panel reports the noiseless case $\sigma_\xi^2 = 0$ with $\mathsf{SNR}^{-1}_{\mu_0} = 0$. These plots show excellent agreement with (19) in that the global minimum of both the theoretical and empirical risk curves is attained roughly at $\eta_* = \mathsf{SNR}^{-1}_{\mu_0}$.

In order to demonstrate the validity of the above phenomenon for 'most' μ_0's, as claimed in Theorem 54, we uniformly generate 500 different μ_0's over $\partial B_n(1)$. For each μ_0, we discretize $\eta \in [0, 1.5]$ into 160 grid points and select the empirical optimal value $\eta^{\#}$ by minimizing the empirical prediction, estimation, and in-sample risks. The difference between the empirical optimal $\eta^{\#}$ and the theoretically optimal tuning η_* is depicted in the right panel of Figure 2 through a boxplot of $\eta^{\#} - \eta_*$. It is easily seen that, for all three risks, these differences are highly concentrated around 0.

We finally explain how the theoretical risk curves $\eta \mapsto \bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ are computed in practice. For each fixed η, we solve the fixed-point system (13) for $(\gamma_{\eta,*}, \tau_{\eta,*})$ as follows. The second equation in (13) involves only the scalar variable τ; under our assumptions, its

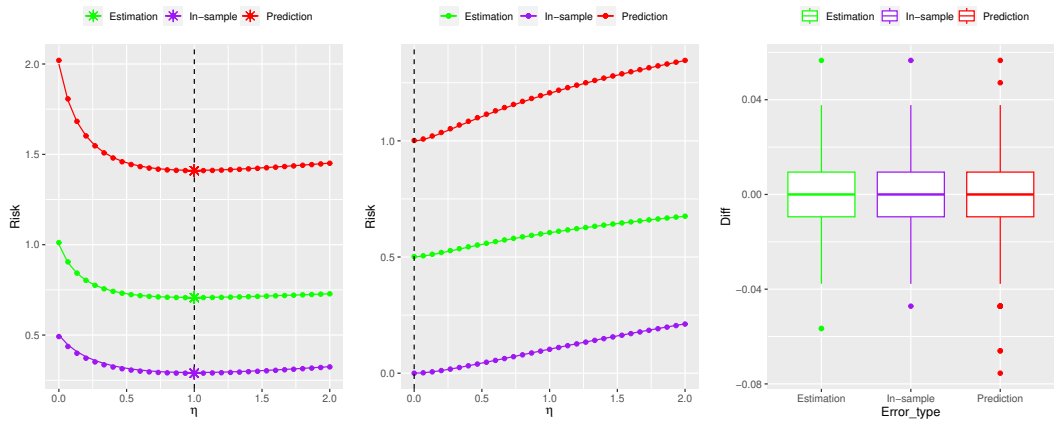

Figure 2: Validation of (19) (see also Theorem 54 for a rigorous formulation). The theoretical risks $\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$ are computed by solving (13), and the empirical risks $R^{\#}_{(\Sigma,\mu_0)}(\eta)$ are computed via Monte Carlo simulation over 200 repetitions. *Left panel*: noisy case with minimal empirical risks attained at $\eta_* = \mathsf{SNR}_{\mu_0}^{-1} = 1$ (marked with $*$). *Middle panel*: noiseless case with all risks minimized at the interpolation regime $\eta_* = \mathsf{SNR}_{\mu_0}^{-1} = 0$. *Right panel*: differences between the global minimizer of the empirical risk curves and the oracle η_* are concentrated around 0 over 500 different μ_0's.

right-hand side is monotone in τ, so the solution $\tau_{\eta,*}$ is unique. We therefore solve this one-dimensional fixed-point equation for $\tau_{\eta,*}$ by a standard bisection method on a prescribed interval, up to a given numerical tolerance. Once $\tau_{\eta,*}$ is obtained, we plug it into the first equation in (13) to compute $\gamma_{\eta,*}$. The resulting pair $(\gamma_{\eta,*}, \tau_{\eta,*})$ is then substituted into the closed-form expressions for $\bar{R}^{\#}_{(\Sigma,\mu_0)}(\eta)$.