Provable Benefits of Overparameterization in Model Compression: From Double Descent to Pruning Neural Networks

Xiangyu Chang[1] **Yingcong Li**[1] **Samet Oymak**[1] **Christos Thrampoulidis**[2*]

[1] University of California, Riverside
[2] University of British Columbia, Vancouver
{cxian008, yli692, soymak}@ucr.edu, cthrampo@ece.ubc.ca

Abstract

Deep networks are typically trained with many more parameters than the size of the training dataset. Recent empirical evidence indicates that the practice of overparameterization not only benefits training large models, but also assists – perhaps counterintuitively – building lightweight models. Specifically, it suggests that overparameterization benefits model pruning / sparsification. This paper sheds light on these empirical findings by theoretically characterizing the high-dimensional asymptotics of model pruning in the overparameterized regime. The theory presented addresses the following core question: "should one train a small model from the beginning, or first train a large model and then prune?". We analytically identify regimes in which, even if the location of the most informative features is known, we are better off fitting a large model and then pruning rather than simply training with the known informative features. This leads to a new double descent in the training of sparse models: growing the original model, while preserving the target sparsity, improves the test accuracy as one moves beyond the overparameterization threshold. Our analysis further reveals the benefit of retraining by relating it to feature correlations. We find that the above phenomena are already present in linear and random-features models. Our technical approach advances the toolset of high-dimensional analysis and precisely characterizes the asymptotic distribution of over-parameterized least-squares. The intuition gained by analytically studying simpler models is numerically verified on neural networks.

1 Introduction

Large model size and overparameterization in deep learning are known to improve generalization performance (Neyshabur et al. 2017), and, state-of-the-art deep neural networks (DNNs) can be outrageously large. However, such large models are not suitable for certain important application domains, such as mobile computing (Tan et al. 2019; Sandler et al. 2018). Pruning algorithms aim to address the challenge of building lightweight DNNs for such domains. While there are several pruning methods, their common goal is to compress large DNN models by removing weak connections/weights with minimal decline in accuracy. Here, a key empirical phenomenon is that *it is often better to train and prune a large model rather than training a small model from scratch*. Unfortunately, the mechanisms behind this phenomenon are poorly understood especially for practical gradient-based algorithms. This paper sheds light on this by answering: *What are the optimization and generalization dynamics of pruning overparameterized models? Does gradient descent naturally select the good weights?*

Contributions: We analytically study the performance of popular pruning strategies. First, we analyze linear models, and then, generalize the results to nonlinear feature maps. Through extensive simulations, we show that our analytical findings predict similar behaviors in more complex settings.
(a) Distributional characterization (DC): The key innovation facilitating our results is a theoretical characterization of the distribution of the solution of overparameterized least-squares. This DC enables us to accurately answer *"what happens to the accuracy if X% of the weights are pruned?"*.
(b) Benefits of overparameterization: Using DC, we obtain rigorous precise characterizations of the pruning performance in linear problems. Furthermore, we use, so called "linear gaussian equivalences", to obtain sharp analytic predictions for nonlinear maps, which we verify via extensive numerical simulations. By training models of growing size and compressing them to fixed sparsity, we identify a novel double descent behavior, where the risk of the pruned model is consistently minimized in the overparameterized regime. Using our theory, we uncover rather surprising scenarios where pruning an overparameterized model is provably better than training a small model with the exact information of optimal nonzero locations.
(c) Benefits of retraining: An important aspect of pruning is retraining with using only the favorable weights identified during the initial training. We show that retraining can actually hurt the performance when features are uncorrelated. However, it becomes critical as correlations increase. Importantly, we devise the DC of the *train→prune→retrain process* (see Figs. 2 and 4 and the discussion around Def. 5 for details), and, we demonstrate that it correctly captures the pruning performance of random features that are known to be good proxies for understanding DNN behavior (Jacot, Gabriel, and Hongler 2018).

We anticipate that our techniques towards establishing the DC of the overparameterized problems might be useful, beyond the context of pruning, in other statistical inference

*The author names are in alphabetical order.

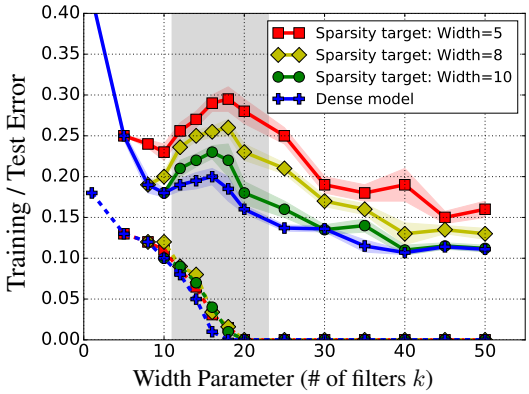


Figure 1: We train sparse ResNet-20 models on the CIFAR-10 dataset with varying width (i.e. number of filters) and sparsity targets. The width parameter controls the overall model size. The solid (resp. dashed) lines are test (resp. training) errors. The blue line corresponds to training of a dense model with width-k. The other three curves correspond to sparsity targets $s \in \{5, 8, 10\}$, for which a dense model of width-k is first pruned to achieve the exact same number of nonzeros as a dense model of width-s and then retrained over the identified nonzero pattern. Surprisingly, all curves interpolate (achieve zero training error) around the same width parameter despite varying sparsity. The best test error is always achieved in the overparameterized regime (large width). Test error curves have two local minima which uncovers a novel double descent phenomena for pruning. The shaded region highlights the transition to zero training error, where the test error peaks.

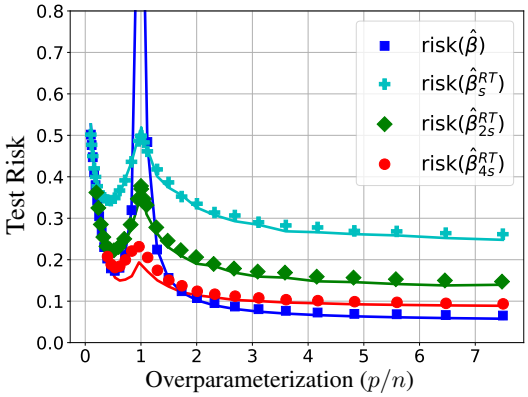


Figure 2: Random feature regression (RFR) with ReLU feature-map $\phi(\boldsymbol{a}) = \text{ReLU}(\boldsymbol{Ra})$. Here $\boldsymbol{R}$ has i.i.d. standard normal entries corresponding to the input layer of a shallow neural net and we regress the output layer. Solid lines follow from our distributional characterization and the markers are obtained by solving random feature regression, which exhibit a good match. The blue line is the performance of usual RFR with growing number of features p. The other lines are obtained by solving RFR with p features and pruning and retraining the solution to fixed sparsity levels $(s, 2s, 4s)$ with $s/n = 0.1$. Importantly, the risks of the retrained models exhibit double descent and are minimized when $p \gg n$ despite fixed model size / sparsity. The slight mismatch of the red curve/markers is explained in Fig. 4.

tasks that require careful distributional studies.

1.1 Prior Art

This work relates to the literature on model compression and overparameterization in deep learning.

Neural network pruning: Large model sizes in deep learning have led to a substantial interest in model pruning/quantization (Han, Mao, and Dally 2015; Hassibi and Stork 1993; LeCun, Denker, and Solla 1990). DNN pruning has a diverse literature with various architectural, algorithmic, and hardware considerations (Sze et al. 2017; Han et al. 2015). The pruning algorithms can be applied before, during, or after training a dense model (Lee, Ajanthan, and Torr 2018; Wang, Zhang, and Grosse 2020; Jin et al. 2016; Oymak 2018) and in this work we focus on after training. Related to over-parameterizarion, (Frankle and Carbin 2019) shows that a large DNN contains a small subset of favorable weights (for pruning), which can achieve similar performance to the original network when trained with the same initialization. (Zhou et al. 2019; Malach et al. 2020; Pensia et al. 2020) demonstrate that there are subsets with good test performance even without any training and provide theoretical guarantees. However, these works do not answer why practical gradient-based algorithms lead to good pruning outcomes. Closer to us, (Li et al. 2020) derives formulas for predicting the pruning performance of over-parameterized least-squares without proofs. In contrast, we provide provable guarantees, and, also obtain DC for more complex problems with general design matrices and nonlinearities.

Benefits of overparameterization: Studies on the optimization and generalization properties of DNNs demonstrate that overparameterization acts as a catalyst for learning. (Arora, Cohen, and Hazan 2018; Neyshabur, Tomioka, and Srebro 2014; Gunasekar et al. 2017; Ji and Telgarsky 2018) argue that gradient-based algorithms are implicitly biased towards certain favorable solutions (even without explicit regularization) to explain benign overfitting (Bartlett et al. 2020; Oymak and Soltanolkotabi 2020; Du et al. 2018; Chizat, Oyallon, and Bach 2019; Belkin, Ma, and Mandal 2018; Belkin, Rakhlin, and Tsybakov 2019; Tsigler and Bartlett 2020; Liang and Rakhlin 2018; Mei and Montanari 2019; Ju, Lin, and Liu 2020). More recently, these studies have led to interesting connections between kernels and DNNs and a flurry of theoretical developments. Closest to us, (Nakkiran et al. 2019; Belkin, Hsu, and Xu 2019; Belkin et al. 2019) uncover a double-descent phenomenon: the test risk has two minima as a function of model size. One minima occurs in the classical underparameterized regime whereas the other minima occurs when the model is overparameterized and the latter risk can in fact be better than former. Closer to our theory, (Dereziński, Liang, and Mahoney 2019; Hastie et al. 2019; Montanari et al. 2019; Deng, Kammoun, and Thrampoulidis 2019; Kini and Thrampoulidis 2020; Liang and Sur 2020; Salehi, Abbasi, and Hassibi 2020; Ju, Lin, and Liu 2020) characterize the asymptotic

performance of overparameterized learning problems. However these works are limited to characterizing the test error of regular (dense) training. In contrast, we use distributional characterization (DC) to capture the performance of more challenging pruning strategies and we uncover novel double descent phenomena (see Fig. 1).

2 Problem Setup

Let us fix the notation. Let $[p] = \{1, 2, \dots, p\}$. Given $\boldsymbol{\beta} \in \mathbb{R}^p$, let $\mathbb{T}_s(\boldsymbol{\beta})$ be the pruning operator that sets the smallest $p - s$ entries in absolute value of $\boldsymbol{\beta}$ to zero. Let $\mathcal{I}(\boldsymbol{\beta}) \subset [p]$ return the index of the nonzero entries of $\boldsymbol{\beta}$. $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. $\boldsymbol{X}^\dagger$ denotes the pseudoinverse of matrix $\boldsymbol{X}$.

Data: Let $(\boldsymbol{a}_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ with i.i.d. input-label pairs. Let $\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}^p$ be a (nonlinear) feature map. We generate $\boldsymbol{x}_i = \phi(\boldsymbol{a}_i)$ and work with the dataset $\mathcal{S} = (\boldsymbol{x}_i, y_i)_{i=1}^n$ coming i.i.d. from some distribution $\mathcal{D}$. As an example, of special interest to the rest of the paper, consider random feature regression, where $\boldsymbol{x}_i = \psi(\boldsymbol{R}\boldsymbol{a}_i)$ for a nonlinear activation function ψ that acts entry-wise and a random matrix $\boldsymbol{R} \in \mathbb{R}^{p \times d}$ with i.i.d. $\mathcal{N}(0, 1)$ entries; see Fig. 2. In matrix notation, we let $\boldsymbol{y} = [y_1 \ \dots \ y_n]^T \in \mathbb{R}^n$ and $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_n]^T \in \mathbb{R}^{n \times p}$ denote the vector of labels and the feature matrix, respectively. Throughout, we focus on regression tasks, in which the training and test risks of a model $\boldsymbol{\beta}$ is defined as

$$\text{Population risk:} \quad \mathcal{L}(\boldsymbol{\beta}) = \mathbb{E}_{\mathcal{D}}[(y - \boldsymbol{x}^T\boldsymbol{\beta})^2]. \tag{1}$$

$$\text{Empirical risk:} \quad \hat{\mathcal{L}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2. \tag{2}$$

During training, we will solve the empirical risk minimization (ERM) problem over a set of selected features $\Delta \subset [p]$, from which we obtain the least-squares solution

$$\hat{\boldsymbol{\beta}}(\Delta) = \arg \min_{\boldsymbol{\beta} \,:\, \mathcal{I}(\boldsymbol{\beta}) = \Delta} \hat{\mathcal{L}}(\boldsymbol{\beta}). \tag{3}$$

For example, regular ERM corresponds to $\Delta = [p]$, and we simply use $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}([p])$ to denote its solution above. Let $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$ be the covariance matrix and $\boldsymbol{b} = \mathbb{E}[y\boldsymbol{x}]$ be the cross-covariance. The parameter minimizing the test error is given by $\boldsymbol{\beta}^\star = \boldsymbol{\Sigma}^{-1}\boldsymbol{b}$. We are interested in training a model over the training set $\mathcal{S}$ that not only achieves small test error, but also, it is sparse. We do this as follows. First, we run stochastic gradient descent (SGD) to minimize the empirical risk (starting from zero initialization). It is common knowledge that SGD on least-squares converges to the minimum ℓ_2 norm solution given by $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\dagger\boldsymbol{y}$. Next, we describe our pruning strategies to compress the model.

Pruning strategies: Given dataset $\mathcal{S}$ and target sparsity level s, a pruning function P takes a model $\boldsymbol{\beta}$ as input and outputs an s-sparse model $\boldsymbol{\beta}_s^P$. Two popular pruning functions are magnitude-based (MP) and Hessian-based (HP) (a.k.a. optimal brain damage) pruning (LeCun, Denker, and Solla 1990). The latter uses a diagonal approximation of the covariance via $\hat{\boldsymbol{\Sigma}} = \text{diag}(\boldsymbol{X}^T\boldsymbol{X})/n$ to capture *saliency* (see (4)). Formally, we have the following definitions:

- *Magnitude-based pruning:* $\boldsymbol{\beta}_s^M = \mathbb{T}_s(\boldsymbol{\beta})$.
- *Hessian-based pruning:* $\boldsymbol{\beta}_s^H = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbb{T}_s(\hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\beta})$.
- *Oracle pruning:* Let $\Delta^\star \subset [p]$ be the optimal s indices so that $\hat{\boldsymbol{\beta}}(\Delta^\star)$ achieves the minimum population risk (in expectation over $\mathcal{S}$) among all $\hat{\boldsymbol{\beta}}(\Delta)$ and any subset Δ in (3). When $\boldsymbol{\Sigma}$ is diagonal and $s < n$, using rather classical results, it can be shown that (see Lemma 7 in the Supplementary Material (SM)) these *oracle indices* are the ones with the top-s saliency score given by

$$\text{Saliency score} = \boldsymbol{\Sigma}_{i,i}\boldsymbol{\beta}_i^{\star 2}. \tag{4}$$

  Oracle pruning employs these latent saliency scores and returns $\boldsymbol{\beta}_s^O$ by pruning the weights of $\boldsymbol{\beta}$ outside of $\Delta^\star$.

We remark that our distributional characterization might allow us to study more complex pruning strategies, such as optimal brain surgeon (Hassibi, Stork, and Wolff 1994). However, we restrict our attention to the three aforementioned core strategies to keep the discussion focused.

Pruning algorithm: To shed light on contemporary pruning practices, we will study the following three-stage *train→prune→retrain* algorithms.

1. Find the empirical risk minimizer $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\dagger\boldsymbol{y}$.
2. Prune $\hat{\boldsymbol{\beta}}$ with strategy P to obtain $\hat{\boldsymbol{\beta}}_s^P$.
3. *Retraining:* Obtain $\hat{\boldsymbol{\beta}}_s^{RT} = \hat{\boldsymbol{\beta}}(\mathcal{I}(\hat{\boldsymbol{\beta}}_s^P))$.

The last step obtains a new s-sparse model by solving ERM in (3) with the features $\Delta = \mathcal{I}(\hat{\boldsymbol{\beta}}_s^P)$ identified by pruning. Figures 1 and 2 illustrate the performance of this procedure for ResNet-20 on the CIFAR-10 dataset and for a random feature regression on a synthetic problem, respectively . Our analytic formulas for RF, as seen in Fig. 1, very closely match the empirical observations (see Sec. 3 for further explanations). Interestingly, the arguably simpler RF model already captures key behaviors (double-descent, better performance in the overparameterized regime, performance of sparse model comparable to large model) in ResNet.

Sections 3 and 4 present numerical experiments on pruning that verify our analytical predictions, as well as, our insights on the fundamental principles behind the roles of overparameterization and retraining. Sec 5 establishes our theory on the DC of $\hat{\boldsymbol{\beta}}$ and provable guarantees on pruning. All proofs are deferred to the Supplementary Material (SM).

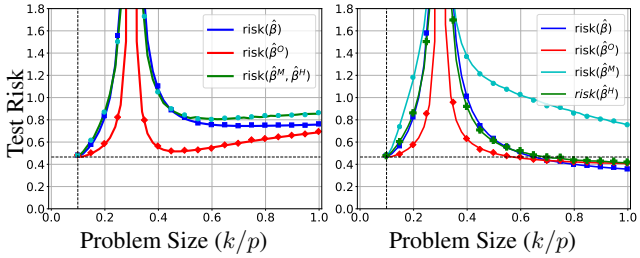
3 Motivating Examples

3.1 Linear Gaussian Problems

We begin our study with linear Gaussian problems (LGP), which we formally define as follows.

Definition 1 (Linear Gaussian Problem (LGP)) *Given latent vector $\boldsymbol{\beta}^\star \in \mathbb{R}^d$, covariance $\boldsymbol{\Sigma}$ and noise level σ, assume that each example in $\mathcal{S}$ is generated independently as $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}^\star + \sigma z_i$ where $z_i \sim \mathcal{N}(0, 1)$ and $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. Additionally, the map $\phi(\cdot)$ is identity and $p = d$.*

Albeit simple, LGPs are of fundamental importance for the following reasons: (1) We show in Sec. 5 that our theoretical



(a) Identity covariance, spiked latent weights.

(b) Spiked covariance, identical latent weights.

Figure 3: Our theoretical predictions for various pruning strategies in linear models with $s/p = 0.1$ and $n/p = 0.3$. We solve ERM using the first k features and then prune to obtain an s-sparse model. The vertical dashed line shows the $k = s$ point. The horizontal dashed line highlights the minimum risk among all underparameterized solutions ($k \leq n$) and all solutions obtained by a final retraining. Retraining curves are omitted here, but they can be found in Fig. 7 of SM.

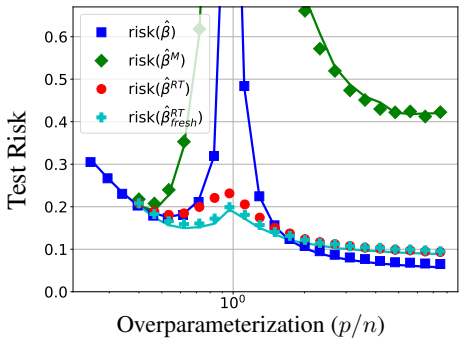


Figure 4: Illustration of the mismatch between pruning with retraining (red markers) and pruning with fresh samples (cyan markers/line). The setting here is exactly the same as in Fig. 2, but we only show the case of sparsity $4s$ for which the mismatch is observed. Observe that our analytical predictions accurately capture the risk of retraining with fresh samples. However, we observe a discrepancy with the true risk of retraining (without fresh samples) around the interpolation threshold. Also shown the risk of the original ERM solution before pruning (in blue) and of the magnitude-pruned model (before any retraining).

framework rigorously characterizes pruning strategies for LGPs; (2) Through a "linear Gaussian equivalence", we will use our results for linear models to obtain analytic predictions for nonlinear random features; (3) Our theoretical predictions and numerical experiments discussed next demonstrate that LGPs already capture key phenomena observed in more complex models (e.g., Fig. 1).

In Fig. 3, we consider LGPs with diagonal covariance $\boldsymbol{\Sigma}$. We set the sparsity level $s/p = 0.1$ and the relative dataset size $n/p = 0.3$. To parameterize the covariance and $\boldsymbol{\beta}^\star$, we use a *spiked* vector $\boldsymbol{\lambda}$, the first s entries of which are set equal to $C = 25 \gg 1$ and the remaining entries equal to 1. $\boldsymbol{\lambda}$ corresponds to the latent saliency score (cf. (4)) of the indices. To understand the role of overparameterization, we vary the number of features used in the optimization. Specifically, we solve (3) with $\Delta = [k]$ and vary the number of features k from 0 to p. Here we consider the *train→prune* algorithm, where we first solve for $\hat{\boldsymbol{\beta}}([k])$ and obtain our pruned model $\hat{\boldsymbol{\beta}}_s^P([k])$ by applying magnitude, Hessian or Oracle pruning (cf. $P \in \{M, H, O\}$). Since $\boldsymbol{\lambda}$ is decreasing, the indices are sorted by saliency score; thus, Oracle pruning always picks the first s indices. Solid lines represent analytic predictions, while markers are empirical results. The vertical dashed line is the sparsity level s/p.

In Fig. 3a, we set $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and $\boldsymbol{\beta}^\star = \sqrt{\boldsymbol{\lambda}}$. Note, that the analytic curves correctly predict the test risk and the double descent behavior. Observe that the Hessian and Magnitude pruning coincide here, since the diagonal of the empirical covariance is essentially identity. In contrast, Fig. 3b emphasizes the role of the feature covariance by setting $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\lambda})$ and $\boldsymbol{\beta}^\star$ to be the all ones vector. In this scenario, we observe that Hessian pruning performs better compared to Fig. 3a and also outperforms Magnitude pruning. This is because the empirical covariance helps distinguish the salient indices. Importantly, for Hessian and Oracle pruning, the optimal sparse model is achieved in the highly overparameterized regime $k = p$. Notably, the achieved performance at $k = p$ is strictly better than the horizontal dashed line, which highlights the optimal risk among all underparameterized solutions $k \leq n$ and all retraining solutions (see also SM Sec. A). This has two striking consequences. First, *retraining can in fact hurt the performance*; because the *train→prune* performance at $k = p$ is strictly better than *train→prune→retrain* for all k. Second, *overparameterized pruning can be provably better than solving the sparse model with the knowledge of the most salient features* as $k = p$ is also strictly better than $k = s$.

3.2 Random Features Regression

We relate an ERM problem (3) with nonlinear map ϕ to an equivalent LGP. This will allows us to use our theoretical results about the latter to characterize the properties of the original nonlinear map. We ensure the equivalence by properly setting up the LGP to exhibit similar second order statistics as the original problem.

Definition 2 (Equivalent Linear Problem) *Given distribution $(\boldsymbol{x}, y) \sim \mathcal{D}$, the equivalent LGP($\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma$) with n samples is given with parameters $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$, $\boldsymbol{\beta}^\star = \boldsymbol{\Sigma}^{-1}\,\mathbb{E}[y\boldsymbol{x}]$ and $\sigma = \mathbb{E}[(y - \boldsymbol{x}^T\boldsymbol{\beta}^\star)^2]^{1/2}$.*

In Section 5, we formalize the DC of LGPs, which enables us to characterize pruning/retraining dynamics. Then, we empirically verify that DC and pruning dynamics of equivalent LGPs can successfully predict the original problem (3) with non-linear features. The idea of setting up and studying equivalent LGPs as a proxy to nonlinear models, has been recently used in the emerging literature of high-dimensional learning, for predicting the performance of the original ERM task (Montanari et al. 2019; Goldt et al. 2020; Abbasi, Salehi, and Hassibi 2019; Dereziński, Liang, and Mahoney 2019). This work goes beyond prior art, which fo-

cuses on ERM, by demonstrating that we can also successfully predict the pruning/retraining dynamics. Formalizing the performance equivalence between LGP and equivalent problem is an important future research avenue and it can presumably be accomplished by building on the recent high-dimensional universality results such as (Oymak and Tropp 2018; Hu and Lu 2020; Abbasi, Salehi, and Hassibi 2019; Goldt et al. 2020).

In Fig. 2, we study random feature regression to approximate a synthetic nonlinear distribution. Specifically, data has the following distribution: Given input $\boldsymbol{a} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, we generate random unit norm $\boldsymbol{\beta}^1 \in \mathbb{R}^d, \boldsymbol{\beta}^2 \in \mathbb{R}^d$ and set the label to be a quadratic function given by $y = \boldsymbol{a}^T\boldsymbol{\beta}^1 + (\boldsymbol{a}^T\boldsymbol{\beta}^2)^2$. Then, we fix $\boldsymbol{R} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and we generate ReLU features $\boldsymbol{x} = \text{ReLU}(\boldsymbol{R}\boldsymbol{a})$, where $\boldsymbol{R}$ corresponds to the input layer of a two-layer network. The markers in Fig. 2 are obtained by solving RFR and pruning and retraining with varying sparsity targets ($s, 2s, 4s$ with $s/n = 10\%$). Here, $d = 10, n = 200$. For each marker, the results are averages of 50 $\boldsymbol{R} \in \mathbb{R}^{p\times d}$ realizations and 10 iterations for each choice of $\boldsymbol{R}$. The lines are obtained via our DC of the equivalent LGP (by using Defs. 4 and 5) where the latent parameter $\boldsymbol{\beta}^\star$, noise σ and the covariance $\boldsymbol{\Sigma}$ of the RFR problem are calculated for fixed realization of the input layer $\boldsymbol{R}$ (similarly averaged over 50 random $\boldsymbol{R}$). Our theory and empirical curves exhibit a good match. The results demonstrate the importance of overparameterization for RF pruning, which corresponds to picking *random features smartly*. Here, the coefficients of least-squares act like a scoring function for the saliency of random features and capture how well they are aligned with the target function. The fact that the risk of the pruned models is minimized in the overparameterized regime implies that least-squares regression succeeds in properly selecting salient random features from a larger candidate set. In the context of deep learning, our discussion can be interpreted as *pruning hidden nodes of the network*.

Predicting retraining performance. As discussed in Sec. 5 and Def. 5, for the retraining stage, our DC is accomplished by assuming that retraining phase uses n fresh training examples (i.e. a new dataset $\mathcal{S}_{\text{fresh}}$). Let us denote the resulting model by $\hat{\boldsymbol{\beta}}^{RT}_{\text{fresh}}$. Perhaps surprisingly, Fig. 2 shows that this DC correctly captures the performance of $\hat{\boldsymbol{\beta}}^{RT}$ with the exception of the red curve ($4s$). Fig. 4 focuses on this instance and shows that our DC indeed perfectly predicts the fresh retraining performance and verifies the slight empirical mismatch between $\hat{\boldsymbol{\beta}}^{RT}$ and $\hat{\boldsymbol{\beta}}^{RT}_{\text{fresh}}$.

3.3 Neural Network Experiments

Finally, we study pruning deep neural networks. Inspired by (Nakkiran et al. 2019), we train ResNet-20 with changeable filters over CIFAR-10. Here, the filter number k is equivalent to the width/channel of the model. As the width of ResNet-20 changes, the fitting performance of the dataset varies. Here, we apply *train→prune→retrain*. Select s as the sparsity target and s-filter ResNet-20 model as the base model with N_s parameters. First, we train a dense model with k filters and N_k parameters, $N_k \gg N_s$, and prune it by only keeping the largest N_s entries in absolute value non-zero. N_k grows approximately quadratically in k. Now, the sparse model shares the same number of parameters amenable to training as the does the base model. Finally, we retain the pruned network and record its performance on the same dataset and same configuration. In Fig. 1, we plot the training and test error of dense and sparse models. All the neural experiments are trained with Adam optimization and 0.001 learning rate for 1000 epochs, with data augmentation. Green, yellow and red lines correspond to 5, 8 or 10 sparsity targets, with around 28,000, 70,000 and 109,000 trainable parameters, respectively. As the width k grows, the training and test error decrease for both 5-,8-,10-filter base models, except for the shaded double descent range. These experiments verify once the main message revealed to us by studying simpler linear and random-feature models, that is, training a larger model, followed by appropriate pruning, can preform better than training a small model from the beginning. Another worth-mentioning observation is that with appropriate sparsity level (here, 10) the pruned model has prediction performance comparable to the dense model. Finally and interestingly, the test error dynamics of the pruned model exhibit a double descent that resembles that of the dense model (previously observed in (Nakkiran et al. 2019)).

3.4 Further Intuitions on The Denoising Effect of Overparameterization

To provide further insights into the pruning benefits of overparameterization, consider a simple linear model (as in Def 1) with $n \geq p \geq s$, noise level $\sigma = 0$ and identity covariance. Suppose our goal is estimating the coefficients $\boldsymbol{\beta}^\star_\Delta$ for some fixed index set $\Delta \subset [p]$ with $|\Delta| = s$. For pruning, we can pick Δ to be the most salient/largest entries. If we solve the smaller regression problem over Δ, $\hat{\boldsymbol{\beta}}(\Delta)$ will only provide a noisy estimate of $\boldsymbol{\beta}^\star_\Delta$. The reason is that, the signal energy of the missing features $[p] - \Delta$ acts as a noise uncorrelated with the features in Δ. Conversely, if we solve ERM with all features (the larger problem), we perfectly recover $\boldsymbol{\beta}^\star$ due to zero noise and invertibility ($n \geq p$). Then one can also perfectly estimate $\boldsymbol{\beta}^\star_\Delta$. This simple argument, which is partly inspired by the missing feature setup in (Hastie et al. 2019), shows that solving the larger problem with more parameters can have a "denoising-like effect" and perform better than the small problem. Our contribution obviously goes well beyond this discussion and theoretically characterizes the exact asymptotics, handles the general covariance model and all (n, p, s) regimes, and also highlights the importance of the overparameterized regime $n \ll p$.

4 Understanding the Benefits of Retraining

On the one hand, the study of LGPs in Fig. 3 and Fig. 7 of SM suggest that retraining can actually hurt the performance. On the other hand, in practice and in the RFR experiments of Fig. 4, retraining is crucial; compare the green $\hat{\boldsymbol{\beta}}^M$ and red $\hat{\boldsymbol{\beta}}^{RT}$ curves and see SM Section A for further DNN experiments. Here, we argue that the benefit of retraining is connected to the correlations between input features. Indeed, the covariance/Hessian matrices associated with RF and DNN regression are not diagonal (as was the case in

Fig. 3). To build intuition, imagine that only a single feature suffices to explain the label. If there are multiple other features that can similarly explain the label, the model prediction will be shared across these features. Then, pruning will lead to a biased estimate, which can be mitigated by retraining. The following lemma formalizes this intuition under an instructive setup, where the features are perfectly correlated.

Lemma 1 *Suppose $\mathcal{S}$ is drawn from an LGP$(\sigma, \boldsymbol{\Sigma}, \boldsymbol{\beta}_\star)$ as in Def. 1 where rank$(\boldsymbol{\Sigma}) = 1$ with $\boldsymbol{\Sigma} = \boldsymbol{\lambda}\boldsymbol{\lambda}^T$ for $\boldsymbol{\lambda} \in \mathbb{R}^p$. Define $\zeta = \mathbb{T}_s(\boldsymbol{\lambda})^2/\|\boldsymbol{\lambda}\|_{\ell_2}^2$. For magnitude and Hessian pruning ($P \in \{M, H\}$) and the associated retraining, we have the following excess risks with respect to $\boldsymbol{\beta}^\star$*

$$\mathbb{E}_{\mathcal{S}}[\mathcal{L}(\hat{\boldsymbol{\beta}}_s^P)] - \mathcal{L}(\boldsymbol{\beta}^\star) = \frac{\zeta^2\sigma^2}{n-2} + \underbrace{(1-\zeta)^2(\boldsymbol{\lambda}^T\boldsymbol{\beta}^\star)^2}_{\text{Error due to bias}} \quad (5)$$

$$\mathbb{E}_{\mathcal{S}}[\mathcal{L}(\hat{\boldsymbol{\beta}}_s^{RT})] - \mathcal{L}(\boldsymbol{\beta}^\star) = \sigma^2/(n-2). \quad (6)$$

The lemma reveals that pruning the model leads to a biased estimator of the label. Specifically, the bias coefficient $1-\zeta$ arises from the missing predictions of the pruned features (which correspond to the small coefficients of $|\boldsymbol{\lambda}|$). In contrast, regardless of s, retraining always results in an unbiased estimator with the exact same risk as the dense model which quickly decays in sample size n. The reason is that retraining enables the remaining features to account for the missing predictions. Here, this is accomplished perfectly, due to the fully correlated nature of the problem. In particular, this is in contrast to the diagonal covariance (Fig. 3), where the missing features act like uncorrelated noise during retraining.

5 Main Results

Here, we present our main theoretical result: a sharp asymptotic characterization of the distribution of the solution to overparameterized least-squares for correlated designs. We further show how this leads to a sharp prediction of the risk of magnitude-based pruning. Concretely, for the rest of this section, we assume the linear Gaussian problem (LGP) of Definition 1, the overparameterized regime $k = p > n$ and the min-norm model

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_{\ell_2} \text{ s.t. } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}. \quad (7)$$

As mentioned in Sec. 2, $\hat{\boldsymbol{\beta}}$ is actually given in closed-form as $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\dagger\boldsymbol{y}$. Interestingly, our analysis of the distribution of $\hat{\boldsymbol{\beta}}$ does not rely on the closed-form expression, but rather follows by viewing $\hat{\boldsymbol{\beta}}$ as the solution to the convex linearly-constrained quadratic program in (7). Specifically, our analysis uses the framework of the convex Gaussian min-max Theorem (CGMT) (Thrampoulidis, Oymak, and Hassibi 2015), which allows to study rather general inference optimization problems such as the one in (7), by relating them with an auxiliary optimization that is simpler to analyze (Stojnic 2013; Oymak, Thrampoulidis, and Hassibi 2013; Thrampoulidis, Oymak, and Hassibi 2015; Thrampoulidis, Abbasi, and Hassibi 2018; Salehi, Abbasi, and Hassibi 2019; Taheri, Pedarsani, and Thrampoulidis 2020). Due to space considerations, we focus here on the more challenging overparameterized regime and defer the analysis of the underparameterized regime to the SM.

5.1 Distributional Characterization of the Overparameterized Linear Gaussian Models

Notation: We first introduce additional notation necessary to state our theoretical results. $\odot$ denotes the entrywise product of two vectors and $\mathbf{1}_p$ is the all ones vector in $\mathbb{R}^p$. The *empirical distribution* of a vector $\boldsymbol{x} \in \mathbb{R}^p$ is given by $\frac{1}{p}\sum_{i=1}^p \delta_{x_i}$, where δ_{x_i} denotes a Dirac delta mass on x_i. Similarly, the empirical joint distribution of vectors $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$ is $\frac{1}{p}\sum_{i=1}^p \delta_{(x_i, x_i')}$. The *Wasserstein-k* ($W_k$) distance between two measures μ and ν is defined as $W_k(\mu,\nu) \equiv \left(\inf_\rho \mathbb{E}_{(X,Y)\sim\rho}|X-Y|^k\right)^{1/k}$, where the infimum is over all the couplings of μ and ν, i.e. all random variables (X, Y) such that $X \sim \nu$ and $Y \sim \nu$ marginally. A sequence of probability distributions ν_p on $\mathbb{R}^m$ converges in W_k to ν, written $\nu_p \overset{W_k}{\Longrightarrow} \nu$, if $W_k(\nu_p, \nu) \to 0$ as $p \to \infty$. Finally, we say that a function $f : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order k, denoted $f \in \mathrm{PL(k)}$, if there is a constant $L > 0$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$, $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le L(1 + \|\boldsymbol{x}\|_{\ell_2}^{k-1} + \|\boldsymbol{y}\|_{\ell_2}^{k-1})\|\boldsymbol{x} - \boldsymbol{y}\|_2$. We call L the PL(k) constant of f. An equivalent definition of W_k convergence is that, for any $f \in \mathrm{PL(k)}$, $\lim_{p\to\infty} \mathbb{E}\, f(X_p) = \mathbb{E}\, f(X)$, where expectation is with respect to $X_p \sim \nu_p$ and $X \sim \nu$. For a sequence of random variables $\mathcal{X}_p$ that converge in probability to some constant c in the limit of Assumption 1 below, we write $\mathcal{X}_p \xrightarrow{P} c$.

Next, we formalize the set of assumption under which our analysis applies. Our asymptotic results hold in the linear asymptotic regime specified below.

Assumption 1 *We focus on a double asymptotic regime where $n, p, s \to \infty$ at fixed overparameterization ratio $\kappa := p/n > 0$ and sparsity level $\alpha := s/p \in (0, 1)$.*

Additionally, we require certain mild assumptions on the behavior of the covariance matrix $\boldsymbol{\Sigma}$ and of the true latent vector $\boldsymbol{\beta}^\star$. For simplicity, we assume here that $\boldsymbol{\Sigma} = \mathrm{diag}([\boldsymbol{\Sigma}_{1,1}, \ldots, \boldsymbol{\Sigma}_{p,p}])$.

Assumption 2 *The covariance matrix $\boldsymbol{\Sigma}$ is diagonal and there exist constants $\Sigma_{\min}, \Sigma_{\max} \in (0, \infty)$ such that: $\Sigma_{\min} \le \boldsymbol{\Sigma}_{i,i} \le \Sigma_{\max}$, for all $i \in [p]$.*

Assumption 3 *The joint empirical distribution of $\{(\boldsymbol{\Sigma}_{i,i}, \sqrt{p}\boldsymbol{\beta}_i^\star)\}_{i\in[p]}$ converges in Wasserstein-k distance to a probability distribution μ on $\mathbb{R}_{>0} \times \mathbb{R}$ for some $k \ge 4$. That is $\frac{1}{p}\sum_{i\in[p]} \delta_{(\boldsymbol{\Sigma}_{i,i}, \sqrt{p}\boldsymbol{\beta}_i^\star)} \overset{W_k}{\Longrightarrow} \mu$.*

With these, we are ready to define, what will turn out to be, the asymptotic DC in the overparameterized regime.

Definition 3 (Asymptotic DC – Overparameterized regime) *Let random variables $(\Lambda, B) \sim \mu$ (where μ is defined in Assumption 3) and fix $\kappa > 1$. Define parameter ξ as the unique positive solution to the following equation*

$$\mathbb{E}_\mu\left[\left(1 + (\xi\cdot\Lambda)^{-1}\right)^{-1}\right] = \kappa^{-1}. \quad (8)$$

Further define the positive parameter γ as follows:

$$\gamma := \left(\sigma^2 + \mathbb{E}_\mu\left[\frac{B^2\Lambda}{(1+\xi\Lambda)^2}\right]\right) \Big/ \left(1 - \kappa\,\mathbb{E}_\mu\left[\frac{1}{(1+(\xi\Lambda)^{-1})^2}\right]\right). \quad (9)$$

With these and $H \sim \mathcal{N}(0,1)$, define the random variable

$$X_{\kappa,\sigma^2}(\Lambda, B, H) := \Big(1 - \frac{1}{1+\xi\Lambda}\Big)B + \sqrt{\kappa}\frac{\sqrt{\gamma}\,\Lambda^{-1/2}}{1+(\xi\Lambda)^{-1}}H, \tag{10}$$

and let Π_{κ,σ^2} be its distribution.

Our main result establishes asymptotic convergence of the empirical distribution of $(\sqrt{p}\hat{\boldsymbol{\beta}}, \sqrt{p}\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})$ for a rich class of test functions. These are the functions within $\mathrm{PL}(3)$ that become $\mathrm{PL}(2)$ when restricted to the first two indices. Formally, we define this class of functions as follows

$$\mathcal{F} := \{f : \mathbb{R}^2 \times \mathcal{Z} \to \mathbb{R},\ f \in \mathrm{PL}(3) \text{ and } \sup_{z\in\mathcal{Z}} \text{“PL(2) constant of f}(\cdot,\cdot,z)\text{”} < \infty\}. \tag{11}$$

For pruning analysis, we set $\mathcal{Z} = [\Sigma_{\min}, \Sigma_{\max}]$ and define

$$\mathcal{F}_{\mathcal{L}} := \{f : \mathbb{R}^2 \times \mathcal{Z} \to \mathbb{R} \;\big|\; f(x,y,z) = z(y-g(x))^2 \text{ where } g(\cdot) \text{ is Lipschitz}\}. \tag{12}$$

As discussed below, $\mathcal{F}_{\mathcal{L}}$ is important for predicting the risk of the (pruned) model. In the SM, we prove that $\mathcal{F}_{\mathcal{L}} \subset \mathcal{F}$. We are now ready to state our main theoretical result.

Theorem 1 (Asymptotic DC – Overparameterized LGP) *Fix $\kappa > 1$ and suppose Assumptions 2 and 3 hold. Recall the solution $\hat{\boldsymbol{\beta}}$ from* (7) *and let*

$$\hat{\Pi}_n(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) := \frac{1}{p}\sum_{i=1}^{p} \delta_{(\sqrt{p}\hat{\boldsymbol{\beta}}_i, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i})}$$

be the joint empirical distribution of $(\sqrt{p}\hat{\boldsymbol{\beta}}, \sqrt{p}\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})$. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a function in $\mathcal{F}$ defined in (11)*. We have that*

$$\frac{1}{p}\sum_{i=1}^{p} f(\sqrt{p}\hat{\boldsymbol{\beta}}_i, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}) \xrightarrow{P} \mathbb{E}\left[f(X_{\kappa,\sigma^2}, B, \Lambda)\right]. \tag{13}$$

As advertised, Theorem 1 fully characterizes the joint empirical distribution of the min-norm solution, the latent vector and the covariance spectrum. The asymptotic DC allows us to precisely characterize several quantities of interest, such as estimation error, generalization error etc.. For example, a direct application of (13) to the function $f(x,y,z) = z(y-x)^2 \in \mathcal{F}_{\mathcal{L}} \subset \mathcal{F}$ directly yields the risk prediction of the min-norm solution recovering (Hastie et al. 2019, Thm. 3) as a special case. Later in this section, we show how to use Theorem 1 towards the more challenging task of precisely characterizing the risk of magnitude-based pruning.

Before that, let us quickly remark on the technical novelty of the theorem. Prior work has mostly applied the CGMT to isotropic features. Out of these, only very few obtain DC, (Thrampoulidis, Xu, and Hassibi 2018; Miolane and Montanari 2018), while the majority focuses on simpler metrics, such as squared-error. Instead, Theorem 1 considers correlated designs and the overparameterized regime. The most closely related work in that respect is (Montanari et al. 2019), which very recently obtained the DC of the max-margin classifier. Similar to us, they use the CGMT, but their analysis of the auxiliary optimization is technically different to ours. Our approach is similar to (Thrampoulidis, Xu, and Hassibi 2018), but extra technical effort is needed to account for correlated designs and the overparameterized regime.

5.2 From DC to Risk Characterization

First, we consider a simpler “threshold-based" pruning method that applies a fixed threshold at every entry of $\hat{\boldsymbol{\beta}}$. Next, we relate this to magnitude-based pruning and obtain a characterization for the performance of the latter. In order to define the threshold-based pruning vector, let

$$\mathcal{T}_t(x) = \begin{cases} x & \text{if } |x| > t \\ 0 & \text{otherwise} \end{cases},$$

be the hard-thresholding function with fixed threshold $t \in \mathbb{R}_+$. Define $\hat{\boldsymbol{\beta}}_t^{\mathcal{T}} := \mathcal{T}_{t/\sqrt{p}}(\hat{\boldsymbol{\beta}})$, where $\mathcal{T}_t$ acts componentwise. Then, the population risk of $\hat{\boldsymbol{\beta}}_t^{\mathcal{T}}$ becomes

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\beta}}_t^{\mathcal{T}}) &= \mathbb{E}_{\mathcal{D}}[(\boldsymbol{x}^T(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}_t^{\mathcal{T}}) + \sigma z)^2] \\ &= \sigma^2 + \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{\Sigma}_{i,i}\big(\sqrt{p}\boldsymbol{\beta}_i^\star - \mathcal{T}_t(\sqrt{p}\hat{\boldsymbol{\beta}}_i)\big)^2 \\ &\xrightarrow{P} \sigma^2 + \mathbb{E}\left[\Lambda\left(B - \mathcal{T}_t(X_{\kappa,\sigma^2})\right)\right].\end{aligned} \tag{14}$$

In the second line above, we note that $\sqrt{p}\mathcal{T}_{t'}(x) = \mathcal{T}_{\sqrt{p}t'}(\sqrt{p}x)$. In the last line, we apply (13), after recognizing that the function $(x,y,z) \mapsto z(y-\mathcal{T}_t(x))^2$ is a member of the $\mathcal{F}_{\mathcal{L}}$ family defined in (12). As in (13), the expectation here is with respect to $(\Lambda, B, H) \sim \mu \otimes \mathcal{N}(0,1)$.

Now, we show how to use (14) and Theorem 1 to characterize the risk of the magnitude-based pruned vector $\boldsymbol{\beta}_s^M := \mathbb{T}_s(\hat{\boldsymbol{\beta}})$. Recall, here from Assumption 1 that $s = \alpha p$. To relate $\hat{\boldsymbol{\beta}}_s^M$ to $\hat{\boldsymbol{\beta}}_t^{\mathcal{T}}$, consider the set $\mathcal{S}_t := \{i \in [p] : \sqrt{p}|\hat{\boldsymbol{\beta}}_i| \geq t\}$ for some constant $t \in \mathbb{R}_+$ (not scaling with n, p, s). Note that the ratio $|\mathcal{S}_t|/p$ is equal to

$$p^{-1}\sum_{i=1}^{p} \mathbb{1}_{[\sqrt{p}|\hat{\boldsymbol{\beta}}_i| \geq t]} \xrightarrow{P} \mathbb{E}[\mathbb{1}_{[|X_{\kappa,\sigma^2}| \geq t]}] = \mathbb{P}\left(|X_{\kappa,\sigma^2}| \geq t\right).$$

Here, $\mathbb{1}$ denotes the indicator function and the convergence follows from Theorem 1 when applied to a sequence of bounded Lipschitz functions approximating the indicator. Thus, by choosing

$$t^\star := \inf\left\{t \in \mathbb{R} : \mathbb{P}(|X_{\kappa,\sigma^2}| \geq t) \geq \alpha\right\}, \tag{15}$$

it holds that $|\mathcal{S}_t|/p \xrightarrow{P} \alpha$. In words, and observing that X_{κ,σ^2} admits a continuous density (due to the Gaussian variable H): for any $\varepsilon > 0$, in the limit of $n, p, s \to \infty$, the vector $\hat{\boldsymbol{\beta}}_{t^\star}^{\mathcal{T}}$ has $(1\pm\varepsilon)\alpha p = (1\pm\varepsilon)s$ non-zero entries, which correspond to the largest magnitude entries of $\hat{\boldsymbol{\beta}}$, with probability approaching1. Since this hold for arbitrarily small $\varepsilon > 0$, recalling $t^\star$ as in (15), we can conclude from (14) that the risk of the magnitude-pruned model converges as follows.

Corollary 1 (Risk of Magnitude-pruning) *Let the same assumptions and notation as in the statement of Theorem 1 hold. Specifically, let $\hat{\boldsymbol{\beta}}$ be the min-norm solution in* (7) *and $\hat{\boldsymbol{\beta}}_s^M := \mathbb{T}_s(\boldsymbol{\beta})$ the magnitude-pruned model at sparsity s. Recall the threshold $t^\star$ from* (15)*. The risk of $\hat{\boldsymbol{\beta}}_s^M$ satisfies the following in the limit of $n, p, s \to \infty$ at rates $\kappa := p/n > 1$ and $\alpha := s/p \in (0,1)$ (cf. Assumption 1):*

$$\mathcal{L}(\hat{\boldsymbol{\beta}}_s^M) \xrightarrow{P} \sigma^2 + \mathbb{E}\left[\Lambda\left(B - \mathcal{T}_{t^\star}(X_{\kappa,\sigma^2})\right)\right],$$

where the expectation is over $(\Lambda, B, H) \sim \mu \otimes \mathcal{N}(0,1)$.

5.3 Non-asymptotic DC and Retraining Formula

While Theorem 1 is stated in the asymptotic regime, during analysis, the DC arises in a non-asymptotic fashion. The following definition is the non-asymptotic counterpart of Def. 3. We remark that this definition applies to arbitrary covariance (not necessarily diagonal) by applying a simple eigen-rotation before and after the DC formula associated with the diagonalized covariance.

Definition 4 (Non-asymptotic DC) *Fix $p > n \geq 1$ and set $\kappa = p/n > 1$. Given $\sigma > 0$, covariance $\boldsymbol{\Sigma} = \boldsymbol{U} diag(\boldsymbol{\lambda}) \boldsymbol{U}^T$ and latent vector $\boldsymbol{\beta}$, set $\bar{\boldsymbol{\beta}} = \boldsymbol{U}^T \boldsymbol{\beta}$ and define the unique non-negative terms $\xi, \gamma, \boldsymbol{\zeta} \in \mathbb{R}^p$ and $\boldsymbol{\phi} \in \mathbb{R}^p$ as follows:*

$$\xi > 0 \quad \textit{is the solution of} \quad \kappa^{-1} = p^{-1} \sum_{i=1}^{p} \left(1 + (\xi \boldsymbol{\lambda}_i)^{-1}\right)^{-1},$$

$$\gamma = \frac{\sigma^2 + \sum_{i=1}^{p} \boldsymbol{\lambda}_i \boldsymbol{\zeta}_i^2 \bar{\boldsymbol{\beta}}_i^2}{1 - \frac{\kappa}{p} \sum_{i=1}^{p} (1 + (\xi \boldsymbol{\lambda}_i)^{-1})^{-2}},$$

$$\boldsymbol{\zeta}_i = (1 + \xi \boldsymbol{\lambda}_i)^{-1} \quad , \quad \boldsymbol{\phi}_i = \kappa \gamma (1 + (\xi \boldsymbol{\lambda}_i)^{-1})^{-2}, \; 1 \leq i \leq p.$$

The non-asymptotic distributional prediction is given by the following $\boldsymbol{U}$-rotated normal distribution

$$\mathcal{D}_{\sigma, \boldsymbol{\Sigma}, \boldsymbol{\beta}} = \boldsymbol{U} \mathcal{N}((\mathbf{1}_p - \boldsymbol{\zeta}) \odot \bar{\boldsymbol{\beta}}, p^{-1} diag(\boldsymbol{\lambda}^{-1} \odot \boldsymbol{\phi})).$$

We remark that this definition is similar in spirit to the concurrent/recent work (Li et al. 2020). However, unlike this work, here we prove the asymptotic correctness of the DC, we use it to rigorously predict the pruning performance and also extend this to retraining DC as discussed next.

Retraining DC. As the next step, we would like to characterize the DC of the solution after retraining, i.e., $\hat{\boldsymbol{\beta}}^{RT}$. We carry out the retraining derivation (for magnitude pruning) as follows. Let $\mathcal{I} \subset [p]$ be the nonzero support of the pruned vector $\hat{\boldsymbol{\beta}}_s^M$. Re-solving (2) restricted to the features over $\mathcal{I}$ corresponds to a linear problem with effective feature covariance $\boldsymbol{\Sigma}_{\mathcal{I}}$ with support of non-zeros restricted to $\mathcal{I} \times \mathcal{I}$. For this feature covariance, we can also calculate the effective noise level and global minima of the population risk $\boldsymbol{\beta}_{\mathcal{I}}^{\star}$. The latter has the closed-form solution $\boldsymbol{\beta}_{\mathcal{I}}^{\star} = \boldsymbol{\Sigma}_{\mathcal{I}}^{\dagger} \boldsymbol{\Sigma} \boldsymbol{\beta}^{\star}$. The effective noise is given by accounting for the risk change due to the missing features via $\sigma_{\mathcal{I}} = (\sigma^2 + \boldsymbol{\beta}^{\star T} \boldsymbol{\Sigma} \boldsymbol{\beta}^{\star} - \boldsymbol{\beta}_{\mathcal{I}}^{\star T} \boldsymbol{\Sigma}_{\mathcal{I}} \boldsymbol{\beta}_{\mathcal{I}}^{\star})^{1/2}$. With these terms in place, fixing $\mathcal{I}$ and using Def. 4, the retraining prediction becomes $\mathcal{D}_{\sigma_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}, \boldsymbol{\beta}_{\mathcal{I}}}$. This process is summarized below.

Definition 5 (Retraining DC) *Consider the setting of Def. 4 with $\sigma, \boldsymbol{\Sigma}, \boldsymbol{\beta}^{\star}$ and sparsity target s. The sample $\hat{\boldsymbol{\beta}}^{RT}$ from the retraining distribution $\mathcal{D}^{RT,s}_{\sigma, \boldsymbol{\Sigma}, \boldsymbol{\beta}^{\star}}$ is constructed as follows. Sample $\hat{\boldsymbol{\beta}} \sim \mathcal{D}_{\sigma, \boldsymbol{\Sigma}, \boldsymbol{\beta}^{\star}}$ and compute the set of the top-s indices $\mathcal{I} = \mathcal{I}(\mathbb{T}_s(\hat{\boldsymbol{\beta}}))$. Given $\mathcal{I}$, obtain the effective covariance $\boldsymbol{\Sigma}_{\mathcal{I}} \in \mathbb{R}^{p \times p}$, population minima $\boldsymbol{\beta}_{\mathcal{I}}^{\star} \in \mathbb{R}^p$, and the noise level $\sigma_{\mathcal{I}} > 0$ as described above. Draw $\hat{\boldsymbol{\beta}}^{RT} \sim \mathcal{D}_{\sigma_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}, \boldsymbol{\beta}_{\mathcal{I}}}$.*

Observe that, the support $\mathcal{I}$ depends on the samples $\mathcal{S}$ via $\hat{\boldsymbol{\beta}}$. Thus, our retraining DC is actually derived for the scenario when the retraining phase uses a fresh set of n samples to break the dependence between $\mathcal{I}, \mathcal{S}$ (which obtains $\hat{\boldsymbol{\beta}}^{RT}_{\text{fresh}}$). Despite this, we empirically observe that the retraining DC predicts the regular retraining (reusing $\mathcal{S}$) performance remarkably well and perfectly predicts $\hat{\boldsymbol{\beta}}^{RT}_{\text{fresh}}$ as discussed in Figs. 2 and 4. Finally, we defer the formalization of the retraining analysis to a future work. This includes proving that $\hat{\boldsymbol{\beta}}^{RT}_{\text{fresh}}$ obeys Def. 5 asymptotically as well as directly studying $\hat{\boldsymbol{\beta}}^{RT}$ by capturing the impact of the $\mathcal{I}, \mathcal{S}$ dependency.

6 Conclusions and Future Directions

This paper sheds light on under-explored phenomena in pruning practices for neural network model compression. On a theoretical level, we prove an accurate distributional characterization (DC) for the solution of overparameterized least-squares for linear models with correlated Gaussian features. Our DC allows to precisely characterize the pruning performance of popular pruning methods, such as magnitude pruning. Importantly, our DC combined with a linear Gaussian equivalence, leads to precise analytic formulas for the pruning performance of nonlinear random feature models. On the experimental side, we provide a thorough study of overparameterization and pruning with experiments on linear models, random features and neural nets with growing complexity. Our experiments reveal striking phenomena such as a novel double descent behavior for model pruning and the power of overparameterization. They also shed light on common practices such as retraining after pruning.

Going forward, there are several exciting directions to pursue. First, it would be insightful to study whether same phenomena occur for other loss functions in particular for cross-entropy. Second, this work focuses on unregularized regression tasks and it is important to identify optimal regularization schemes for pruning purposes. For instance, should we use classical ℓ_1/ℓ_2 regularization or can we refine them by injecting problem priors such as covariance information? Finally, going beyond pruning, using DC, one can investigate other compression techniques that processes the output of the initial overparameterized learning problem, such as model quantization and distillation.

Acknowledgments

S. Oymak is partially supported by the NSF award CNS-1932254. C. Thrampoulidis is partially supported by the NSF under Grant Numbers CCF-2009030.

Potential Ethical Impacts

While deep learning is transformative in wide swath of applications, it comes at a cost: State-of-the-art deep learning models tend to be very large and consume significant energy during inference. The race for larger and better models and growing list of applications exacerbates this carbon footprint problem. Thus there is an urgent need for better and more principled model compression methods to help build environmentally friendly ML models. This work responds to this need by establishing the fundamental algorithmic principles and guarantees behind the contemporary model compression algorithms and by shedding light on the design of lightweight energy- and compute-efficient neural networks. We do not see an ethical concern associated with this work.

References

Abbasi, E.; Salehi, F.; and Hassibi, B. 2019. Universality in learning from linear measurements. In *Advances in Neural Information Processing Systems*, 12372–12382.

Andersen, P. K.; and Gill, R. D. 1982. Cox's regression model for counting processes: a large sample study. *The annals of statistics* 1100–1120.

Arora, S.; Cohen, N.; and Hazan, E. 2018. On the optimization of deep networks: Implicit acceleration by overparameterization. In *35th International Conference on Machine Learning*.

Bartlett, P. L.; Long, P. M.; Lugosi, G.; and Tsigler, A. 2020. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences* .

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

Belkin, M.; Hsu, D.; Ma, S.; and Mandal, S. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* 116(32): 15849–15854.

Belkin, M.; Hsu, D.; and Xu, J. 2019. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571* .

Belkin, M.; Ma, S.; and Mandal, S. 2018. To Understand Deep Learning We Need to Understand Kernel Learning. In *International Conference on Machine Learning*, 541–549.

Belkin, M.; Rakhlin, A.; and Tsybakov, A. B. 2019. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1611–1619.

Chizat, L.; Oyallon, E.; and Bach, F. 2019. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2933–2943.

Deng, Z.; Kammoun, A.; and Thrampoulidis, C. 2019. A Model of Double Descent for High-dimensional Binary Linear Classification. *arXiv preprint arXiv:1911.05822* .

Dereziński, M.; Liang, F.; and Mahoney, M. W. 2019. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv preprint arXiv:1912.04533* .

Du, S. S.; Lee, J. D.; Li, H.; Wang, L.; and Zhai, X. 2018. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804* .

Fan, K. 1953. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America* 39(1): 42.

Frankle, J.; and Carbin, M. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*.

Goldt, S.; Reeves, G.; Mézard, M.; Krzakala, F.; and Zdeborová, L. 2020. The Gaussian equivalence of generative models for learning with two-layer neural networks. *arXiv preprint arXiv:2006.14709* .

Gunasekar, S.; Woodworth, B. E.; Bhojanapalli, S.; Neyshabur, B.; and Srebro, N. 2017. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, 6151–6159.

Han, S.; Mao, H.; and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* .

Han, S.; Pool, J.; Tran, J.; and Dally, W. 2015. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, 1135–1143.

Hassibi, B.; and Stork, D. G. 1993. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in neural information processing systems*, 164–171.

Hassibi, B.; Stork, D. G.; and Wolff, G. 1994. Optimal brain surgeon: Extensions and performance comparisons. In *Advances in neural information processing systems*, 263–270.

Hastie, T.; Montanari, A.; Rosset, S.; and Tibshirani, R. J. 2019. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560* .

Hu, H.; and Lu, Y. M. 2020. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669* .

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, 8571–8580.

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

Ji, Z.; and Telgarsky, M. 2018. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300* .

Jin, X.; Yuan, X.; Feng, J.; and Yan, S. 2016. Training skinny deep neural networks with iterative hard thresholding methods. *arXiv preprint arXiv:1607.05423* .

Ju, P.; Lin, X.; and Liu, J. 2020. Overfitting Can Be Harmless for Basis Pursuit, But Only to a Degree. *Advances in Neural Information Processing Systems* 33.

Kini, G.; and Thrampoulidis, C. 2020. Analytic study of double descent in binary classification: The impact of loss. *arXiv preprint arXiv:2001.11572* .

LeCun, Y.; Denker, J. S.; and Solla, S. A. 1990. Optimal brain damage. In *Advances in neural information processing systems*, 598–605.

Lee, N.; Ajanthan, T.; and Torr, P. H. 2018. Snip: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340* .

Li, M.; Sattar, Y.; Thrampoulidis, C.; and Oymak, S. 2020. Exploring Weight Importance and Hessian Bias in Model Pruning. *arXiv preprint arXiv:2006.10903* .

Liang, T.; and Rakhlin, A. 2018. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387* .

Liang, T.; and Sur, P. 2020. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586* .

Malach, E.; Yehudai, G.; Shalev-Shwartz, S.; and Shamir, O. 2020. Proving the Lottery Ticket Hypothesis: Pruning is All You Need. *arXiv preprint arXiv:2002.00585* .

Mei, S.; and Montanari, A. 2019. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355* .

Miolane, L.; and Montanari, A. 2018. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212* .

Montanari, A.; Ruan, F.; Sohn, Y.; and Yan, J. 2019. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544* .

Montanari, A.; and Venkataramanan, R. 2017. Estimation of low-rank matrices via approximate message passing. *arXiv preprint arXiv:1711.01682* .

Nakkiran, P.; Kaplun, G.; Bansal, Y.; Yang, T.; Barak, B.; and Sutskever, I. 2019. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292* .

Newey, W. K.; and McFadden, D. 1994. Large sample estimation and hypothesis testing. *Handbook of econometrics* 4: 2111–2245.

Neyshabur, B.; Tomioka, R.; Salakhutdinov, R.; and Srebro, N. 2017. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071* .

Neyshabur, B.; Tomioka, R.; and Srebro, N. 2014. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614* .

Oymak, S. 2018. Learning Compact Neural Networks with Regularization. *International Conference on Machine Learning* .

Oymak, S.; and Soltanolkotabi, M. 2020. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory* .

Oymak, S.; Thrampoulidis, C.; and Hassibi, B. 2013. The squared-error of generalized lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1002–1009. IEEE.

Oymak, S.; and Tropp, J. A. 2018. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA* 7(3): 337–446.

Pensia, A.; Rajput, S.; Nagle, A.; Vishwakarma, H.; and Papailiopoulos, D. 2020. Optimal Lottery Tickets via SubsetSum: Logarithmic Over-Parameterization is Sufficient. *arXiv preprint arXiv:2006.07990* .

Salehi, F.; Abbasi, E.; and Hassibi, B. 2019. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, 12005–12015.

Salehi, F.; Abbasi, E.; and Hassibi, B. 2020. The Performance Analysis of Generalized Margin Maximizers on Separable Data. In *International Conference on Machine Learning*, 8417–8426. PMLR.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Stojnic, M. 2013. A framework to characterize performance of LASSO algorithms. *arXiv preprint arXiv:1303.7291* .

Sze, V.; Chen, Y.-H.; Yang, T.-J.; and Emer, J. S. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE* 105(12): 2295–2329.

Taheri, H.; Pedarsani, R.; and Thrampoulidis, C. 2020. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. *arXiv preprint arXiv:2006.08917* .

Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; and Le, Q. V. 2019. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2820–2828.

Thrampoulidis, C.; Abbasi, E.; and Hassibi, B. 2018. Precise Error Analysis of Regularized M-Estimators in High Dimensions. *IEEE Transactions on Information Theory* 64(8): 5592–5628.

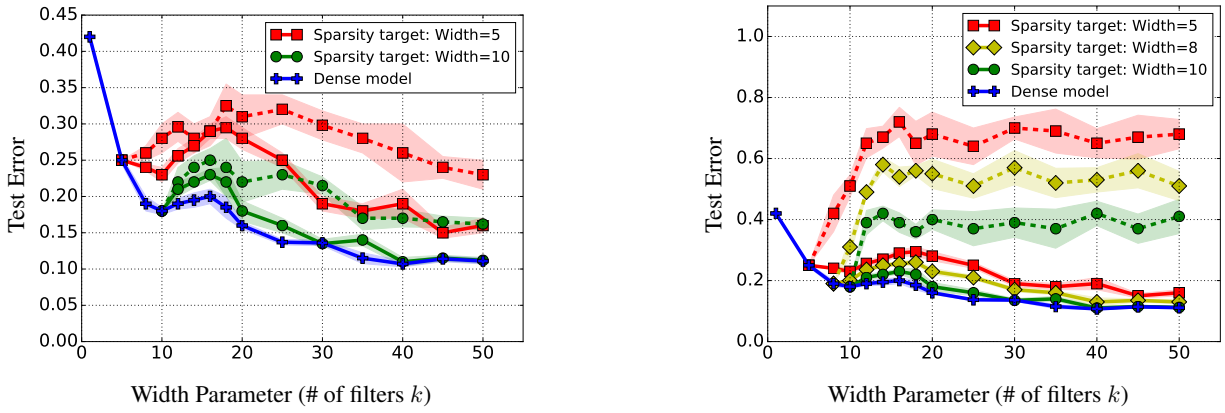
Thrampoulidis, C.; Oymak, S.; and Hassibi, B. 2015. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, 1683–1709.

Thrampoulidis, C.; Xu, W.; and Hassibi, B. 2018. Symbol error rate performance of box-relaxation decoders in massive mimo. *IEEE Transactions on Signal Processing* 66(13): 3377–3392.

Tsigler, A.; and Bartlett, P. L. 2020. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286* .

Wang, C.; Zhang, G.; and Grosse, R. 2020. Picking winning tickets before training by preserving gradient flow. *arXiv preprint arXiv:2002.07376* .

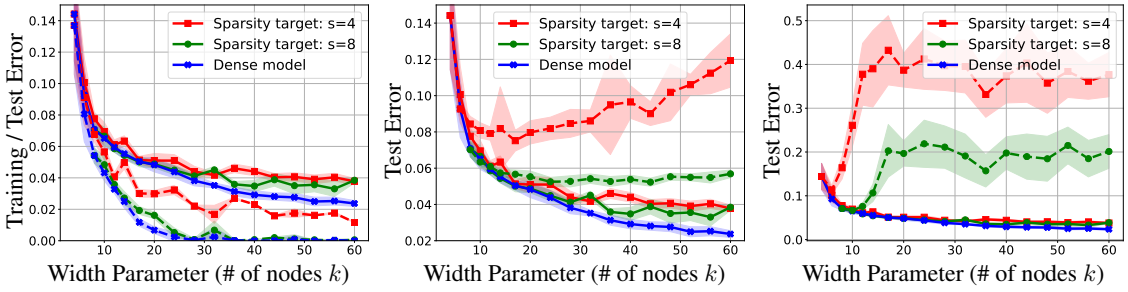
Zhou, H.; Lan, J.; Liu, R.; and Yosinski, J. 2019. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems*, 3592–3602.



(a) Magnitude-based and randomly pruned models.

(b) With and without retraining pruned models.

Figure 5: We train and prune ResNet-20 models on CIFAR-10 and also add randomly pruning and non-retraining curves. Solid lines here are exactly the same as Figure 1 which show test errors of dense and s-sparse models. Dotted lines are test errors of random-based pruned models in (a) and of magnitude-based pruned models however without retraining in (b).



(a) Dense and magnitude-based pruned and retrained models.

(b) Magnitude-based and randomly pruned models.

(c) With and without retraining pruned models.

Figure 6: Here we use the simplest neural network consisting of two fully-connected layers to train MNIST with various width and sparsity targets. In this architecture the width parameter equivalent to the number of hidden layer nodes controls model size. Same as Figure 1 the blue line corresponds to training dense models with width k. As for the red and green lines we choose 4- and 8-width models as base models respectively and prune k-width dense model to corresponding sparsity targets. (a) shows both training and test errors after magnitude-based pruning and retraining. In (b) solid and dotted lines are test errors of magnitude- and random-based pruned models with retraining. To verify the importance of retraining in neural networks we present the test errors of magnitude-based models without retraining in (c).

Organization of the Supplementary Material

The supplementary material (SM) is organized as follows.

1. In Section A we provide additional experiments on neural networks and linear models further supporting the main results.
2. In Section B we prove our main Theorem 1.
3. Section C provides our analysis and results on magnitude and Hessian pruning.
4. Section D provides further supporting technical results used in our proofs.
5. Section E provides our asymptotic analysis of overdetermined problems complementing our main results on overparameterized problems.
6. In Section F we prove Lemma 1.

A Further Experiments and Intuitions

A.1 Further discussion and experiments on CIFAR-10

First, we provide further discussion on Figure 1. Recall that, in this figure, we apply *train→prune→retrain* to obtain the sparse neural networks. The test error and the training errors for the sparse models are evaluated at the end of the retraining phase. Thus, it is rather surprising that sparse models manage to achieve zero training error around the same width parameter k as the dense model because while parameter count of the dense model increases in k, it is fixed for sparse models.

Secondly, we complement Figure 1 with two additional experiments. The first experiment assesses the benefit of pruning compared to using a random nonzero pattern with the same sparsity. The second experiment assesses the benefit of retraining by comparing the curves in Fig 1 with the test errors obtained without retraining. These two experiments are shown in Figure 5a and 5b and all of them are trained over same dataset and configured as given in Section 3.3. Instead of applying magnitude-based pruning, dotted red and green lines in Fig 5a are sparse models over 5 or 10 sparsity targets, pruned randomly to achieve same number of nonzeros as magnitude-based pruning strategy. Although the double descent phenomenon and downward trend are still present on the dotted lines, the performance is worse than magnitude-base pruning method. Fig. 5b shows how retraining benefits pruning ResNet-20 models. The results agree with our intuition that the non-retrained (dotted) lines achieve much bigger test error than retrained (solid) lines and overparameterization does not help in improving performance.

A.2 MNIST Experiments with two layers

In Figure 6 we train the simplest neural model with only 2 fully-connected layers over MNIST with various number of nodes to explore properties of magnitude-based pruning, random pruning and non-retraining on simple neural networks. Here, the number of nodes is equivalent to the width of the model, which directly controls the model size. Same as in Section 3.3, we select an s-width model as base model and prune trained dense models to the same sparsity. All experiments are trained with Adam optimization, 0.001 learning rate and 200 epochs under MNIST. Solid red, green and blue lines in Figure 6 correspond to test error of 4-, 8-sparsity targets and dense models. In Figure 6a dotted lines are training errors of dense and s-sparse models. As the width k grows, the training and test error decrease for all dense and sparse models. The behavior is similar to Figure 1 and training larger models benefit pruned-model accuracy however double descent is not really visible. We suspect that this may be because of the simpler nature of the MNIST dataset and LeNet architecture compared to the CIFAR10 dataset and ResNet-20 model. In Figure 6b, dotted lines apply randomly pruning. Different to magnitude pruning, where training bigger models and then pruning results in better performance, randomly pruning hurts when the sparsity level s/k is relatively low. This is because under magnitude-based pruning, we can identity most of optimal entries of weights from trained dense model and retraining with these entries achieves lower errors. In constrast, random pruning learns nothing from the trained model and as the sparsity level decreases, the probability that random operator selects the limited optimal entries by chance also decreases, leading to worse performance. Dotted lines in Figure 6c show test errors of sparse models before retraining which educes the same conclusion in Section A.1, that is retraining is crucial to improve the performance in neural networks.

A.3 Experiments on LGP

In Figure 7, we carry out the identical experiments as in Figure 3. The difference is that we display two more figures which are the retraining curves for Magnitude- and Hessian-based pruning strategies shown in purple and yellow lines respectively. Figure 7a is the counterpart of Figure 3a and Figure 7b is the counterpart of Figure 3b. The main message in these experiments is that *retraining hurts the performance*. This performance degradation is more emphasized in the overparameterized regime. Specifically, both retrained versions of Magnitude and Hessian pruning $\hat{\boldsymbol{\beta}}^{RT,M}$ and $\hat{\boldsymbol{\beta}}^{RT,H}$ perform consistently worse compared to their pruning-only counterparts $\hat{\boldsymbol{\beta}}^{M}$ and $\hat{\boldsymbol{\beta}}^{H}$. Observe that, the only regime where retraining outperforms the pruning-only approach is at the peak of double descent. This is the region where pruning-only risk diverges to infinity whereas retraining attains finite risk. This is because the retraining stage solves a well-conditioned problem and avoids the ill-conditioning occurring at $n = k$. Recall that, in light of Lemma 1, unlike the rank-one covariance case, retraining hurts because covariance is diagonal; thus, features are uncorrelated and do not have overlapping predictions.

B Proofs for overparameterized least-squares

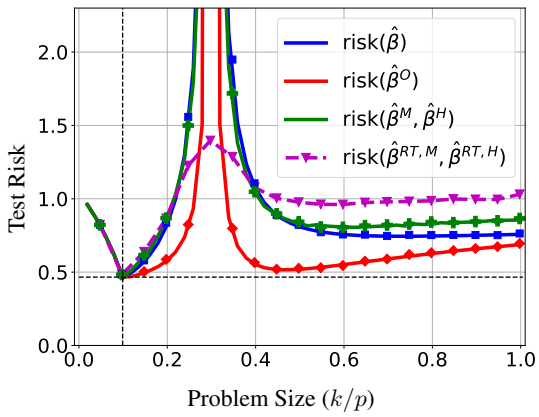
In this section, we assume the linear Gaussian problem (LGP) of Definition 1, the overparameterized regime $k = p > n$ and the min-norm model $\hat{\boldsymbol{\beta}}$ of (7). We prove Theorem 1 that derives the asymptotic DC of $\hat{\boldsymbol{\beta}}$ and we show how this leads to sharp formulae for the risk of the Magnitude- and Hessian-pruned models.

B.1 Notation and Assumptions

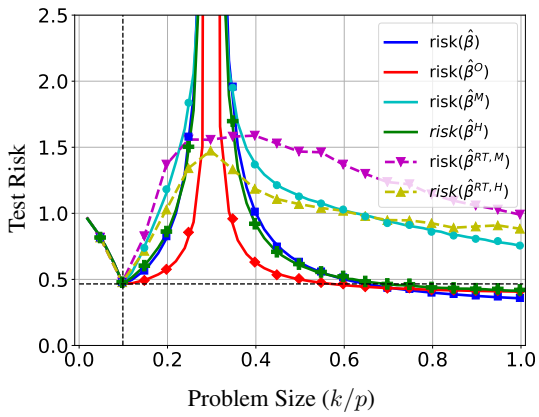
For the reader's convenience, we recall some necessary notation and assumptions from Section 5. We say that a function $f : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order k, denoted $f \in \mathrm{PL(k)}$, if there is a constant $L > 0$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$, $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L(1 + \|\boldsymbol{x}\|_{\ell_2}^{k-1} + \|\boldsymbol{y}\|_{\ell_2}^{k-1})\|\boldsymbol{x} - \boldsymbol{y}\|_2$ (See also Section D). We say that a sequence of probability distributions ν_p on $\mathbb{R}^m$ converges in W_k to ν, written $\nu_p \overset{W_k}{\Longrightarrow} \nu$, if $W_k(\nu_p, \nu) \to 0$ as $p \to \infty$. An equivalent definition is that, for any $f \in \mathrm{PL(k)}$, $\lim_{p\to\infty} \mathbb{E}\, f(X_p) = \mathbb{E}\, f(X)$, where expectation is with respect to $X_p \sim \nu_p$ and $X \sim \nu$ (e.g., (Montanari and Venkataramanan 2017)). Finally, recall that a sequence of probability distributions ν_n on $\mathbb{R}^m$ *converges weakly* to ν, if for any bounded Lipschitz function f: $\lim_{p\to\infty} \mathbb{E}\, f(X_p) = \mathbb{E}\, f(X)$, where expectation is with respect to $X_p \sim \nu_p$ and $X \sim \nu$. Throughout, we use C, C', c, c' to denote absolute constants (not depending on n, p) whose value might change from line to line.

We focus on a double asymptotic regime where:

$$n, p, s \to \infty \text{ at fixed overparameterization ratio } \kappa := p/n > 1 \text{ and sparsity level } \alpha := s/p \in (0, 1).$$



(a) Identity covariance, spiked latent weights.



(b) Spiked covariance, identical latent weights.

Figure 7: Same as Figure 3 however retraining curves are included. Our asymptotic prediction for various pruning strategies in linear gaussian models with $s/p = 0.1$ and $n/p = 0.3$. We solve ERM using the first k features and then prune to obtain an s-sparse model. The vertical dashed line shows the $k = s$ point. The horizontal dashed line highlights the minimum risk among all underparameterized solutions including retraining. Retraining curves are not displayed.

For a sequence of random variables $\mathcal{X}_p$ that converge in probability to some constant c in the limit of Assumption 1 below, we write $\mathcal{X}_p \xrightarrow{P} c$. For a sequence of event $\mathcal{E}_p$ for which $\lim_{p\to} \mathbb{P}(\mathcal{E}_p) = 1$, we say that $\mathcal{E}_p$ occurs *with probability approaching 1*. For this, we will often use the shorthand "wpa 1".

Next, we recall the set of assumption under which our analysis applies:

Assumption 2 *The covariance matrix $\mathbf{\Sigma}$ is diagonal and there exist constants $\Sigma_{\min}, \Sigma_{\max} \in (0, \infty)$ such that: $\Sigma_{\min} \le \mathbf{\Sigma}_{i,i} \le \Sigma_{\max}$, for all $i \in [p]$.*

Assumption 3 *The joint empirical distribution of $\{(\mathbf{\Sigma}_{i,i}, \sqrt{p}\boldsymbol{\beta}^\star_i)\}_{i\in[p]}$ converges in Wasserstein-k distance to a probability distribution μ on $\mathbb{R}_{>0} \times \mathbb{R}$ for some $k \ge 4$. That is $\frac{1}{p}\sum_{i\in[p]} \delta_{(\mathbf{\Sigma}_{i,i}, \sqrt{p}\boldsymbol{\beta}^\star_i)} \overset{W_k}{\Longrightarrow} \mu$.*

We remark that Assumption 3 above implies (see (Bayati and Montanari 2011, Lem. 4) and (Javanmard and Montanari 2013, Lem. A3)) that for any pseudo-Lipschitz function $\psi : \mathbb{R}^2 \to \mathbb{R}$ of order 4, i.e., $\psi \in \mathrm{PL}(4)$:

$$\frac{1}{p}\sum_{i=1}^{p} \psi(\mathbf{\Sigma}_{i,i}, \sqrt{p}\boldsymbol{\beta}^\star_i) \xrightarrow{P} \mathbb{E}_{(\Lambda,B)\sim\mu}\left[\psi(\Lambda, B)\right].$$

B.2 Asymptotic distribution and risk characterizations

In this section, we prove our main result Theorem 1. Recall that $\hat{\boldsymbol{\beta}}$ is the min-norm solution. Since the distribution of $\hat{\boldsymbol{\beta}}$ depends on the problem dimensions (as it is a function of $\boldsymbol{X}, \boldsymbol{y}$), when necessary, we will use $\hat{\boldsymbol{\beta}}_n$ notation to make its dimension dependence explicit. Let $\hat{\boldsymbol{\beta}}^P = \mathcal{P}(\hat{\boldsymbol{\beta}})$ be a pruned version of the min-norm solution $\hat{\boldsymbol{\beta}}$. Recall from Section 5.2, that the first crucial step in characterizing the risk $\mathcal{L}(\hat{\boldsymbol{\beta}}^P)$ is studying the risk $\mathcal{L}(\hat{\boldsymbol{\beta}}^{\mathcal{T}}_t)$ of a threshold-based pruned vector.

To keep things slightly more general, consider $\hat{\boldsymbol{\beta}}^g$ defined such that $\sqrt{p}\hat{\boldsymbol{\beta}}^g = g(\sqrt{p}\hat{\boldsymbol{\beta}})$, where g is a Lipschitz function acting entry-wise on $\hat{\boldsymbol{\beta}}$ (for example, g can be the (arbitrarily close Lipschitz approximation of the) thresholding operator $\mathcal{T}_t$ of Section 5.2). Then, the risk of $\hat{\boldsymbol{\beta}}^g$ can be written as

$$\begin{aligned}
\mathcal{L}(\hat{\boldsymbol{\beta}}^g) &= \mathbb{E}_{\mathcal{D}}[(\boldsymbol{x}^T(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^g) + \sigma z)^2] = \sigma^2 + (\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^g)^T\mathbf{\Sigma}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^g) \\
&= \sigma^2 + \frac{1}{p}\sum_{i=1}^{p} \mathbf{\Sigma}_{i,i}\big(\sqrt{p}\boldsymbol{\beta}^\star_i - g(\sqrt{p}\hat{\boldsymbol{\beta}}_i)\big)^2 \\
&=: \sigma^2 + \frac{1}{p}\sum_{i=1}^{p} f\big(\sqrt{p}\hat{\boldsymbol{\beta}}_i, \sqrt{p}\boldsymbol{\beta}^\star_i, \mathbf{\Sigma}_{i,i}\big),
\end{aligned} \tag{16}$$

where in the last line, we defined the function f as $f = f_{\mathcal{L}} \in \mathcal{F}_{\mathcal{L}} \subset \mathcal{F}$ given by

$$f_{\mathcal{L}}(x, y, z) := z(y - g(x))^2 \quad \text{where } g \text{ is Lipschitz.}$$

Here, recall the definition of the families $\mathcal{F}_{\mathcal{L}}$ in (12) and $\mathcal{F}$ in (11).

The following theorem establishes the asymptotic limit of (16). For the reader's convenience, we repeat the notation introduced in Definition 3. Let random variables $(\Lambda, B) \sim \mu$ (where μ is defined in Assumption 3) and fix $\kappa > 1$. Define parameter ξ as the unique positive solution to the following equation

$$\mathbb{E}_\mu \left[\left(1 + (\xi \cdot \Lambda)^{-1}\right)^{-1} \right] = \kappa^{-1} .$$

Further define the positive parameter γ as follows:

$$\gamma := \left(\sigma^2 + \mathbb{E}_\mu \left[\frac{B^2 \Lambda}{(1+\xi\Lambda)^2}\right]\right) \Big/ \left(1 - \kappa\, \mathbb{E}_\mu \left[\frac{1}{(1+(\xi\Lambda)^{-1})^2}\right]\right).$$

With these and $H \sim \mathcal{N}(0,1)$, define the random variable

$$X_{\kappa,\sigma^2} := X_{\kappa,\sigma^2}(\Lambda, B, H) := \left(1 - \frac{1}{1+\xi\Lambda}\right) B + \sqrt{\kappa} \frac{\sqrt{\gamma}\, \Lambda^{-1/2}}{1 + (\xi\Lambda)^{-1}} H,$$

and let Π_{κ,σ^2} be its distribution.

Theorem 1 (Asymptotic DC – Overparameterized LGP) *Fix $\kappa > 1$ and suppose Assumptions 2 and 3 hold. Recall the solution $\hat{\boldsymbol{\beta}}$ from* (7) *and let*

$$\hat{\Pi}_n(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) := \frac{1}{p} \sum_{i=1}^{p} \delta_{(\sqrt{p}\hat{\boldsymbol{\beta}}_i, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i})}$$

be the joint empirical distribution of $(\sqrt{p}\hat{\boldsymbol{\beta}}, \sqrt{p}\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})$. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a function in $\mathcal{F}$ defined in (11)*. We have that*

$$\frac{1}{p} \sum_{i=1}^{p} f(\sqrt{p}\hat{\boldsymbol{\beta}}_i, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}) \xrightarrow{P} \mathbb{E}\left[f(X_{\kappa,\sigma^2}, B, \Lambda)\right]. \tag{13}$$

Before we prove the theorem, let us show how it immediately leads to a sharp prediction of the risk behavior. Indeed, a direct application of (13) for $f = f_{\mathcal{L}}$ to (16) shows that

$$\mathcal{L}(\hat{\boldsymbol{\beta}}^g) \xrightarrow{P} \sigma^2 + \mathbb{E}_{(\Lambda,B,H)\sim\mu\otimes\mathcal{N}(0,1)} \left[f_{\mathcal{L}}(X_{\kappa,\sigma^2}, B, \Lambda)\right] = \sigma^2 + \mathbb{E}_{(\Lambda,B,H)\sim\mu\otimes\mathcal{N}(0,1)} \left[\Sigma\left(B - g(X_{\kappa,\sigma^2})\right)^2\right]. \tag{17}$$

We further remark on the following two consequences of Theorem 1.

First, since (13) holds for any $\mathrm{PL}(2)$ function, we have essentially shown that $\hat{\Pi}_n(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma})$ converges in Wasserstein-2 distance to $\Pi_{\kappa,\sigma^2} \otimes \mu$, where recall that Π_{κ,σ^2} is the distribution of the random variable X_{κ,σ^2}.

Second, the theorem implies that the empirical distribution of $\sqrt{p}\hat{\boldsymbol{\beta}}_n$ converges weakly to Π_{κ,σ^2}. To see this, apply (13) for the $\mathrm{PL}(2)$ function $f(x,y,z) = \psi(x)$ where $\psi : \mathbb{R} \to \mathbb{R}$ is a bounded Lipschitz test function.

B.3 Proof of Theorem 1

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ have zero-mean and normally distributed rows with a diagonal covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$. Given a ground-truth vector $\boldsymbol{\beta}^\star$ and labels $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \sigma\boldsymbol{z},\ \boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_n)$, we consider the least-squares problem subject to the minimum Euclidian norm constraint (as $\kappa = p/n > 1$) given by

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|_{\ell_2}^2 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}. \tag{18}$$

It is more convenient to work with the following change of variable:

$$\boldsymbol{w} := \sqrt{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star). \tag{19}$$

With this, the optimization problem in (7) can be rewritten as

$$\Phi(\boldsymbol{X}) = \min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}^\star\|_{\ell_2}^2 \quad \text{subject to} \quad \bar{\boldsymbol{X}}\boldsymbol{w} = \sigma\boldsymbol{z}, \tag{20}$$

where we set $\bar{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. First, using standard arguments, we show that the solution of (20) is bounded. Hence, we can constraint the optimization in a sufficiently large compact set without loss of generality.

Lemma 2 (Boundedness of the solution) *Let $\widehat{\boldsymbol{w}}_n := \widehat{\boldsymbol{w}}_n(\boldsymbol{X}, \boldsymbol{z})$ be the minimizer in* (20)*. Then, with probability approaching 1, it holds that $\widehat{\boldsymbol{w}}_n \in \mathcal{B}$, where*

$$\mathcal{B} := \{\boldsymbol{w} \,|\, \|\boldsymbol{w}\|_2 \le B_+\}, \qquad B_+ := 5\sqrt{\frac{\Sigma_{\max}}{\Sigma_{\min}}} \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} (\sqrt{\Sigma_{\max}\,\mathbb{E}\left[B^2\right]} + \sigma).$$

Proof First, we show that the min-norm solution $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}$ of (18) is bounded. Note that $\kappa > 1$, thus $\boldsymbol{X}\boldsymbol{X}^T$ is invertible wpa 1. We have,

$$\|\hat{\boldsymbol{\beta}}_n\|_{\ell_2}^2 = \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y} \leq \frac{\|\boldsymbol{y}\|_{\ell_2}^2}{\lambda_{\min}(\boldsymbol{X}\boldsymbol{X}^T)} = \frac{\|\boldsymbol{y}\|_{\ell_2}^2}{\lambda_{\min}(\bar{\boldsymbol{X}}\boldsymbol{\Sigma}\bar{\boldsymbol{X}}^T)} \leq \frac{\|\boldsymbol{y}\|_{\ell_2}^2}{\lambda_{\min}(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T)\,\Sigma_{\min}} = \frac{\|\boldsymbol{y}\|_{\ell_2}^2}{\sigma_{\min}^2(\bar{\boldsymbol{X}})\,\Sigma_{\min}}. \tag{21}$$

But, wpa 1, $\sigma_{\min}(\bar{\boldsymbol{X}})/\sqrt{n} \geq \frac{1}{2}\left(\sqrt{\kappa}-1\right)$. Furthermore, $\|\boldsymbol{y}\|_2 \leq \|\bar{\boldsymbol{X}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^\star\|_2 + \sigma\|\boldsymbol{z}\|_2 \leq \sigma_{\max}(\bar{\boldsymbol{X}})\sqrt{\Sigma_{\max}}\|\boldsymbol{\beta}^\star\|_2 + \sigma\|\boldsymbol{z}\|_2$. Hence, wpa 1,

$$\|\boldsymbol{y}\|_2/\sqrt{n} \leq 2(\sqrt{\kappa}+1)\sqrt{\Sigma_{\max}}\sqrt{\mathbb{E}\left[B^2\right]} + 2\sigma,$$

where we used the facts that wpa 1: $\|z\|_2/\sqrt{n} \xrightarrow{P} 1$, $\sigma_{\max}(\bar{\boldsymbol{X}}) < \sqrt{2n}(\sqrt{\kappa}+1)$ and by Assumption 3:

$$\|\boldsymbol{\beta}^\star\|_2^2 = \frac{1}{p}\sum_{i=1}^{p}(\sqrt{p}\boldsymbol{\beta}_i^\star)^2 \xrightarrow{P} \mathbb{E}\left[B^2\right].$$

Put together in (21), shows that

$$\|\hat{\boldsymbol{\beta}}_n\|_{\ell_2} < \frac{2(\sqrt{\kappa}+1)\sqrt{\Sigma_{\max}}\sqrt{\mathbb{E}\left[B^2\right]} + 2\sigma}{\sqrt{\Sigma_{\min}}(\sqrt{\kappa}-1)/2} =: \tilde{B}_+. \tag{22}$$

Recalling that $\widehat{\boldsymbol{w}}_n = \sqrt{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}}_n - \sqrt{\boldsymbol{\Sigma}}\boldsymbol{\beta}^\star$, we conclude, as desired, that wpa 1, $\|\widehat{\boldsymbol{w}}_n\|_{\ell_2} \leq \sqrt{\Sigma_{\max}}\tilde{B}_+ + \sqrt{\Sigma_{\max}}\sqrt{\mathbb{E}[B^2]} \leq B_+$. ■

Lemma 2 implies that nothing changes in (20) if we further constrain $\boldsymbol{w} \in \mathcal{B}$ in (20). Henceforth, with some abuse of notation, we let

$$\Phi(\boldsymbol{X}) := \min_{\boldsymbol{w}\in\mathcal{B}} \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}^\star\|_{\ell_2}^2 \quad \text{subject to} \quad \bar{\boldsymbol{X}}\boldsymbol{w} = \sigma\boldsymbol{z}, \tag{23}$$

Next, in order to analyze the primary optimization (PO) problem in (23) in apply the CGMT (Thrampoulidis, Oymak, and Hassibi 2015; Thrampoulidis, Abbasi, and Hassibi 2018). Specifically, we use the constrained formulation of the CGMT given by Theorem 2. Specifically, the auxiliary problem (AO) corresponding to (23) takes the following form with $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_n)$, $\boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{I}_p)$, $h \sim \mathcal{N}(0,1)$:

$$\phi(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{B}} \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}^\star\|_{\ell_2}^2 \quad \text{subject to} \quad \|\boldsymbol{g}\|_{\ell_2}\|\boldsymbol{w}\;\sigma\|_{\ell_2} \leq \boldsymbol{h}^T\boldsymbol{w} + \sigma h. \tag{24}$$

We will prove the following techincal result about the AO problem.

Lemma 3 (Properties of the AO – Overparameterized regime) *Let the assumptions of Theorem 1 hold. Let $\phi_n = \phi(\boldsymbol{g},\boldsymbol{h})$ be the optimal cost of the minimization in* (24)*. Define $\bar{\phi}$ as the optimal cost of the following deterministic min-max problem*

$$\bar{\phi} := \max_{u\geq 0}\min_{\tau>0}\; \mathcal{D}(u,\tau) := \frac{1}{2}\left(u\tau + \frac{u\sigma^2}{\tau} - u^2\kappa\;\mathbb{E}\left[\frac{1}{\Lambda^{-1}+\frac{u}{\tau}}\right] - \mathbb{E}\left[\frac{B^2}{1+\frac{u}{\tau}\Lambda}\right]\right). \tag{25}$$

The following statements are true.
(i). The AO minimization in (24) *is $\frac{1}{\Sigma_{\max}}$-strongly convex and has a unique minimizer $\hat{\boldsymbol{w}}_n^{\mathrm{AO}} := \hat{\boldsymbol{w}}_n^{\mathrm{AO}}(\boldsymbol{g},\boldsymbol{h})$.*
(ii). In the limit of $n, p \to \infty, p/n = \kappa$, it holds that $\phi(\boldsymbol{g},\boldsymbol{h}) \xrightarrow{P} \bar{\phi}$, i.e., for any $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left(|\phi(\boldsymbol{g},\boldsymbol{h}) - \bar{\phi}| > \varepsilon\right) = 0.$$

(iii). The max-min optimization in (25) *has a unique saddle point (u_*, τ_*) satisfying the following:*

$$u_*/\tau_* = \xi \quad \textit{and} \quad \tau_* = \gamma,$$

where ξ, γ are defined in Definition 3.
(iv). Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a function in PL(3)*. Let $\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}} = \hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}(\boldsymbol{g},\boldsymbol{h}) = \boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{w}}_n^{\mathrm{AO}} + \boldsymbol{\beta}^\star$. Then,*

$$\frac{1}{p}\sum_{i=1}^{p} f\left(\sqrt{p}\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}, \sqrt{p}\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}\right) \xrightarrow{P} \mathbb{E}_{(B,\Lambda,H)\sim\mu\otimes\mathcal{N}(0,1)}\left[f\left(X_{\kappa,\sigma^2}(B,\Lambda,H), B, \Lambda\right)\right].$$

In particular, this holds for all functions $f \in \mathcal{F}$ defined in (11)*.*
(v). The empirical distribution of $\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}$ converges weakly to the measure of X_{κ,σ^2}, and also, for some absolute constant $C > 0$:

$$\|\hat{\boldsymbol{\beta}}^{\mathrm{AO}}\|_{\ell_2}^2 < C \quad \textit{wpa 1.} \tag{26}$$

We prove Lemma 3 in Section B.4. We remark that Assumption 3 on W_4-convergence of the joint empirical distribution of $\{(\boldsymbol{\Sigma}_{i,i}, \sqrt{p}\boldsymbol{\beta}_i^\star)\}_{i\in[p]}$ is required in the proof of the statement (iv) above. More generally if W_k-convergence is known for some integer k, then statement (iv) above holds for test functions $f \in \mathrm{PL(k-1)}$. This is the first place in the proof of Theorem 1, where we use the assumption $f \in \mathcal{F}$; indeed, we show in Lemma 4 that $\mathcal{F}_{\mathcal{L}} \subset \mathcal{F} \subset \mathrm{PL}(3)$. The second part is in proving the perturbation result in (33) below. Unlike the former, when proving the perturbation result, the requirement $f \in \mathcal{F}$ cannot be relaxed (e.g. to $f \in \mathrm{PL(k-1)}$) by simply increasing the order of W_k-convergence in Assumption 3.

Finalizing the proof of Theorem 1: Here, we show how Lemma 3 leads to the proof of Theorem 1 when combined with the CGMT framework (Thrampoulidis, Oymak, and Hassibi 2015; Thrampoulidis, Abbasi, and Hassibi 2018).

Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a function in $\mathcal{F}$, where $\mathcal{F}$ was defined in (11). For convenience, define

$$F_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) := \frac{1}{p}\sum_{i=1}^{p} f\left(\sqrt{p}\hat{\boldsymbol{\beta}}_{n,i}, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{ii}\right) \quad \text{and} \quad \alpha_* := \mathbb{E}_\mu\left[f(X_{\kappa,\sigma^2}(\Lambda, B, H), B, \Lambda)\right].$$

Fix any $\varepsilon > 0$ and define the set

$$\mathcal{S} = \mathcal{S}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) = \left\{\boldsymbol{w} = \sqrt{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star) \in \mathcal{B} \;\middle|\; |F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - \alpha_*| \geq 2\varepsilon\right\}. \tag{27}$$

With this definition, observe that, it suffices to prove that the solution $\hat{\boldsymbol{w}}_n = \sqrt{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star)$ of the PO in (18) satisfies $\hat{\boldsymbol{w}}_n \notin \mathcal{S}$ wpa 1.

To prove the desired, we need to consider the "perturbed" PO and AO problems (compare to (20) and (24)) as:

$$\Phi_S(\boldsymbol{X}) = \min_{\boldsymbol{w}\in\mathcal{S}} \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}^\star\|_{\ell_2}^2 \quad \text{subject to} \quad \bar{\boldsymbol{X}}\boldsymbol{w} = \sigma\boldsymbol{z}, \tag{28}$$

and

$$\phi_S(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}} \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}^\star\|_{\ell_2}^2 \quad \text{subject to} \quad \|\boldsymbol{g}\|_{\ell_2}\|\boldsymbol{w}\,\sigma\|_{\ell_2} \leq \boldsymbol{h}^T\boldsymbol{w} + \sigma h. \tag{29}$$

Recall here, that $\bar{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_n)$, $\boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{I}_p)$, $h \sim \mathcal{N}(0,1)$ and we have used the change of variables $\boldsymbol{w} := \sqrt{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)$ for convenience.

Using (Thrampoulidis, Abbasi, and Hassibi 2018, Theorem 6.1(iii)) it suffices to find costants $\bar{\phi}, \bar{\phi}_S$ and $\eta > 0$ such that the following three conditions hold:

1. $\bar{\phi}_S \geq \bar{\phi} + 3\eta$,
2. $\phi(\boldsymbol{g}, \boldsymbol{h}) \leq \bar{\phi} + \eta$, with probability approaching 1,
3. $\phi_S(\boldsymbol{g}, \boldsymbol{h}) \geq \bar{\phi}_S - \eta$, with probability approaching 1.

In what follows, we explicitly find $\bar{\phi}, \bar{\phi}_S, \eta$ such that the three conditions above hold.

Satisfying Condition 2: Recall the deterministic min-max optimization in (25). Choose $\bar{\phi} = \mathcal{D}(u_*, \tau_*)$ be the optimal cost of this optimization. From Lemma 3(ii), $\phi(\boldsymbol{g}, \boldsymbol{h}) \xrightarrow{P} \bar{\phi}$. Thus, for any $\eta > 0$, with probability approaching 1:

$$\bar{\phi} + \eta \geq \phi(\boldsymbol{g}, \boldsymbol{h}) \geq \bar{\phi} - \eta. \tag{30}$$

Clearly then, Condition 2 above holds for any $\eta > 0$.

Satisfying Condition 3: Next, we will show that the third condition holds for appropriate $\bar{\phi}$. Let $\hat{\boldsymbol{w}}_n^{\mathrm{AO}} = \hat{\boldsymbol{w}}_n^{\mathrm{AO}}(\boldsymbol{g}, \boldsymbol{h})$ be the unique minimizer of (24) as per Lemma 3(i), i.e., $\frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{w}}_n^{\mathrm{AO}} + \boldsymbol{\beta}\|_{\ell_2}^2 = \phi(\boldsymbol{g}, \boldsymbol{h})$. Again from Lemma 3, the minimization in (24) is $1/\Sigma_{\max}$-strongly convex in $\boldsymbol{w}$. Here, $\Sigma_{\max}$ is the upper bound on the eigenvalues of $\boldsymbol{\Sigma}$ as per Assumption 3. Thus, for any $\tilde{\varepsilon} > 0$ and any feasible $\boldsymbol{w}$ the following holds (deterministically):

$$\frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}\|_{\ell_2}^2 \geq \phi(\boldsymbol{g}, \boldsymbol{h}) + \frac{\tilde{\epsilon}^2}{2\Sigma_{\max}}, \text{ provided that } \|\boldsymbol{w} - \hat{\boldsymbol{w}}_n^{\mathrm{AO}}\|_2 \geq \tilde{\epsilon}. \tag{31}$$

Now, we argue that wpa 1,

$$\text{for all } \boldsymbol{w} \in \mathcal{S} \text{ it holds that } \|\boldsymbol{w} - \hat{\boldsymbol{w}}_n^{\mathrm{AO}}\|_2 \geq \tilde{\varepsilon}, \tag{32}$$

for an appropriate value of a constant $\tilde{\varepsilon} > 0$.

Consider any $\boldsymbol{w} \in \mathcal{S}$.

First, by definition in (27), for $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}^\star$ we have that

$$|F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - \alpha_*| \geq 2\varepsilon.$$

Second, by Lemma 3(iv), with probability approaching 1,

$$|F(\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - \alpha_*| \le \epsilon.$$

Third, we will show that wpa 1, there exists universal constant $C > 0$ such that

$$|F_n(\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| \le C\|\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}} - \boldsymbol{\beta}\|_2. \tag{33}$$

Before proving (33), let us argue how combining the above three displays shows the desired. Indeed, in that case, wpa 1,

$$\begin{aligned}
2\varepsilon \le |F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - \alpha_*| &\le |F_n(\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| + |F_n(\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - \alpha_*| \\
&\le \epsilon + C\,\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}\|_2. \\
&\implies \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}\|_2 \ge \varepsilon/C =: \hat{\varepsilon} \\
&\implies \|\boldsymbol{w} - \hat{\boldsymbol{w}}_n^{\mathrm{AO}}\|_2 \ge \hat{\varepsilon}\sqrt{\Sigma_{\min}} =: \tilde{\varepsilon}.
\end{aligned}$$

In the last line above, we recalled that $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}^\star$ and $\boldsymbol{\Sigma}_{i,i} \ge \Sigma_{\min},\ i \in [p]$ by Assumption 3. This proves (32).

Next, combining (32) and (31), we find that wpa 1, $\frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}\|_{\ell_2}^2 \ge \phi(\boldsymbol{g}, \boldsymbol{h}) + \frac{\tilde{\epsilon}^2}{2\Sigma_{\max}},$ for all $\boldsymbol{w} \in \mathcal{S}$. Thus,

$$\phi_S(\boldsymbol{g}, \boldsymbol{h}) \ge \phi(\boldsymbol{g}, \boldsymbol{h}) + \frac{\tilde{\epsilon}^2}{2\Sigma_{\max}}.$$

When combined with (30), this shows that

$$\phi_S(\boldsymbol{g}, \boldsymbol{h}) \ge \bar{\phi} + \frac{\tilde{\epsilon}^2}{2\Sigma_{\max}} - \eta. \tag{34}$$

Thus, choosing $\bar{\phi}_S = \bar{\phi} + \frac{\tilde{\epsilon}^2}{2\Sigma_{\max}}$ proves the Condition 3 above.

Perturbation analysis via Pseudo-Lipschitzness (Proof of (33)**).** To complete the proof, let us now show (33). Henceforth, C is used to denote a universal constant whose value can change from line to line. Recall that $f \in \mathcal{F}$ where $\mathcal{F} : \mathbb{R}^2 \times \mathcal{Z} \to \mathbb{R}$ is the set of $\mathrm{PL}(3)$ functions such that $f(\cdot, \cdot, z)$ is $\mathrm{PL}(2)$ for all $z \in \mathcal{Z}$. Suppose that the $\mathrm{PL}(2)$ constant of $f(\cdot, \cdot, z)$ is upper bounded over $z \in \mathcal{Z}$ by some $C > 0$. We also let C change from line to line for notational simplicity. Then, we have the following chain of inequalities:

$$\begin{aligned}
|F_n(\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}(\boldsymbol{g}, \boldsymbol{h}), \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| &= \frac{1}{p}\sum_{i=1}^p |f(\sqrt{p}\hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}}, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}) - f(\sqrt{p}\boldsymbol{\beta}_i, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i})| \\
&\le \frac{C}{p}\sum_{i=1}^p (1 + \|\sqrt{p}[\boldsymbol{\beta}_i^\star, \hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}}]\|_2 + \|\sqrt{p}[\boldsymbol{\beta}_i^\star, \boldsymbol{\beta}_i]\|_2)\sqrt{p}|\hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}} - \boldsymbol{\beta}_i| \\
&\le C\Big(1 + \frac{1}{\sqrt{p}}\big(\sum_{i=1}^p \|\sqrt{p}[\boldsymbol{\beta}_i^\star, \hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}}]\|_2^2\big)^{1/2} + \frac{1}{\sqrt{p}}\big(\sum_{i=1}^p \|\sqrt{p}[\boldsymbol{\beta}_i^\star, \boldsymbol{\beta}_i]\|_2^2\big)^{1/2}\Big)\|\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}} - \boldsymbol{\beta}\|_2 \\
&\le C\left(1 + \max\{\|\boldsymbol{\beta}^\star\|_2^2, \|\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}\|_2^2, \|\boldsymbol{\beta}\|_2^2\}^{1/2}\right)\|\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}} - \boldsymbol{\beta}\|_2.
\end{aligned} \tag{35}$$

In the second line above, we used the fact that $f(\cdot, \cdot, z)$ is $\mathrm{PL}(2)$. The third line follows by Cauchy-Schwartz inequality. Finally, in the last line, we used the elementary fact that $a + b + c \le 3\max\{a, b, c\}$ for $a = 2\sum_{i=1}^p (\boldsymbol{\beta}_i^\star)^2$ and $b = \sum_{i=1}^p (\hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}})^2$ and $c = \sum_{i=1}^p \boldsymbol{\beta}_i^2$.

Hence, it follows from (35) that in order to prove (33), we need to show boundedness of the following terms: $\|\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}\|_2$, $\|\boldsymbol{\beta}^\star\|_2$ and $\|\boldsymbol{\beta}\|_2$. By feasibility of $\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}$ and $\boldsymbol{\beta}$, we know that $\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}, \boldsymbol{\beta} \in \mathcal{B}$. Thus, the desired $\|\boldsymbol{\beta}\|_2 < \infty$ and $\|\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}\|_2 < \infty$ follow directly by Lemma 2 (Alternatively, for $\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}$ we conclude the desired by directly applying Lemma 3(v)). Finally, to prove $\|\boldsymbol{\beta}^\star\|_2 < \infty$, note that

$$\|\boldsymbol{\beta}^\star\|_2^2 = \frac{1}{p}\sum_{i=1}^p (\sqrt{p}\boldsymbol{\beta}_i^\star)^2,$$

which is bounded wpa 1 by Assumption 3, which implies bounded second moments of $\sqrt{p}\boldsymbol{\beta}^\star$. This completes the proof of (33), as desired.

Satisfying Condition 1: To prove Condition 1, we simply pick η to satisfy the following

$$\bar{\phi}_S > \bar{\phi} + 3\eta \ \Leftarrow\ \bar{\phi} + \frac{\tilde{\epsilon}^2}{2\Sigma_{\max}} - \eta \ge \bar{\phi} + 3\eta \ \Leftarrow\ \eta \le \frac{\tilde{\epsilon}^2}{8\Sigma_{\max}}.$$

This completes the proof of Theorem 1.

B.4 Proof of Lemma 3

Proof of (i). Strong convexity of the objective function in (24) is easily verified by the second derivative test and use of Assumption 3 that $\boldsymbol{\Sigma}_{i,i} \leq \Sigma_{\max},\ i \in [p]$. Uniqueness of the solution follows directly from strong convexity.

Proof of (ii). Using Lagrangian formulation, the solution $\hat{\boldsymbol{w}}_n^{\mathrm{AO}}$ to (24) is the same as the solution to the following:

$$\left(\hat{\boldsymbol{w}}_n^{\mathrm{AO}}, u_n\right) := \arg\min_{\boldsymbol{w}\in\mathcal{B}} \max_{u\geq 0}\ \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}^\star\|_{\ell_2}^2 + u\left(\sqrt{\|\boldsymbol{w}\|_{\ell_2}^2 + \sigma^2}\|\bar{\boldsymbol{g}}\|_{\ell_2} - \sqrt{\kappa}\,\bar{\boldsymbol{h}}^T\boldsymbol{w} + \frac{\sigma h}{\sqrt{n}}\right) \tag{36}$$

where we have: (i) set $\bar{\boldsymbol{g}} := \boldsymbol{g}/\sqrt{n}$ and $\bar{\boldsymbol{h}} := \boldsymbol{h}/\sqrt{p}$; (ii) recalled that $p/n = \kappa$; and, (iii) used $\left(\hat{\boldsymbol{w}}_n^{\mathrm{AO}}, u_n\right)$ to denote the optimal solutions in (36). The subscript n emphasizes the dependence of $\left(\hat{\boldsymbol{w}}_n^{\mathrm{AO}}, u_n\right)$ on the problem dimensions. Also note that (even though not explicit in the notation) $\left(\hat{\boldsymbol{w}}_n^{\mathrm{AO}}, u_n\right)$ are random variables depending on the realizations of $\bar{\boldsymbol{g}}, \bar{\boldsymbol{h}}$ and h.

Notice that the objective function above is convex in $\boldsymbol{w}$ and linear (thus, concave) in u. Also, $\mathcal{B}$ is compact. Thus, strong duality holds and we can flip the order of min-max (Fan 1953). Moreover, in order to make the objective easy to optimize with respect to $\boldsymbol{w}$, we use the following variational expression for the square-root term $\sqrt{\|\boldsymbol{w}\|_{\ell_2}^2 + \sigma^2}$:

$$\|\bar{\boldsymbol{g}}\|_{\ell_2}\sqrt{\|\boldsymbol{w}\|_{\ell_2}^2 + \sigma^2} = \|\bar{\boldsymbol{g}}\|_{\ell_2} \cdot \min_{\tau\in[\sigma,\sqrt{\sigma^2+B_+^2}]}\left\{\frac{\tau}{2} + \frac{\|\boldsymbol{w}\|_{\ell_2}^2 + \sigma^2}{2\tau}\right\} = \min_{\tau\in[\sigma,\sqrt{\sigma^2+B_+^2}]}\left\{\frac{\tau\|\bar{\boldsymbol{g}}\|_{\ell_2}^2}{2} + \frac{\sigma^2}{2\tau} + \frac{\|\boldsymbol{w}\|_{\ell_2}^2}{2\tau}\right\},$$

where B_+ is defined in Lemma 2. For convenience define the constraint set for the variable τ as $\mathcal{T}' := [\sigma, \sqrt{\sigma^2 + B_+^2}]$. For reasons to be made clear later in the proof (see proof of statement (iii)), we consider the (possibly larger) set:

$$\mathcal{T} := [\sigma, \max\{\sqrt{\sigma^2 + B_+^2}, 2\tau_*\}]$$

where τ_* is as in the statement of the lemma.

The above lead to the following equivalent formulation of (36),

$$\left(\hat{\boldsymbol{w}}_n^{\mathrm{AO}}, u_n, \tau_n\right) = \max_{u\geq 0}\min_{\boldsymbol{w}\in\mathcal{B},\tau\in\mathcal{T}}\ \frac{u\tau\|\bar{\boldsymbol{g}}\|_{\ell_2}^2}{2} + \frac{u\sigma^2}{2\tau} + \frac{u\sigma h}{\sqrt{n}} + \min_{\boldsymbol{w}\in\mathcal{B}}\left\{\frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w} + \boldsymbol{\beta}^\star\|_{\ell_2}^2 + \frac{u}{2\tau}\|\boldsymbol{w}\|_{\ell_2}^2 - u\sqrt{\kappa}\,\bar{\boldsymbol{h}}^T\boldsymbol{w}\right\}. \tag{37}$$

The minimization over $\boldsymbol{w}$ is easy as it involves a strongly convex quadratic function. First, note that the unconstrained optimal $\boldsymbol{w}' := \boldsymbol{w}'(\tau, u)$ (for fixed (τ, u)) is given by

$$\boldsymbol{w}' := \boldsymbol{w}'(\tau, u) = -\left(\boldsymbol{\Sigma}^{-1} + \frac{u}{\tau}\boldsymbol{I}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}^\star - u\sqrt{\kappa}\bar{\boldsymbol{h}}\right), \tag{38}$$

and (37) simplifies to

$$(u_n, \tau_n) = \max_{u\geq 0}\min_{\tau\in\mathcal{T}}\ \frac{u\tau\|\bar{\boldsymbol{g}}\|_{\ell_2}^2}{2} + \frac{u\sigma^2}{2\tau} + \frac{u\sigma h}{\sqrt{n}} - \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}^\star - u\sqrt{\kappa}\bar{\boldsymbol{h}}\right)^T\left(\boldsymbol{\Sigma}^{-1} + \frac{u}{\tau}\boldsymbol{I}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}^\star - u\sqrt{\kappa}\bar{\boldsymbol{h}}\right) =: \mathcal{R}(u, \tau). \tag{39}$$

It can be checked by direct differentiation and the second-derivative test that the objective function in (39) is strictly convex in τ and strictly concave in u over the domain $\{(u, \tau) \in \mathbb{R}_+ \times \mathbb{R}_+\}$ [1]. Thus, the saddle point (u_n, τ_n) is unique. Specifically, this implies that the optimal $\hat{\boldsymbol{w}}_n^{\mathrm{AO}}$ in (37) is given by (cf. (38))

$$\hat{\boldsymbol{w}}_n^{\mathrm{AO}} = \boldsymbol{w}'(\tau_n, u_n) = -\left(\boldsymbol{\Sigma}^{-1} + \frac{u_n}{\tau_n}\boldsymbol{I}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}^\star - u_n\sqrt{\kappa}\bar{\boldsymbol{h}}\right). \tag{40}$$

In Lemma 3(v) we will prove that wpa 1, in the limit of $p \to \infty$, $\|\hat{\boldsymbol{w}}_n^{\mathrm{AO}}\|_2 \leq C$ for sufficiently large absolute constant $C > 0$. Thus, by choosing the upper bound in the definition of $\mathcal{B}$ in Lemma 2 strictly larger than C, guarantees that the unconstrained $\hat{\boldsymbol{w}}_n^{\mathrm{AO}}$ in (40) is feasible in (37).

Asymptotic limit of the key quantities τ_n, u_n: In what follows, we characterize the high-dimensional limit of the optimal pair (u_n, τ_n) in the limit $n, p \to \infty,\ p/n \to \kappa$. We start by analyzing the (point-wise) convergence of $\mathcal{R}(u, \tau)$. For the first three summands in (39), we easily find that

$$\left\{\frac{u\tau\|\bar{\boldsymbol{g}}\|_{\ell_2}^2}{2} + \frac{u\sigma^2}{2\tau} + \frac{u\sigma h}{\sqrt{n}}\right\} \xrightarrow{P} \left\{\frac{u\tau}{2} + \frac{u\sigma^2}{2\tau}\right\}.$$

[1] To analyze the matrix-vector product term in (39) for (τ, u) one can use the fact that $\boldsymbol{\Sigma}$ is diagonal. This way, as a function of u and τ the analysis reduces to the properties of relatively simple functions. For instance, for τ, this function is in the form $f(\tau) = -(a + b/\tau)^{-1}$ for $a, b > 0$, which is strictly convex.

Next, we study the fourth summand. First, note that

$$
\begin{aligned}
(u\sqrt{\kappa}\bar{\boldsymbol{h}})^T \left(\boldsymbol{\Sigma}^{-1} + \frac{u}{\tau}\boldsymbol{I}\right)^{-1} (u\sqrt{\kappa}\bar{\boldsymbol{h}}) &= u^2\kappa \, \frac{1}{p}\boldsymbol{h}^T \left(\boldsymbol{\Sigma}^{-1} + \frac{u}{\tau}\boldsymbol{I}\right)^{-1} \boldsymbol{h} \\
&= u^2\kappa \, \frac{1}{p} \sum_{i=1}^{p} \frac{\boldsymbol{h}_i^2}{\boldsymbol{\Sigma}_{i,i}^{-1} + \frac{u}{\tau}} \\
&\xrightarrow{P} u^2\kappa \, \mathbb{E}\left[\frac{1}{\Lambda^{-1} + \frac{u}{\tau}}\right]. \qquad (41)
\end{aligned}
$$

In the last line, Λ is a random variable as in Definition 3. Also, we used Assumption 3 together with the facts that $\boldsymbol{h}$ is independent of $\boldsymbol{\Sigma}$ and that the function $(x_1, x_2) \mapsto x_1^2(x_2^{-1} + u/\tau)^{-1}$ is PL(3) assuming x_2 is bounded (see Lemma 5 for proof). Second, we find that

$$
\begin{aligned}
(\boldsymbol{\beta}^\star)^T \boldsymbol{\Sigma}^{-1/2} \left(\boldsymbol{\Sigma}^{-1} + \frac{u}{\tau}\boldsymbol{I}\right)^{-1} \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}^\star &= \frac{1}{p}(\sqrt{p}\boldsymbol{\beta}^\star)^T \left(\boldsymbol{I} + \frac{u}{\tau}\boldsymbol{\Sigma}\right)^{-1} (\sqrt{p}\boldsymbol{\beta}^\star) \\
&= \frac{1}{p} \sum_{i=1}^{p} \frac{\left(\sqrt{p}\boldsymbol{\beta}_i^\star / \sqrt{\boldsymbol{\Sigma}_{i,i}}\right)^2}{\boldsymbol{\Sigma}_{i,i}^{-1} + \frac{u}{\tau}} \\
&\xrightarrow{P} \mathbb{E}\left[\frac{B^2\Lambda^{-1}}{\Lambda^{-1} + \frac{u}{\tau}}\right]. \qquad (42)
\end{aligned}
$$

Here, Λ, B are random variables as in Definition 3 and we also used Assumption 3 together with the fact that the function $(x_1, x_2) \mapsto x_1^2 x_2^{-1}(x_2^{-1} + u/\tau)^{-1}$ is PL(3) assuming x_2 is bounded (see Lemma 5 for proof). Third, by independence of $(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})$ from $\boldsymbol{h}$

$$
(u\sqrt{\kappa}\bar{\boldsymbol{h}})^T \left(\boldsymbol{\Sigma}^{-1} + \frac{u}{\tau}\boldsymbol{I}\right)^{-1} \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}^\star = u\sqrt{\kappa} \cdot \frac{1}{p} \sum_{i=1}^{p} \frac{\boldsymbol{h}_i \boldsymbol{\Sigma}_{i,i}^{-1/2}(\sqrt{p}\boldsymbol{\beta}_i^\star)}{\boldsymbol{\Sigma}_{i,i}^{-1} + \frac{u}{\tau}\boldsymbol{I}} \xrightarrow{P} 0. \qquad (43)
$$

Putting these together, the objective $\mathcal{R}(u, \tau)$ in (39) converges point-wise in u, τ to

$$
\mathcal{R}(u, \tau) \xrightarrow{P} \mathcal{D}(u, \tau) := \frac{1}{2}\left(u\tau + \frac{u\sigma^2}{\tau} - u^2\kappa \, \mathbb{E}\left[\frac{1}{\Lambda^{-1} + \frac{u}{\tau}}\right] - \mathbb{E}\left[\frac{B^2\Lambda^{-1}}{\Lambda^{-1} + \frac{u}{\tau}}\right]\right). \qquad (44)
$$

Note that $\mathcal{R}(u, \tau)$ (and thus, $\mathcal{D}(u, \tau)$) is convex in τ and concave in u. Thus, the convergence in (44) is in fact uniform (e.g., (Andersen and Gill 1982)) and we can conclude that

$$
\phi(\boldsymbol{g}, \boldsymbol{h}) \xrightarrow{P} \max_{u \geq 0} \min_{\tau \in \mathcal{T}} \mathcal{D}(u, \tau). \qquad (45)
$$

and using strict concave/convexity of $\mathcal{D}(u, \tau)$, we also have the parameter convergence (Newey and McFadden 1994, Lem. 7.75)

$$
(u_n, \tau_n) \xrightarrow{P} (u_*, \tau_*) := \arg\max_{u \geq 0} \min_{\tau \in \mathcal{T}} \mathcal{D}(u, \tau). \qquad (46)
$$

In the proof of statement (iii) below, we show that the saddle point of (45) is (u_*, τ_*). In particular, τ_* is strictly in the interior of $\mathcal{T}$, which combined with convexity implies that

$$
\max_{u \geq 0} \min_{\tau \in \mathcal{T}} \mathcal{D}(u, \tau) = \max_{u \geq 0} \min_{\tau > 0} \mathcal{D}(u, \tau) =: \bar{\phi}.
$$

This, together with the first display above proves the second statement of the lemma.

Proof of (iii). Next, we compute the saddle point (u_*, τ_*) by studying the first-order optimality conditions of the strictly concave-convex $\mathcal{D}(u, \tau)$. Specifically, we consider the unconstrained minimization over τ and we will show that the minimum is achieved in the strict interior of $\mathcal{T}$. Direct differentiation of $\mathcal{D}(u, \tau)$ gives

$$
\tau + \frac{\sigma^2}{\tau} - 2u\kappa \, \mathbb{E}\left[\frac{1}{\Lambda^{-1} + \frac{u}{\tau}}\right] + \frac{u^2}{\tau}\kappa \, \mathbb{E}\left[\frac{1}{\left(\Lambda^{-1} + \frac{u}{\tau}\right)^2}\right] + \frac{1}{\tau} \, \mathbb{E}\left[\frac{B^2\Lambda^{-1}}{\left(\Lambda^{-1} + \frac{u}{\tau}\right)^2}\right] = 0, \qquad (47a)
$$

$$
u - \frac{u\sigma^2}{\tau^2} - \frac{u^3}{\tau^2}\kappa \, \mathbb{E}\left[\frac{1}{\left(\Lambda^{-1} + \frac{u}{\tau}\right)^2}\right] - \frac{u}{\tau^2} \, \mathbb{E}\left[\frac{B^2\Lambda^{-1}}{\left(\Lambda^{-1} + \frac{u}{\tau}\right)^2}\right] = 0, \qquad (47b)
$$

Multiplying (47b) with $\frac{\tau}{u}$ and adding to (47a) results in the following equation

$$\tau = u\kappa\,\mathbb{E}\left[\frac{1}{\Lambda^{-1}+\frac{u}{\tau}}\right] \;\Leftrightarrow\; \mathbb{E}\left[\frac{1}{(\frac{u}{\tau}\Lambda)^{-1}+1}\right] = \frac{1}{\kappa}\,. \tag{48}$$

Thus, we have found that the ratio $\frac{u_*}{\tau_*}$ is the unique solution to the equation in (48). Note that this coincides with the Equation (8) that defines the parameter ξ in Definition 3. The fact that (48) has a unique solution for all $\kappa > 1$ can be easily seen as $F(x) = \mathbb{E}\left[\frac{1}{(x\Lambda)^{-1}+1}\right], x \in \mathbb{R}_+$ has range $(0,1)$ and is strictly increasing (by differentiation).

Thus, we call $\xi = \frac{u_*}{\tau_*}$. Moreover, multiplying (47b) with u leads to the following equation for τ_*:

$$u_*^2 = \sigma^2\xi^2 + u_*^2\xi^2\kappa\,\mathbb{E}\left[\frac{1}{\left(\Lambda^{-1}+\xi\right)^2}\right] + \xi^2\,\mathbb{E}\left[\frac{B^2\Lambda^{-1}}{\left(\Lambda^{-1}+\xi\right)^2}\right] \;\Rightarrow\; \tau_*^2 = \frac{\sigma^2+\mathbb{E}\left[\frac{B^2\Lambda^{-1}}{(\Lambda^{-1}+\xi)^2}\right]}{1-\xi^2\kappa\,\mathbb{E}\left[\frac{1}{(\Lambda^{-1}+\xi)^2}\right]} = \frac{\sigma^2+\mathbb{E}\left[\frac{B^2\Lambda^{-1}}{(\Lambda^{-1}+\xi)^2}\right]}{1-\kappa\,\mathbb{E}\left[\frac{1}{((\xi\Lambda)^{-1}+1)^2}\right]}. \tag{49}$$

Again, note that this coincides with Equation (9) that determines the parameter γ in Definition 3, i.e., $\tau_*^2 = \gamma$. By definition of $\mathcal{T}$ and of τ_*, it is clear that $\tau_\star$ is in the strict interior of $\mathcal{T}$.

Proof of (iv). For convenience, define

$$F_n(\hat{\boldsymbol{\beta}}^{\mathrm{AO}}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) := \frac{1}{p}\sum_{i=1}^{p} f\left(\sqrt{p}\hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}}, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{ii}\right) \quad \text{and} \quad \alpha_* := \mathbb{E}_\mu\left[f(X_{\kappa,\sigma^2}(\Lambda,B,H),B,\Lambda)\right].$$

Recall from (40) the explicit expression for $\hat{\boldsymbol{w}}_n^{\mathrm{AO}}$, repeated here for convenience.

$$\hat{\boldsymbol{w}}_n^{\mathrm{AO}} = -\left(\boldsymbol{\Sigma}^{-1}+\frac{u_n}{\tau_n}\boldsymbol{I}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}^\star - u_n\sqrt{\kappa}\bar{\boldsymbol{h}}\right).$$

Also, recall that $\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}} = \boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{w}}_n^{\mathrm{AO}} + \boldsymbol{\beta}^\star$. Thus, (and using the fact that $\bar{\boldsymbol{h}}$ is distributed as $-\bar{\boldsymbol{h}}$),

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}} &= \boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{\Sigma}^{-1}+\frac{u_n}{\tau_n}\boldsymbol{I}\right)^{-1}u_n\sqrt{\kappa}\bar{\boldsymbol{h}} + \left(\boldsymbol{I}-\boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{\Sigma}^{-1}+\frac{u_n}{\tau_n}\boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}^{-1/2}\right)\boldsymbol{\beta}^\star \\
\Longrightarrow \hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}} &= \frac{\boldsymbol{\Sigma}_{i,i}^{-1/2}}{1+(\xi_n\boldsymbol{\Sigma}_{i,i})^{-1}}\sqrt{\kappa}\tau_n\bar{\boldsymbol{h}}_i + \left(1-\frac{1}{1+\xi_n\boldsymbol{\Sigma}_{i,i}}\right)\boldsymbol{\beta}_i^\star.
\end{aligned} \tag{50}$$

For $i \in [p]$, define

$$\boldsymbol{v}_{n,i} = \frac{\boldsymbol{\Sigma}_{i,i}^{-1/2}}{1+(\xi_*\boldsymbol{\Sigma}_{i,i})^{-1}}\sqrt{\kappa}\tau_*\bar{\boldsymbol{h}}_i + \left(1-\frac{1}{1+\xi_*\boldsymbol{\Sigma}_{i,i}}\right)\boldsymbol{\beta}_i^\star \tag{51}$$

In the above, for convenience, we have denoted $\xi_n := u_n/\tau_n$ and recall that $\xi_* := u_*/\tau_*$.

The proof proceeds in two steps. In the first step, we use the fact that $\xi_n \xrightarrow{P} \xi_*$ and $u_n \xrightarrow{P} u_\star$ (see (46)) to prove that for any $\varepsilon \in (0, \xi_*/2)$, there exists an absolute constant $C > 0$ such that wpa 1:

$$|F_n(\sqrt{p}\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}, \sqrt{p}\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - F_n(\sqrt{p}\boldsymbol{v}_n, \sqrt{p}\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| \le C\varepsilon. \tag{52}$$

In the second step, we use pseudo-Lipschitzness of f and Assumption 3 to prove that

$$|F_n(\boldsymbol{v}_n, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| \xrightarrow{P} \alpha_*. \tag{53}$$

The desired follows by combining (52) and (53). Thus, in what follows, we prove (52) and (53).

Proof of (52). Fix some $\varepsilon \in (0, \xi_*/2)$. From (46), we know that w.p.a. 1 $|\xi_n - \xi_*| \le \varepsilon$ and $|u_n - u_*| \le \varepsilon$. Thus, $\hat{\boldsymbol{w}}_n^{\mathrm{AO}}$ is close to $\boldsymbol{v}_n$. Specifically, in this event, for every $i \in [p]$, it holds that:

$$\begin{aligned}
|\hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}} - \boldsymbol{v}_{n,i}| &\le |\boldsymbol{\beta}_i^\star|\left|\frac{1}{1+\xi_n\boldsymbol{\Sigma}_{i,i}} - \frac{1}{1+\xi_*\boldsymbol{\Sigma}_{i,i}}\right| + \sqrt{\kappa}\frac{|\bar{\boldsymbol{h}}_i|}{\sqrt{\boldsymbol{\Sigma}_{i,i}}}\left|\frac{\tau_n}{1+(\xi_n\boldsymbol{\Sigma}_{i,i})^{-1}} - \frac{\tau_*}{1+(\xi_*\boldsymbol{\Sigma}_{i,i})^{-1}}\right| \\
&= |\boldsymbol{\beta}_i^\star|\left|\frac{1}{1+\xi_n\boldsymbol{\Sigma}_{i,i}} - \frac{1}{1+\xi_*\boldsymbol{\Sigma}_{i,i}}\right| + \sqrt{\kappa}\frac{|\bar{\boldsymbol{h}}_i|}{\sqrt{\boldsymbol{\Sigma}_{i,i}}}\left|\frac{u_n}{\xi_n+\boldsymbol{\Sigma}_{i,i}^{-1}} - \frac{u_*}{\xi_*+\boldsymbol{\Sigma}_{i,i}^{-1}}\right| \\
&\le |\boldsymbol{\beta}_i^\star|\frac{|\boldsymbol{\Sigma}_{i,i}||\xi_n-\xi_*|}{|1+\xi_n\boldsymbol{\Sigma}_{i,i}||1+\xi_*\boldsymbol{\Sigma}_{i,i}|} + \sqrt{\kappa}\frac{|\bar{\boldsymbol{h}}_i|}{\sqrt{\boldsymbol{\Sigma}_{i,i}}}\frac{u_*|\xi_n-\xi_*|}{(\xi_n+\boldsymbol{\Sigma}_{i,i}^{-1})(\xi_*+\boldsymbol{\Sigma}_{i,i}^{-1})} + \sqrt{\kappa}\frac{|\bar{\boldsymbol{h}}_i|}{\sqrt{\boldsymbol{\Sigma}_{i,i}}}\frac{|u_n-u_*|}{\xi_n+\boldsymbol{\Sigma}_{i,i}^{-1}} \\
&\le |\boldsymbol{\beta}_i^\star|\Sigma_{\max}\varepsilon + \sqrt{\kappa}|\bar{\boldsymbol{h}}_i|u_*\Sigma_{\max}^{3/2}\varepsilon + \sqrt{\kappa}|\bar{\boldsymbol{h}}_i|\Sigma_{\max}^{1/2}\varepsilon \\
&\le C\varepsilon\left(|\bar{\boldsymbol{h}}_i| + |\boldsymbol{\beta}_i^\star|\right).
\end{aligned} \tag{54}$$

where $C = C(\Sigma_{\max}, \kappa, u_*)$ is an absolute constant. In the second line above, we recalled that $u_n = \tau_n \xi_n$ and $u_* = \tau_* \xi_*$. In the third line, we used the triangle inequality. In the fourth line, we used that $\xi_* > 0, 0 < \boldsymbol{\Sigma}_{i,i} \leq \Sigma_{\max}$ and $\xi_n \geq \xi_* - \varepsilon \geq \xi_*/2 > 0$.

Now, we will use this and Lipschitzness of f to argue that there exists absolute constant $C > 0$ such that wpa 1,

$$|F_n(\sqrt{p}\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}, \sqrt{p}\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - F_n(\sqrt{p}\boldsymbol{v}_n, \sqrt{p}\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| \leq C\varepsilon.$$

Denote, $\boldsymbol{a}_i = (\sqrt{p}\hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}}, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i})$ and $\boldsymbol{b}_i = (\sqrt{p}\boldsymbol{v}_{n,i}, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i})$. Following the exact same argument as in (33) (just substitute $\boldsymbol{\beta} \leftrightarrow \boldsymbol{v}_n$ in the derivation), we have that for some absolute constant $C > 0$ wpa 1:

$$|F_n(\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}(\boldsymbol{g}, \boldsymbol{h}), \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - F_n(\boldsymbol{v}_n, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| \leq C\|\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}} - \boldsymbol{v}_n\|_2. \tag{55}$$

From this and (54), we find that

$$\begin{aligned}|F_n(\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}(\boldsymbol{g}, \boldsymbol{h}), \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - F_n(\boldsymbol{v}_n, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| &\leq C\varepsilon \left(\sum_{i=1}^p \left(|\bar{\boldsymbol{h}}_i| + |\boldsymbol{\beta}_i^\star|\right)^2\right)^{1/2} \\ &\leq C\varepsilon\sqrt{2}\sqrt{\|\boldsymbol{\beta}^\star\|_2^2 + \|\bar{\boldsymbol{h}}\|_2^2}. \end{aligned} \tag{56}$$

But, recall that $\|\boldsymbol{\beta}^\star\|_2^2 = \frac{1}{p}\sum_{i=1}^p (\sqrt{p}\boldsymbol{\beta}_i^\star)^2 < \infty$, as $p \to \infty$ by Assumption 3. Also, since $\bar{\boldsymbol{h}}_i \sim \mathcal{N}(0, 1/p)$, it holds that $\|\bar{\boldsymbol{h}}\|_2^2 \leq 2$, wpa 1 as $p \to \infty$. Therefore, from (56), wpa 1, there exists constant $C > 0$ such that

$$|F_n(\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}(\boldsymbol{g}, \boldsymbol{h}), \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) - F_n(\boldsymbol{v}_n, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| \leq C \cdot \varepsilon,$$

as desired.

Proof of (53). Next, we will use Assumption 3 to show that

$$|F_n(\boldsymbol{v}_n, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma})| \xrightarrow{P} \alpha_*. \tag{57}$$

Notice that $\boldsymbol{v}_n$ is a function of $\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}, \bar{\boldsymbol{h}}$. Concretely, define $\widetilde{g} : \mathbb{R}^3 \to \mathbb{R}$, such that

$$\widetilde{g}(x_1, x_2, x_3) := \frac{x_2^{-1/2}}{1 + (\xi_* x_2)^{-1}}\sqrt{\kappa}\tau_* x_3 + (1 - (1 + \xi_* x_2)^{-1})x_1,$$

and notice that

$$\sqrt{p}\boldsymbol{v}_{n,i} = \widetilde{g}\left(\sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}, \boldsymbol{h}_i\right) = \frac{\boldsymbol{\Sigma}_{i,i}^{-1/2}}{1 + (\xi_* \boldsymbol{\Sigma}_{i,i})^{-1}}\sqrt{\kappa}\tau_* \boldsymbol{h}_i + \left(1 - \frac{1}{1 + \xi_* \boldsymbol{\Sigma}_{i,i}}\right)\sqrt{p}\boldsymbol{\beta}_i^\star.$$

Thus,

$$F_n(\boldsymbol{v}_n, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) = \frac{1}{p}\sum_{i=1}^p f\left(g\left(\sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}, \boldsymbol{h}_i\right), \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}\right) =: \frac{1}{p}\sum_{i=1}^p h\left(\boldsymbol{h}_i, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}\right),$$

where we have defined $h : \mathbb{R}^3 \to \mathbb{R}$:

$$h(x_1, x_2, x_3) := f\left(\widetilde{g}(x_2, x_3, x_1), x_2, x_3\right). \tag{58}$$

We will prove that $h \in \mathrm{PL}(4)$. Indeed, if that were the case, then Assumption 3 gives

$$\frac{1}{p}\sum_{i=1}^p h\left(\boldsymbol{h}_i, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}\right) \xrightarrow{P} \mathbb{E}_{\mathcal{N}(0,1)\otimes\mu}\left[h(H, B, \Lambda)\right] = \mathbb{E}_{\mathcal{N}(0,1)\otimes\mu}\left[f\left(\widetilde{g}(B, \Lambda, H), B, \Lambda\right)\right] \tag{59}$$

$$= \mathbb{E}[f\left(X_{\kappa,\sigma^2}(\Lambda, B, H), B, \Lambda\right)] = \alpha_*, \tag{60}$$

where the penultimate equality follows by recognizing that (cf. Eqn. (93))

$$\widetilde{g}(B, \Lambda, H) = (1 - (1 + \xi_*\Lambda)^{-1})B + \sqrt{\kappa}\frac{\tau_*\Lambda^{-1/2}}{1 + (\xi_*\Lambda)^{-1}}H = X_{\kappa,\sigma^2}(\Lambda, B, H).$$

It remains to show that $h \in \mathrm{PL}(4)$. Lemma 6 in Section D shows that if $f \in \mathrm{PL(k)}$, then $h \in \mathrm{PL(k+1)}$ for all integers $k \geq 2$. Using this and the fact that $\mathcal{F} \subset \mathrm{PL}(3)$, for any $f \in \mathcal{F}$, we find that $h \in \mathrm{PL}(4)$. This completes the proof of (53).

Proof of (v): Let $\psi : \mathbb{R} \to \mathbb{R}$ be any bounded Lipschitz function. The function $f(a, b, c) = \psi(a)$ is trivially PL(2). Thus, by directly applying statement (iv) of the lemma, we find that

$$\frac{1}{p}\sum_{i=1}^{p} \psi(\sqrt{p}\hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}}(\boldsymbol{g}, \boldsymbol{h})) \xrightarrow{P} \mathbb{E}\left[\psi(X_{\kappa,\sigma^2})\right].$$

Since this holds for any bounded Lipschitz function, we have shown that the empirical distribution of $\hat{\boldsymbol{\beta}}_n^{\mathrm{AO}}$ converges weakly to the distribution of X_{κ,σ^2}. It remains to prove boundedness of the 2nd moment as advertised in (26). Recall from (50) that

$$\sqrt{p}\hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}} = \frac{\boldsymbol{\Sigma}_{i,i}^{-1/2}}{1 + (\xi_n \boldsymbol{\Sigma}_{i,i})^{-1}}\sqrt{\kappa}\tau_n \boldsymbol{h}_i + \left(1 - \frac{1}{1 + \xi_n \boldsymbol{\Sigma}_{i,i}}\right)(\sqrt{p}\boldsymbol{\beta}_i^\star).$$

Using this, boundedness of $\boldsymbol{\Sigma}_{i,i}$ from Assumption 3, and the fact that $\tau_n \xrightarrow{P} \tau_\star, \xi_n \xrightarrow{P} \xi_\star$, there exists constant $C = C(\Sigma_{\max}, \Sigma_{\min}, k, \tau_\star, \xi_\star)$ such that wpa 1,

$$\frac{1}{p}\sum_{i=1}^{p} |\sqrt{p}\hat{\boldsymbol{\beta}}_{n,i}^{\mathrm{AO}}|^2 \le C\left(\frac{1}{p}\sum_{i=1}^{p} |\boldsymbol{h}_i|^2 + \frac{1}{p}\sum_{i=1}^{p} |\sqrt{p}\boldsymbol{\beta}_i^\star|^2\right).$$

But the two summands in the expression above are finite in the limit of $p \to \infty$. Specifically, (i) from Assumption 3, $\frac{1}{p}\sum_{i=1}^{p} |\sqrt{p}\boldsymbol{\beta}_i^\star|^2 \xrightarrow{P} \mathbb{E}[B^2] < \infty$; (ii) $\frac{1}{p}\sum_{i=1}^{p} |\boldsymbol{h}_i|^2 \xrightarrow{P} \mathbb{E}[H^2] = 1$, using the facts that $\boldsymbol{h}_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ and $H \sim \mathcal{N}(0, 1)$. This proves (26), as desired.

C Asymptotic formulas on Magnitude- and Hessian- pruning

Here, we use Theorem 1 to characterize the risk of the magnitude- and Hessian- pruned solutions. This section supplements the discussion in Section 5.2. For completeness, first, we recall the magnitude-pruning results of Section 5.2 and restate Corollary 1 below. This corollary characterizes the performance of magnitude pruning. Following this, we shall further discuss Hessian pruning.

C.1 Magnitude-based pruning

We begin with the following necessary definitions. Define the hard-thresholding function with fixed threshold $t \in \mathbb{R}_+$ as follows:

$$\mathcal{T}_t(x) = \begin{cases} x & \text{if } |x| > t \\ 0 & \text{otherwise} \end{cases}. \tag{61}$$

Further, given model sparsity target $1 > \alpha > 0$, define the threshold $t^\star$ as follows:

$$t^\star := \inf\left\{t \in \mathbb{R} : \Pr(|X_{\kappa,\sigma^2}| \ge t) \ge \alpha\right\}. \tag{62}$$

Corollary 1 (Risk of Magnitude-pruning) *Let the same assumptions and notation as in the statement of Theorem 1 hold. Specifically, let $\hat{\boldsymbol{\beta}}$ be the min-norm solution in* (7) *and $\hat{\boldsymbol{\beta}}_s^M := \mathbb{T}_s(\hat{\boldsymbol{\beta}})$ the magnitude-pruned model at sparsity s. Recall the threshold $t^\star$ from* (15)*. The risk of $\hat{\boldsymbol{\beta}}_s^M$ satisfies the following in the limit of $n, p, s \to \infty$ at rates $\kappa := p/n > 1$ and $\alpha := s/p \in (0, 1)$ (cf. Assumption 1):*

$$\mathcal{L}(\hat{\boldsymbol{\beta}}_s^M) \xrightarrow{P} \sigma^2 + \mathbb{E}\left[\Lambda\left(B - \mathcal{T}_{t^\star}(X_{\kappa,\sigma^2})\right)\right],$$

where the expectation is over $(\Lambda, B, H) \sim \mu \otimes \mathcal{N}(0, 1)$.

The proof of the corollary above, is given in Section 5.2. Below, we extend the results to Hessian-based pruning.

C.2 Hessian-based pruning

Let $\hat{\boldsymbol{\beta}}$ be the min-norm solution in (7). Recall that the Hessian-pruned model (via Optimal Brain Damage) $\boldsymbol{\beta}_s^H$ at sparsity s is given by

$$\boldsymbol{\beta}_s^H = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbb{T}_s(\hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\beta}), \tag{63}$$

where $\hat{\boldsymbol{\Sigma}} = \mathrm{diag}(\boldsymbol{X}^T\boldsymbol{X})/n$ the diagonal of the empirical covariance matrix.

We will argue that the following formula characterizes the asymptotic risk of the Hessian pruning solution. Recall the notation in (61) and define

$$t^\diamond := \inf\left\{t \in \mathbb{R} : \Pr(|\Lambda^{1/2}X_{\kappa,\sigma^2}| \ge t) \ge \alpha\right\}. \tag{64}$$

The risk of the Hessian-pruned model satisfies the following in the limit of $n, p, s \to \infty$ at rates $\kappa := p/n > 1$ and $\alpha := s/p \in (0, 1)$ (cf. Assumption 1):

$$\mathcal{L}(\hat{\boldsymbol{\beta}}_s^H) \xrightarrow{P} \sigma^2 + \mathbb{E}\left[\left(\Lambda^{1/2}B - \mathcal{T}_{t^\diamond}(\Lambda^{1/2}X_{\kappa,\sigma^2})\right)^2\right], \tag{65}$$

where the expectation is over $(\Lambda, B, H) \sim \mu \otimes \mathcal{N}(0, 1)$. In our LGP experiments, we used this formula (65) to accurately predict the Hessian-based pruning performance.

Recall the definition of the hard-thresholding operator $\mathcal{T}_t(x)$. Similar to Section 5.2, we consider a threshold-based pruning vector

$$\hat{\boldsymbol{\beta}}_t^{\mathcal{T},H} := \hat{\boldsymbol{\Sigma}}^{-1/2}\mathcal{T}_{t/\sqrt{p}}(\hat{\boldsymbol{\Sigma}}^{1/2}\hat{\boldsymbol{\beta}}),$$

where $\mathcal{T}_t$ acts component-wise. Further define

$$\hat{\boldsymbol{\beta}}_t^{\mathcal{T},H^\star} := \boldsymbol{\Sigma}^{-1/2}\mathcal{T}_{t/\sqrt{p}}(\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\beta}}).$$

Note that $\hat{\boldsymbol{\beta}}_t^{\mathcal{T},H^\star}$ uses the true (diagonal) covariance matrix $\boldsymbol{\Sigma}$ instead of its sample estimate $\hat{\boldsymbol{\Sigma}}$. For later reference, note here that $\hat{\boldsymbol{\Sigma}}$ concentrates (entry-wise) to $\boldsymbol{\Sigma}$. Using boundedness of $\boldsymbol{\Sigma}$ and standard concentration of sub-exponential random variables.

First, we compute the limiting risk of $\hat{\boldsymbol{\beta}}_t^{\mathcal{T},H^\star}$. Then, we will use the fact that $\hat{\boldsymbol{\Sigma}}$ concentrates (entry-wise) to $\boldsymbol{\Sigma}$, to show that the risks of $\hat{\boldsymbol{\beta}}_t^{\mathcal{T},H^\star}$ and $\hat{\boldsymbol{\beta}}_t^{\mathcal{T},H}$ are arbitrarily close as $p \to \infty$.

Similar to (14),

$$\begin{aligned}
\mathcal{L}(\hat{\boldsymbol{\beta}}_t^{\mathcal{T},H^\star}) &= \sigma^2 + (\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}_t^{\mathcal{T},H^\star})^T\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}_t^{\mathcal{T},H^\star}) \\
&= \sigma^2 + \frac{1}{p}\sum_{i=1}^p \boldsymbol{\Sigma}_{i,i}\big(\sqrt{p}\boldsymbol{\beta}_i^\star - \boldsymbol{\Sigma}_{i,i}^{-1/2}\mathcal{T}_t(\boldsymbol{\Sigma}_{i,i}^{1/2}\sqrt{p}\hat{\boldsymbol{\beta}}_i)\big)^2 = \sigma^2 + \frac{1}{p}\sum_{i=1}^p \big(\boldsymbol{\Sigma}_{i,i}^{1/2}\sqrt{p}\boldsymbol{\beta}_i^\star - \mathcal{T}_t(\boldsymbol{\Sigma}_{i,i}^{1/2}\sqrt{p}\hat{\boldsymbol{\beta}}_i)\big)^2 \\
&= \sigma^2 + \frac{1}{p}\sum_{i=1}^p \big(\boldsymbol{\Sigma}_{i,i}^{1/2}\sqrt{p}\boldsymbol{\beta}_i^\star - \mathcal{T}_t(\widehat{\boldsymbol{w}}_i + \boldsymbol{\Sigma}_{i,i}^{1/2}\sqrt{p}\boldsymbol{\beta}_i^\star\big)^2 \\
&=: \sigma^2 + \frac{1}{p}\sum_{i=1}^p \big(\sqrt{p}\boldsymbol{\lambda}_i^\star - \mathcal{T}_t(\widehat{\boldsymbol{w}}_i + \boldsymbol{\lambda}_i^\star)\big)^2.
\end{aligned} \tag{66}$$

In the second line above, we used that $\sqrt{p}\mathcal{T}_{t/\sqrt{p}}(x) = \mathcal{T}_t(\sqrt{p}x)$. In the third line, we recalled the change of variable in (19), i.e., $\widehat{\boldsymbol{w}}$ is the solution to (20). Finally, in the last line we defined $\boldsymbol{\lambda}^\star := \sqrt{\boldsymbol{\Sigma}}\boldsymbol{\beta}^\star$ (note that this is related to the saliency score $\boldsymbol{\lambda}$ defined in (4)).

To evaluate the limit of the empirical average in (66), we proceed as follows. First, we claim that the empirical distribution of $\sqrt{p}\boldsymbol{\lambda}^\star$ converges weakly to the distribution of the random variable $B\sqrt{\Lambda}$, where $(\Lambda, B) \sim \mu$. Note that this convergence is already implied by the proof of Theorem 1 by setting the g function to be zero in (12). For an explicit proof, take any bounded Lipschitz test function $\psi : \mathbb{R} \to \mathbb{R}$ and call $\psi'(x, y) := \psi(\sqrt{x}y)$. Then,

$$\begin{aligned}
|\psi'(\boldsymbol{\Sigma}_{i,i}, \sqrt{p}\boldsymbol{\beta}_i^\star) - \psi'(\boldsymbol{\Sigma}'_{i,i}, \sqrt{p}\boldsymbol{\beta}_i^{\star\prime})| &= |\psi(\sqrt{\boldsymbol{\Sigma}_{i,i}}\sqrt{p}\boldsymbol{\beta}_i^\star) - \psi(\sqrt{\boldsymbol{\Sigma}'_{i,i}}\sqrt{p}\boldsymbol{\beta}_i^{\star\prime})| \le C|\sqrt{\boldsymbol{\Sigma}_{i,i}}\sqrt{p}\boldsymbol{\beta}_i^\star - \sqrt{\boldsymbol{\Sigma}'_{i,i}}\sqrt{p}\boldsymbol{\beta}_i^{\star\prime}| \\
&\le C|\sqrt{p}\boldsymbol{\beta}_i^\star - \sqrt{p}\boldsymbol{\beta}_i^{\star\prime}| + C'|\sqrt{p}\boldsymbol{\beta}_i^{\star\prime}||\sqrt{\boldsymbol{\Sigma}_{i,i}} - \sqrt{\boldsymbol{\Sigma}'_{i,i}}| \\
&\le C(|\sqrt{p}\boldsymbol{\beta}_i^\star - \sqrt{p}\boldsymbol{\beta}_i^{\star\prime}| + |\sqrt{p}\boldsymbol{\beta}_i^{\star\prime}||\boldsymbol{\Sigma}_{i,i} - \boldsymbol{\Sigma}'_{i,i}|) \\
&\le C(1 + \|[\sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}]\|_2 + \|[\sqrt{p}\boldsymbol{\beta}_i^{\star\prime}, \boldsymbol{\Sigma}'_{i,i}]\|_2)\sqrt{|\sqrt{p}\boldsymbol{\beta}_i^\star - \sqrt{p}\boldsymbol{\beta}_i^{\star\prime}|^2 + |\boldsymbol{\Sigma}_{i,i} - \boldsymbol{\Sigma}'_{i,i}|^2}.
\end{aligned}$$

Thus, ψ' is PL(2). Hence, from Assumption 3,

$$\frac{1}{p}\sum_{i=1}^p \psi(\sqrt{p}\boldsymbol{\lambda}_i^\star) = \frac{1}{p}\sum_{i=1}^p \psi'(\boldsymbol{\Sigma}_{i,i}, \sqrt{p}\boldsymbol{\beta}_i^\star) \xrightarrow{P} \mathbb{E}[\psi'(B, \sqrt{\Lambda})] = \mathbb{E}[\psi(B\sqrt{\Lambda})] \tag{67}$$

Besides, from Theorem 1 applied for $f_{\mathcal{L}}(x, y, z) = zy^2$ (i.e., set g the zero function in (12)), we have that

$$\frac{1}{p}\sum_{i=1}^p \sqrt{p}(\boldsymbol{\lambda}_i^\star)^2 \xrightarrow{P} \mathbb{E}[B^2\Lambda].$$

Therefore, convergence in (67) actually holds for any $\psi \in \mathrm{PL}(2)$. Next, observe that, $\boldsymbol{\omega}$ of (40) can be written in terms of $\boldsymbol{\lambda}^\star$ via

$$\boldsymbol{w}_n = \boldsymbol{w}'(\tau_n, u_n) = -\left(\boldsymbol{I} + \frac{u_n}{\tau_n}\boldsymbol{\Sigma}\right)^{-1}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^\star - u_n\sqrt{\kappa}\boldsymbol{\Sigma}\bar{\boldsymbol{h}}\right) \tag{68}$$

$$= -\left(\boldsymbol{I} + \frac{u_n}{\tau_n}\boldsymbol{\Sigma}\right)^{-1}\left(\boldsymbol{\lambda}^\star - u_n\sqrt{\kappa}\boldsymbol{\Sigma}\bar{\boldsymbol{h}}\right). \tag{69}$$

After this observation, the convergence proof can be finalized by using a modified version of Assumption 3 as follows.

Assumption 4 (Empirical distribution for saliency) *Set $\boldsymbol{\lambda}_i^\star = \sqrt{\boldsymbol{\Sigma}_{i,i}}\boldsymbol{\beta}_i^\star$. The joint empirical distribution of $\{(\boldsymbol{\Sigma}_{i,i}, \boldsymbol{\lambda}_i^\star)\}_{i\in[p]}$ converges in Wasserstein-k distance to a probability distribution $\mu = \mu(\Lambda, S)$ on $\mathbb{R}_{>0} \times \mathbb{R}$ for some $k \geq 4$. That is $\frac{1}{p}\sum_{i\in[p]}\delta_{(\boldsymbol{\Sigma}_{i,i}, \sqrt{p}\boldsymbol{\lambda}_i^\star)} \overset{W_k}{\Longrightarrow} \mu$.*

Under this assumption, it can be verified that, the exact same proof strategy we used for magnitude-based pruning would apply to Hessian-based pruning by replacing $\boldsymbol{\beta}^\star$ with $\boldsymbol{\lambda}^\star$. The reason is that the Hessian pruning risk is a $\mathrm{PL}(4)$ function of $\boldsymbol{h}, \boldsymbol{\lambda}^\star, \boldsymbol{\Sigma}$ (e.g. can be shown in a similar fashion to Lemma 6). Observe that Assumption 4 is a reasonable assumption given the naturalness of the saliency score. If we only wish to use the earlier Assumption 3 rather than Assumption 4, one can obtain the equivalent result by modifying 3 to enforce a slightly higher order bounded moment and convergence condition. Finally, one needs to address the perturbation due to the finite sample estimation of the covariance. Note that, even if the empirical covariance doesn't converge to the population, its diagonal weakly converges to the population (as we assumed the population is diagonal). The (asymptotically vanishing) deviation due to the finite sample affects can be addressed in an identical fashion to the deviation analysis of τ_n, u_n at (52) and (54). While these arguments are reasonably straightforward and our Hessian pruning formula accurately predicts the empirical performance, the fully formal proof of the Hessian-based pruning is rather lengthy to write and does not provide additional insights.

D Useful results about pseudo-Lipschitz functions and CGMT

For $k \geq 1$ we say a function $f : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order k and denote it by $f \in \mathrm{PL(k)}$ if there exists a constant $L > 0$ such that, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\left(1 + \|\boldsymbol{x}\|_2^{k-1} + \|\boldsymbol{y}\|_2^{k-1}\right)\|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{70}$$

In particular, when $f \in \mathrm{PL(k)}$, the following properties hold; see (Bayati and Montanari 2011):

1. There exists a constant L' such that for all $\boldsymbol{x} \in \mathbb{R}^n$: $|f(\boldsymbol{x})| \leq L'(1 + \|\boldsymbol{x}\|_2^k)$.
2. f is locally Lipschitz, that is for any $M > 0$, there exists a constant $L_{M,m} < \infty$ such that for all $x, y \in [-M, M]^m$, $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L_{M,m}\|\boldsymbol{x} - \boldsymbol{y}\|_2$. Further, $L_{M,m} \leq c(1 + (M\sqrt{m})^{k-1})$ for some costant c.

Using the above properties, we prove the following two technical lemmas used in the proof of Theorem 1

Lemma 4 *Let $g : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function. Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined as follows:*

$$f(\boldsymbol{x}) = x_3(x_2 - g(x_1))^2.$$

Then, $f \in \mathrm{PL}(3)$. If additionally, $f : \mathbb{R}^2 \times \mathcal{Z} \to \mathbb{R}$ for a bounded set $\mathcal{Z} \subset \mathbb{R}$, then $f \in \mathcal{F} \subset \mathrm{PL}(3)$. Specifically, setting $\mathcal{Z} = [\Sigma_{\min}, \Sigma_{\max}]$ (as per Assumption 3), we find that $\mathcal{F}_{\mathcal{L}} \subset \mathcal{F}$, where $\mathcal{F}_{\mathcal{L}}$ is defined in (12)*.*

Proof We first prove that $f \in \mathrm{PL}(3)$.

Let $h : \mathbb{R}^2 \to \mathbb{R}$ defined as $h(\boldsymbol{u}) = (\boldsymbol{u}_2 - g(\boldsymbol{u}_1))^2$. The function $(\boldsymbol{u}_1, \boldsymbol{u}_2) \mapsto \boldsymbol{u}_2 - g(\boldsymbol{u}_1)$ is clearly Lipschitz. Thus, $h \in \mathrm{PL}(2)$, i.e.,

$$|h(\boldsymbol{u}) - h(\boldsymbol{v})| \leq C(1 + \|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2)\|\boldsymbol{u} - \boldsymbol{v}\|_2 \quad \text{and} \quad |h(\boldsymbol{v})| \leq C'(1 + \|\boldsymbol{v}\|_2^2). \tag{71}$$

Therefore, letting $\boldsymbol{x} = (\boldsymbol{u}, x_3) \in \mathbb{R}^3$ and $\boldsymbol{y} = (\boldsymbol{v}, y_3) \in \mathbb{R}^3$, we have that

$$\begin{aligned}
|f(\boldsymbol{x}) - f(\boldsymbol{y})| &= |x_3 h(\boldsymbol{u}) - y_3 h(\boldsymbol{v})| \leq |x_3||h(\boldsymbol{u}) - h(\boldsymbol{v})| + |h(\boldsymbol{v})||x_3 - y_3| \\
&\leq C|x_3|(1 + \|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2)\|\boldsymbol{u} - \boldsymbol{v}\|_2 + C'(1 + \|\boldsymbol{v}\|_2^2)|x_3 - y_3| \\
&\leq C(|x_3|^2 + (1 + \|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2)^2)\|\boldsymbol{u} - \boldsymbol{v}\|_2 + C'(1 + \|\boldsymbol{v}\|_2^2)|x_3 - y_3| \\
&\leq C(1 + \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2)\|\boldsymbol{u} - \boldsymbol{v}\|_2 + C'(1 + \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2)|x_3 - y_3| \\
&\leq C(1 + \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2)\|\boldsymbol{x} - \boldsymbol{y}\|_2.
\end{aligned} \tag{72}$$

In the second line, we used (71). In the third line, we used $2xy \leq x^2 + y^2$. In the fourth line, we used Cauchy-Schwarz inequality. $C, C' > 0$ are absolute constants that may change from line to line.

Now, we shall prove that $f \in \mathcal{F}$. To accomplish this, we simply need to show that for all $z \in \mathcal{Z}$, $f(\cdot,\cdot,z)$ is PL(2) (for some uniform PL constant). This can be shown as follows. Let $C = \sup_{z\in\mathcal{Z}} |z|$. Then

$$|f(x,y,z) - f(x',y',z)| = |z(y-g(x))^2 - z(y'-g(x'))^2| \tag{73}$$
$$\leq C|(y-g(x))^2 - (y'-g(x'))^2| \tag{74}$$
$$\leq C(1 + \|\boldsymbol{u}\|_{\ell_2} + \|\boldsymbol{v}\|_{\ell_2})\|\boldsymbol{u}-\boldsymbol{v}\|_{\ell_2}, \tag{75}$$

where $\boldsymbol{u} = [x,y]$ and $\boldsymbol{v} = [x',y']$. In the second line above, we used boundedness of $\mathcal{Z}$. In the third line, we used the fact that the function $\psi(x,y) = (y-g(x))^2$ is PL(2) as it is a quadratic of a Lipschitz function.

This completes the proof of the lemma. ∎

Lemma 5 (PL with Bounded Variables) *Let $f : \mathbb{R}^{d_1} \to \mathbb{R}$ be a $PL(k)$ function, $\mathcal{M} \subset \mathbb{R}^{d_2}$ be a compact set and g be a continuously differentiable function over $\mathcal{M}$. Then $h(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{x})g(\boldsymbol{y})$ is $PL(k+1)$ over $\mathbb{R}^{d_1}\times\mathcal{M}$.*

Proof First observe that g has continuous derivatives and is continuous over a compact set. Thus g and its gradient are bounded and g is Lipschitz over $\mathcal{M}$. Let $B = \sup_{\boldsymbol{x}\in\mathcal{M}} \max|g(x)|, \|\nabla g(\boldsymbol{x})\|_{\ell_2}$. To proceed, given pairs $(\boldsymbol{x},\boldsymbol{y})$ and $(\boldsymbol{x}',\boldsymbol{y}')$ over $\mathbb{R}\times\mathcal{M}$, we have that

$$|h(\boldsymbol{x},\boldsymbol{y}) - h(\boldsymbol{x}',y')| \leq |h(\boldsymbol{x},\boldsymbol{y}) - h(\boldsymbol{x}',\boldsymbol{y})| + |h(\boldsymbol{x}',\boldsymbol{y}) - h(\boldsymbol{x}',\boldsymbol{y}')| \tag{76}$$
$$\leq |f(\boldsymbol{x}) - f(\boldsymbol{x}')||g(\boldsymbol{y})| + |f(\boldsymbol{x}')||g(\boldsymbol{y}) - g(\boldsymbol{y}')| \tag{77}$$
$$\leq B|f(\boldsymbol{x}) - f(\boldsymbol{x}')| + B\|\boldsymbol{y}-\boldsymbol{y}'\|_{\ell_2}|f(\boldsymbol{x}')| \tag{78}$$
$$\leq B(1 + \|\boldsymbol{x}'\|_{\ell_2}^{k-1} + \|\boldsymbol{x}\|_{\ell_2}^{k-1})\|\boldsymbol{x}-\boldsymbol{x}'\|_{\ell_2} + B\|\boldsymbol{y}-\boldsymbol{y}'\|_{\ell_2}(1+\|\boldsymbol{x}'\|_{\ell_2}^{k}) \tag{79}$$
$$\leq C(1 + \|\boldsymbol{z}\|_{\ell_2}^{k} + \|\boldsymbol{z}'\|_{\ell_2}^{k})\|\boldsymbol{z}-\boldsymbol{z}'\|_{\ell_2}, \tag{80}$$

where $\boldsymbol{z} = [\boldsymbol{x}\ \boldsymbol{y}]^T$ and C an absolute constant. This shows the desired PL$(k+1)$ guarantee. ∎

The following lemma is in similar spirit to Lemma 5 and essentially follows from similar lines of arguments (i.e. using Lipschitzness induced by boundedness).

Lemma 6 *Let functions $f, g : \mathbb{R}^3 \to \mathbb{R}$ such that $f \in$ PL(k) and*

$$g(x,y,z) := \frac{y^{-1/2}}{1+(\xi_* y)^{-1}}\sqrt{\kappa}\tau_* z + (1-(1+\xi_* y)^{-1})x.$$

Here, ξ_*, τ_*, κ are positive constants. Further define

$$h(x,y,z) := f\left(g(y,z,x), y, z\right), \tag{81}$$

and assume that y take values on a fixed bounded compact set $\mathcal{M}\subset\mathbb{R}^+$. Then, it holds that $h \in$ PL(k + 1).

Proof Since f is PL(k), for some $L > 0$, (70) holds. Fix $x, x' \in \mathbb{R}$, $\boldsymbol{a} = [y,z] \in \mathbb{R}^2$ and $\boldsymbol{a}' = [y',z'] \in \mathbb{R}^2$. Let $\boldsymbol{b} = [\boldsymbol{g},\boldsymbol{a}] = [\boldsymbol{g},y,z] \in \mathbb{R}^3$ where $\boldsymbol{g} = g(y,z,x) \in \mathbb{R}$ and define accordingly $\boldsymbol{b}' = [\boldsymbol{g}',\boldsymbol{a}']$ and $\boldsymbol{g}' = g(y',z',x')$. We have that

$$|h([x,\boldsymbol{a}]) - h([x',\boldsymbol{a}'])| = |f(\boldsymbol{b}) - f(\boldsymbol{b}')|$$
$$\leq L\left(1 + \|\boldsymbol{b}\|_2^{k-1} + \|\boldsymbol{b}'\|_2^{k-1}\right)\|\boldsymbol{b}-\boldsymbol{b}'\|_2$$
$$\leq C\left(1 + \|\boldsymbol{a}\|_2^{k-1} + \|\boldsymbol{a}'\|_2^{k-1} + |\boldsymbol{g}|^{k-1} + |\boldsymbol{g}'|^{k-1}\right)(\|\boldsymbol{a}-\boldsymbol{a}'\|_2 + |\boldsymbol{g}-\boldsymbol{g}'|), \tag{82}$$

for some constant $C > 0$. In the last inequality we have repeatedly used the inequality $\left(\sum_{i=1}^m \|\boldsymbol{v}_i\|_2^2\right)^{\frac{d}{2}} \leq C(m)\cdot\sum_{i=1}^m \|\boldsymbol{v}_i\|_2^d$. Next, we need to bound the $\boldsymbol{g}$ term in terms of $(x,\boldsymbol{a})$. This is accomplished as follows

$$|\boldsymbol{g}|^{k-1} = \left|\frac{y^{-1/2}}{1+(\xi y)^{-1}}\sqrt{\kappa}\tau_* z + (1-(1+\xi_* y)^{-1})x\right|^{k-1}$$
$$\leq C(|z|+|x|)^{k-1}$$
$$\leq C\left(|x|^{k-1} + |z|^{k-1}\right) \leq C(|x|^{k-1} + \|\boldsymbol{a}\|_2^{k-1}). \tag{83}$$

Here, the value of the constant $C > 0$ may change from line to line. Secondly and similarly, we have the following perturbation bound on the $\boldsymbol{g}-\boldsymbol{g}'$. Recall that $y, y' \subset \mathcal{M}$ are bounded. Additionally, since $\mathcal{M} \subset \mathbb{R}^+$ and is compact, $\mathcal{M}$ is strictly bounded away from 0. Let

$$g_1(y) = \sqrt{\kappa}\tau_* \frac{y^{-1/2}}{1+(\xi_* y)^{-1}} \quad \text{and} \quad g_2(y) = 1-(1+\xi_* y)^{-1}.$$

It can be seen that g_1, g_2 are continuously differentiable functions over $\mathcal{M}$. Thus g_1, g_2 are bounded and have bounded derivatives over $\mathcal{M}$. We will prove the following sequence of inequalities

$$\begin{aligned}
|\boldsymbol{g} - \boldsymbol{g}'| &= |g(y,z,x) - g(y',z',x')| \\
&\le |g_1(y)x - g_1(y')x'| + |g_2(y)z - g_2(y')z'| \\
&\le |g_1(y)x - g_1(y)x'| + |g_1(y)x' - g_1(y')x'| \\
&\quad + |g_2(y)z - g_2(y)z'| + |g_2(y)z' - g_2(y')z'| \\
&\le C_1|x - x'| + C_2|x'||y - y'| + C_3|z - z'| + C_4|z'||y - y'| \\
&\le C(1 + |x'| + |z'|)(|x - x'| + |z - z'| + |y - y'|) \qquad (84) \\
&\le C\sqrt{3}(1 + |x'| + |z'|)\|[\boldsymbol{a}, x] - [\boldsymbol{a}', x']\|_2. \qquad (85)
\end{aligned}$$

In the fourth inequality above, we used the fact that $|g_i(y)|, |g_i'(y)|$ are bounded. In the last line, we used Cauchy-Scwhartz.

Substituting (83) and (85) in (82) gives:

$$\begin{aligned}
|h(x,y,z) - h(x',y',z')| &\le C\left(1 + \|\boldsymbol{a}\|_2^{k-1} + \|\boldsymbol{a}'\|_2^{k-1} + |x|^{k-1} + |x'|^{k-1}\right)(1 + |x'| + |z'|)\|[\boldsymbol{a}, x] - [\boldsymbol{a}', x']\|_2 \\
&\le C\left(1 + \|[\boldsymbol{a}, x]\|_2^{k} + \|[\boldsymbol{a}', x']\|_2^{k}\right)\|[\boldsymbol{a}, x] - [\boldsymbol{a}', x']\|_2. \qquad (86)
\end{aligned}$$

Thus, $h \in \mathrm{PL}(\mathrm{k}+1)$, as desired. ■

The following theorem replaces the compactness constraint with closedness in the CGMT and is borrowed from (Li et al. 2020). For related statements see also (Deng, Kammoun, and Thrampoulidis 2019, App. A).

Theorem 2 (CGMT with Closedness Constrains) *Let ψ be a convex function obeying $\lim_{\|\boldsymbol{w}\|_{\ell_2}\to\infty} \psi(\boldsymbol{w}) = \infty$. Given a closed set $\mathcal{S}$, define*

$$\Phi_\lambda(\boldsymbol{X}) = \min_{\boldsymbol{w}\in\mathcal{S}} \lambda\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2} + \psi(\boldsymbol{w}) \qquad (87)$$

$$\phi_\lambda(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}} \lambda(\|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2} - \boldsymbol{h}^T\boldsymbol{w})_+ + \psi(\boldsymbol{w}), \qquad (88)$$

and

$$\Phi_\infty(\boldsymbol{X}) = \min_{\boldsymbol{w}\in\mathcal{S}, \boldsymbol{X}\boldsymbol{w}=0} \psi(\boldsymbol{w}) \qquad (89)$$

$$\phi_\infty(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}, \|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2} \le \boldsymbol{h}^T\boldsymbol{w}} \psi(\boldsymbol{w}). \qquad (90)$$

For all $\lambda \in [0,\infty) \cup \{\infty\}$, we have that

- $\mathbb{P}(\Phi_\lambda(\boldsymbol{X}) < t) \le 2\mathbb{P}(\phi_\lambda(\boldsymbol{X}) \le t)$.
- *If $\mathcal{S}$ is additionally convex, we additionally have that $\mathbb{P}(\Phi_\lambda(\boldsymbol{X}) > t) \le 2\mathbb{P}(\phi_\lambda(\boldsymbol{X}) \ge t)$. Combining with the first statement, this implies that for any $\mu, t > 0$*

$$\mathbb{P}(|\Phi_\lambda(\boldsymbol{X}) - \mu| > t) \le 2\mathbb{P}(|\phi_\lambda(\boldsymbol{X}) - \mu| \ge t)$$

E Underparameterized analysis

This section provides our results for the asymptotic DC in the underparameterized regime. This results establish direct counterparts of the overparameterized results Definition 3 and Theorem 1. However, underparameterized DC is substantially less involved compared to overparameterized. A key reason is that underparameterized least-squares returns an unbiased estimate of the ground-truth parameter. Similar to Section B, for simplicity, we assume diagonal covariance however results can be translated to arbitrary covariance via eigen-rotation trick (e.g. recall Def. 4). Throughout, we solve the following problem

$$\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\dagger\boldsymbol{y} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2 \qquad (91)$$

where $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \sigma\boldsymbol{z}$ and $\boldsymbol{X} = \bar{\boldsymbol{X}}\sqrt{\boldsymbol{\Sigma}}$. Now, set $\boldsymbol{\omega} = \sqrt{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)$ as previously. We can rewrite

$$\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\dagger\boldsymbol{y} = \boldsymbol{\beta}^\star + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{w}^\star \quad \text{where} \quad \boldsymbol{w}^\star = \arg\min_{\boldsymbol{\omega}} \|\sigma\boldsymbol{z} - \bar{\boldsymbol{X}}\boldsymbol{\omega}\|_{\ell_2}^2. \qquad (92)$$

We will prove the following DC for the underparameterized problem with $n < p$ and $p/n = \kappa < 1$.

Definition 6 (Asymptotic DC – Underparameterized regime) *Let random variables $(B, \Lambda) \sim \mu$ (where μ is defined in Assumption 3) and fix $\kappa < 1$. Let $H \sim \mathcal{N}(0,1)$ and define the random variable*

$$X_{\kappa,\sigma^2}(B, H) := B + \sigma\frac{\Lambda^{-1/2}H}{\sqrt{\kappa^{-1} - 1}}, \qquad (93)$$

and let Π_{κ,σ^2} be its distribution.

We are now ready to state our main theoretical result.

Theorem 3 (Asymptotic DC – Underparameterized LGP) *Fix $\kappa < 1$. Let Assumptions 2 and 3 hold. Consider $\hat{\boldsymbol{\beta}}$ as in* (91) *and $\hat{\Pi}_n(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) := \frac{1}{p}\sum_{i=1}^p \delta_{\sqrt{p}\hat{\boldsymbol{\beta}}_i, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}}$, the joint empirical distribution of $(\sqrt{p}\hat{\boldsymbol{\beta}}, \sqrt{p}\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})$. Recall the definition of the measure Π_{κ,σ^2} in Def. 6. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a function in $\mathcal{F}$ where $\mathcal{F}$ is defined in* (11)*. We have that*

$$\frac{1}{p}\sum_{i=1}^p f(\sqrt{p}\hat{\boldsymbol{\beta}}_i, \sqrt{p}\boldsymbol{\beta}_i^\star, \boldsymbol{\Sigma}_{i,i}) \xrightarrow{P} \mathbb{E}_{(\Lambda,B,H)\sim\mu\otimes\mathcal{N}(0,1)}\left[f(X_{\kappa,\sigma^2}, B, \Lambda)\right]. \tag{94}$$

Specifically, the asymptotic test risk of $\hat{\boldsymbol{\beta}}$ is given by $\frac{\sigma^2}{1-\kappa}$.

Proof To avoid repetition, we will not provide the full proof as the technical details of the proofs for over/under-parameterized overlap to a significant extent. Instead, we will provide the part of the proof that deviates from the overparameterized.

Since $\kappa < 1$, the problem has a unique solution. Set $\boldsymbol{\omega} = \sqrt{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)$. Define $\boldsymbol{X}_{\boldsymbol{z}} = [\bar{\boldsymbol{X}}\ \boldsymbol{z}]$ and $\boldsymbol{\omega}_\sigma = [\boldsymbol{\omega}\ \sigma]$. This leads to the optimization problem

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \|\bar{\boldsymbol{X}}\boldsymbol{\omega} + \sigma\boldsymbol{z}\|_{\ell_2} = \arg\min_{\boldsymbol{\omega}} \|\boldsymbol{X}_{\boldsymbol{z}}\boldsymbol{\omega}_\sigma\|_{\ell_2}.$$

Fix $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_p), \boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{I}_n), g \sim \mathcal{N}(0,1)$. Applying CGMT leads to the following Auxiliary Optimization

$$\phi(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{\omega}} \max_{\|\boldsymbol{a}\|_{\ell_2}\le 1} \boldsymbol{h}^T\boldsymbol{a}\|\boldsymbol{\omega}_\sigma\|_{\ell_2} - \|\boldsymbol{a}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{\omega} + g\sigma \tag{95}$$

$$= \min_{\boldsymbol{\omega}} \|\boldsymbol{h}\|_{\ell_2}\|\boldsymbol{\omega}_\sigma\|_{\ell_2} - \boldsymbol{g}^T\boldsymbol{\omega} + g\sigma. \tag{96}$$

Solving for optimal $\boldsymbol{\omega}$ leads to the solution

$$\boldsymbol{\omega}^{\text{AO}} = \arg\min_{\boldsymbol{\omega}} \|\boldsymbol{h}\|_{\ell_2}\|\boldsymbol{\omega}_\sigma\|_{\ell_2} - \boldsymbol{g}^T\boldsymbol{\omega} \implies \|\boldsymbol{h}\|_{\ell_2}\frac{\boldsymbol{\omega}^{\text{AO}}}{\sqrt{\|\boldsymbol{\omega}^{\text{AO}}\|_{\ell_2}^2 + \sigma^2}} - \boldsymbol{g} \implies \boldsymbol{\omega}^{\text{AO}} = \frac{\sigma\boldsymbol{g}}{\sqrt{\|\boldsymbol{h}\|_{\ell_2}^2 - \|\boldsymbol{g}\|_{\ell_2}^2}}.$$

Observing $\|\boldsymbol{\omega}^{\text{AO}}\|_{\ell_2}^2 + \sigma^2 = \frac{\sigma^2\|\boldsymbol{h}\|_{\ell_2}^2}{\|\boldsymbol{h}\|_{\ell_2}^2 - \|\boldsymbol{g}\|_{\ell_2}^2}$ and plugging $\boldsymbol{\omega}^{\text{AO}}$ in, we find

$$\sigma^{-1}\phi(\boldsymbol{g}, \boldsymbol{h}) = \frac{\|\boldsymbol{h}\|_{\ell_2}^2 - \|\boldsymbol{g}\|_{\ell_2}^2}{\sqrt{\|\boldsymbol{h}\|_{\ell_2}^2 - \|\boldsymbol{g}\|_{\ell_2}^2}} + g = \sqrt{\|\boldsymbol{h}\|_{\ell_2}^2 - \|\boldsymbol{g}\|_{\ell_2}^2} + g.$$

Thus, in the asymptotic regime $\phi(\boldsymbol{g}, \boldsymbol{h})$ converges to the objective

$$\phi(\boldsymbol{g}, \boldsymbol{h}) \xrightarrow{P} \bar{\phi} = \sigma\sqrt{n-p}.$$

The remaining arguments are same as in Lemma 3. First, the problem is strongly convex with $\sigma_{\min}^2(\boldsymbol{X})$, which satisfies $\sigma_{\min}^2(\boldsymbol{X})/p \gtrsim 1$ wpa. 1. Thus, the solution $\boldsymbol{w}^\star$ of the primary problem (92) will not deviate from $\boldsymbol{\omega}^{\text{AO}}$. Secondly, the empirical distribution of

$$\sqrt{p}\boldsymbol{\omega}^\star = \frac{\sigma\sqrt{p}\boldsymbol{g}}{\sqrt{\|\boldsymbol{h}\|_{\ell_2}^2 - \|\boldsymbol{g}\|_{\ell_2}^2}} \xrightarrow{P} \frac{\sigma\sqrt{p}\boldsymbol{g}}{\sqrt{n-p}} = \frac{\sigma\boldsymbol{g}}{\sqrt{n/p-1}} = \frac{\sigma\boldsymbol{g}}{\sqrt{\kappa^{-1}-1}},$$

converges to $\sigma H/\sqrt{\kappa^{-1}-1}$. By Assumption 3, the empirical distribution of $\sqrt{p}\hat{\boldsymbol{\beta}} = \sqrt{p}(\boldsymbol{\beta}^\star + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\omega}^\star)$ converges to $B + \sigma\Lambda^{-1/2}H/\sqrt{\kappa^{-1}-1} \sim \Pi_{\kappa,\sigma^2}$. Finally, again by Assumption 3, for any $f \in \mathcal{F}$, we obtain the advertised result (94). The asymptotic test risk is given by

$$\mathcal{L}(\hat{\boldsymbol{\beta}}) = \mathbb{E}[\|\boldsymbol{g}^T\sqrt{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) + \sigma z\|_{\ell_2}^2] = \sigma^2 + \sum_{i=1}^p (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}^\star)^2\boldsymbol{\Sigma}_{i,i} \xrightarrow{P} \sigma^2 + \frac{\sigma^2}{\kappa^{-1}-1} = \frac{\sigma^2}{1-\kappa}.$$

■

In the main body of the paper, we claim that the optimal s features to use in the underparameterized regime is given by the features with the maximum saliency score. This is proven below.

Lemma 7 (Optimal s features to use) *Fix a sequence of sets $\Delta_p \subset [p]$ of size s such that $\sum_{i\in\Delta_p}\boldsymbol{\beta}_i^{\star 2}\boldsymbol{\Sigma}_{i,i} \xrightarrow{P} B(\Delta)$. Set $\kappa = s/n$. Under same assumptions as in Thm 3, the asymptotic test risk of $\hat{\boldsymbol{\beta}}(\Delta)$ is given by*

$$\mathcal{L}(\hat{\boldsymbol{\beta}}(\Delta)) \xrightarrow{P} \frac{B - B(\Delta) + \sigma^2}{1-\kappa}.$$

Thus, the optimal feature set Δ chooses the indices with maximum Saliency Score (4) *which maximizes $B(\Delta)$.*

Proof The key idea is the fact that we can treat the missing features as uncorrelated noise. First, due to diagonal covariance, observe that, over the feature set Δ, the optimal population model (i.e. infinite sample) is $\boldsymbol{\beta}^\star_\Delta$. Thus, the s feature problem minimized by $\hat{\boldsymbol{\beta}}(\Delta)$ can be written as the dataset model

$$y = \boldsymbol{x}_\Delta^T \boldsymbol{\beta}^\star_\Delta + \sigma^2_\Delta,$$

where the noise level is given by

$$\mathbb{E}[(y - \boldsymbol{x}_\Delta \boldsymbol{\beta}^\star_\Delta)^2] = \sigma^2 + \mathbb{E}[(\boldsymbol{x}_{\bar{\Delta}} \boldsymbol{\beta}^\star_{\bar{\Delta}})^2] = \sigma^2 + \sum_{i \notin \Delta} \boldsymbol{\Sigma}_{i,i} \boldsymbol{\beta}^{\star 2}_i.$$

The latter quantity converges to $B - B(\Delta)$ wpa. 1. Thus, applying Theorem 3, $\hat{\boldsymbol{\beta}}(\Delta)$ achieves the advertised asymptotic risk. ■

F Proof of Lemma 1

Lemma 8 (Lemma 1 restated) *Suppose $\mathcal{S}$ is drawn from an LGP$(\sigma, \boldsymbol{\Sigma}, \boldsymbol{\beta}_\star)$ as in Def. 1 where rank$(\boldsymbol{\Sigma}) = 1$ with $\boldsymbol{\Sigma} = \boldsymbol{\lambda}\boldsymbol{\lambda}^T$ for $\boldsymbol{\lambda} \in \mathbb{R}^p$. Define $\zeta = \mathbb{T}_s(\boldsymbol{\lambda})^2 / \|\boldsymbol{\lambda}\|^2_{\ell_2}$. For magnitude and Hessian pruning ($P \in \{M, H\}$) and the associated retraining, we have the following excess risks with respect to $\boldsymbol{\beta}^\star$*

$$\mathbb{E}_{\mathcal{S}}[\mathcal{L}(\hat{\boldsymbol{\beta}}^P_s)] - \mathcal{L}(\boldsymbol{\beta}^\star) = \frac{\zeta^2 \sigma^2}{n-2} + \underbrace{(1-\zeta)^2 (\boldsymbol{\lambda}^T \boldsymbol{\beta}^\star)^2}_{\textit{Error due to bias}} \tag{97}$$

$$\mathbb{E}_{\mathcal{S}}[\mathcal{L}(\hat{\boldsymbol{\beta}}^{RT}_s)] - \mathcal{L}(\boldsymbol{\beta}^\star) = \sigma^2/(n-2). \tag{98}$$

Proof **Retraining analysis:** We claim that for any feature set Δ with $\boldsymbol{\lambda}_\Delta \neq 0$, the test risk of $\hat{\boldsymbol{\beta}}(\Delta)$ is exactly identical. Secondly, pruning is guaranteed to pick a nonzero support satisfying $\boldsymbol{\lambda}_\Delta \neq 0$ [2]. Thus, as described next, retraining always achieves a fixed risk. Set $c^\star = \boldsymbol{\lambda}^T \boldsymbol{\beta}^\star$. By definition, each input example $\boldsymbol{x}_i$ has the form $\boldsymbol{x}_i = g_i \boldsymbol{\lambda}$ and $y_i = g_i c^\star + \sigma z_i$. Set $\boldsymbol{g} = [g_1 \ \dots \ g_n]^T$ and $\bar{\boldsymbol{g}} = \boldsymbol{g}/\|\boldsymbol{g}\|_{\ell_2}$. Thus, we have $\boldsymbol{X} = \boldsymbol{g}\boldsymbol{\lambda}^T$ and $\boldsymbol{y} = \boldsymbol{g}\boldsymbol{\lambda}^T \boldsymbol{\beta}^\star + \sigma \boldsymbol{z}$. Decompose $\boldsymbol{z} = \bar{\boldsymbol{z}} + \bar{\boldsymbol{g}}^T \boldsymbol{z} \bar{\boldsymbol{g}}$ where $\bar{\boldsymbol{z}}$ is orthogonal to $\boldsymbol{g}$. When solving the regression of Δ, we have that

$$\boldsymbol{X}_\Delta = \boldsymbol{g}\boldsymbol{\lambda}_\Delta^T, \ \boldsymbol{y} = c^\star \boldsymbol{g} + \sigma(\bar{\boldsymbol{z}} + \bar{\boldsymbol{g}}^T \boldsymbol{z} \bar{\boldsymbol{g}})$$

The least-squares solution has the form $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\Delta) = \hat{c} \boldsymbol{\lambda}_\Delta / \|\boldsymbol{\lambda}_\Delta\|^2_{\ell_2}$ where

$$\hat{c} = \arg\min_c \|(c^\star - c)\boldsymbol{g} + \sigma \boldsymbol{z}\|_{\ell_2} \implies \hat{c} = c^\star + \sigma\gamma. \tag{99}$$

where $\gamma = \frac{\bar{\boldsymbol{g}}^T \boldsymbol{z}}{\|\boldsymbol{g}\|_{\ell_2}}$. Observe that $\sqrt{p}\gamma$ has Student's t-distribution with p degrees of freedom. Set $h, \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Now, observe that a fresh test sample with $y = \boldsymbol{x}^T \boldsymbol{\beta}^\star + \sigma\epsilon$ with $\boldsymbol{x} = g\boldsymbol{\lambda}$, the test error obeys

$$\mathcal{L}(\hat{\boldsymbol{\beta}}(\Delta)) = \mathbb{E}(y - \boldsymbol{x}_\Delta^T \hat{\boldsymbol{\beta}}(\Delta))^2 \tag{100}$$

$$= \mathbb{E}[((c^\star - \hat{c})g + \sigma\epsilon)^2] \tag{101}$$

$$= (\gamma^2 + 1)\sigma^2 \tag{102}$$

Now, observe that the minimum risk is obviously $\mathcal{L}(\boldsymbol{\beta}^\star) = \sigma^2$. Thus, the excess retraining risk becomes

$$\mathcal{L}(\hat{\boldsymbol{\beta}}(\Delta)) - \mathcal{L}(\boldsymbol{\beta}^\star) = \gamma^2 \sigma^2.$$

regardless of choice of Δ. Finally, averaging this risk over $\mathcal{S}$ returns $\mathbb{E}[\sigma^2\gamma^2] = \sigma^2/(n-2)$. Thus retraining risk has the fixed excess risk same as the one advertised in Lemma 1.

Pruning analysis: For pruning setting $\Delta = [p]$ above, we have that $\hat{\boldsymbol{\beta}} = \hat{c}\boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|^2_{\ell_2}$. This means that, for both Magnitude and Hessian pruning[3], pruned vector takes the form $\hat{\boldsymbol{\beta}}_s = \hat{c}\mathbb{T}_s(\boldsymbol{\lambda})/\|\boldsymbol{\lambda}\|^2_{\ell_2}$. Using the fact that $\boldsymbol{\lambda}^T \mathbb{T}_s(\boldsymbol{\lambda}) = \|\mathbb{T}_s(\boldsymbol{\lambda})\|^2_{\ell_2} = \zeta\|\boldsymbol{\lambda}\|^2_{\ell_2}$, we find

$$\mathcal{L}(\hat{\boldsymbol{\beta}}_s) - \mathcal{L}(\boldsymbol{\beta}^\star) = \mathbb{E}[(c^\star - \hat{c}\frac{\boldsymbol{\lambda}^T \mathbb{T}_s(\boldsymbol{\lambda})}{\|\boldsymbol{\lambda}\|^2_{\ell_2}})g + \sigma\epsilon)^2] - \sigma^2 \tag{103}$$

$$= \mathbb{E}[(c^\star - \hat{c}\zeta)g + \sigma\epsilon)^2] - \sigma^2 \tag{104}$$

$$= \mathbb{E}[(((1-\zeta)c^\star - \zeta\sigma\gamma)g + \sigma\epsilon)^2] - \sigma^2 \tag{105}$$

$$= ((1-\zeta)c^\star - \zeta\sigma\gamma)^2. \tag{106}$$

[2] This is because $\hat{\boldsymbol{\beta}} = c\boldsymbol{\lambda}$ for some scalar $c \neq 0$ as $\hat{\boldsymbol{\beta}}$ lies in the row space of $\boldsymbol{X}$. Then, Hessian/Magnitude-pruning would pick a nonzero support of $\hat{\boldsymbol{\beta}}$ which corresponds to the nonzero support of $\boldsymbol{\lambda}$.

[3] They yield the same result since diagonal covariance is proportional to $\boldsymbol{\lambda}$ in magnitude.

Finally, using zero-mean γ, we find

$$\mathbb{E}_{\mathcal{S}}[\mathcal{L}(\hat{\boldsymbol{\beta}}_s)] - \mathcal{L}(\boldsymbol{\beta}^\star) = \mathbb{E}_{\mathcal{S}}[((1-\zeta)c^\star - \zeta\sigma\gamma)^2] = (1-\zeta)^2 c^{\star 2} + \frac{\zeta^2\sigma^2}{n-2},$$

which concludes the proof after observing $c^{\star 2} = (\boldsymbol{\lambda}^T\boldsymbol{\beta}^\star)^2$. Here, we call $(1-\zeta)^2 c^{\star 2}$ "the error due to bias". The reason is that the predictable signal in the data is the noiseless component $\boldsymbol{x}^T\boldsymbol{\beta}^\star$. Pruning leads to an error in this predictable component by resulting in a biased estimate of the label (when conditioned on the random variable g which controls the signal). ■