Exploring Weight Importance and Hessian Bias in Model Pruning

Mingchen Li* Yahya Sattar† Christos Thrampoulidis‡ Samet Oymak§

June 19, 2020

Abstract

Model pruning is an essential procedure for building compact and computationally-efficient machine learning models. A key feature of a good pruning algorithm is that it accurately quantifies the relative importance of the model weights. While model pruning has a rich history, we still don't have a full grasp of the pruning mechanics even for relatively simple problems involving linear models or shallow neural nets. In this work, we provide a principled exploration of pruning by building on a natural notion of importance. For linear models, we show that this notion of importance is captured by covariance scaling which connects to the well-known Hessian-based pruning. We then derive asymptotic formulas that allow us to precisely compare the performance of different pruning methods. For neural networks, we demonstrate that the importance can be at odds with larger magnitudes and proper initialization is critical for magnitude-based pruning. Specifically, we identify settings in which weights become more important despite becoming smaller, which in turn leads to a catastrophic failure of magnitude-based pruning. Our results also elucidate that implicit regularization in the form of Hessian structure has a catalytic role in identifying the important weights, which dictate the pruning performance.

1 Introduction

Contemporary machine learning models such as deep neural networks often achieve good statistical accuracy at the expanse of large model sizes. On the other hand, a growing list of application domains demand compact and energy efficient machine learning models. Model pruning (i.e. sparsification) techniques are critical for addressing the challenge of building models that are simultaneously accurate and efficient. In this work, we investigate the fundamental principles of model pruning by exploring optimization dynamics and high-dimensional behavior of pruning approaches.

Pruning methods have a rich history and the literature on neural network pruning goes back to 1980's [44, 37, 27]. A fundamental approach in pruning is the accurate quantification of importance of each weight (i.e. connections) so that when a weight is removed, we can know how much the model will suffer. An intuitive approach is pruning by the weight magnitude, i.e. removing the weights below a certain threshold. A more principled approach is developing an importance (i.e. saliency) criteria which captures the sensitivity of the loss with respect to the weights. For instance, Optimal Brain Damage (OBD) [37] and Optimal Brain Surgeon [27, 28] calculate a Hessian-based importance criteria by adjusting the magnitudes. Despite its practical significance, a statistical understanding of pruning presents interesting challenges. Deep networks are often trained in an over-parameterized regime where the network size is well beyond what is necessary for achieving zero training error. Thus, network weights can interpolate the data in many ways and it is not immediately clear which weight gets the credit for learning. Pruning typically happens after training this large initial network possibly without any ℓ_1, ℓ_2 regularization. Deep nets may also converge to different solutions under different initialization or data preprocessing. These motivate a careful study of pruning mechanics: Which

*Email: mli176@ucr.edu. Computer Science and Engineering, University of California, Riverside.
†Email: ysatt001@ucr.edu. Electrical and Computer Engineering, University of California, Riverside.
‡Email: cthrampo@ucsb.edu. Electrical and Computer Engineering, University of California, Santa Barbara.
§Email: oymak@ece.ucr.edu. Electrical and Computer Engineering, University of California, Riverside.

approach works when? What is the role of initialization? Does over-parameterization affect the outcome and if so, can it be quantified?

Contributions: In this work, we explore model pruning, importance quantification and the role of Hessian structure in the pruning performance. We study three different importance criteria and corresponding pruning methods: (i) Hessian-based importance (HI) and pruning (HP), (ii) Magnitude-based importance (MI) and pruning (MP), and a third notion, which we call (iii) Natural importance (NI) and pruning (NP). For linear models and shallow neural-networks, we design a class of *equivalent problems* which enable us to assess the role of *Hessian structure* on the robustness and performance of different importance measures. Our specific contributions are as follows.

• **Understanding covariance bias and pruning performance:** For linear models, we introduce a class of problems where Hessian, which corresponds to the feature covariance matrix, is varied using diagonal scaling, while preserving target labels. We show that for over-determined problems HI and NI exhibit scaling invariance, whereas, MI is highly brittle. For over-parameterized problems, we show that scaling invariance no longer holds and the covariance/Hessian structure dictates the eventual pruning performance. We introduce analytical performance formulas, precisely capturing these phenomena, revealing that *implicit bias (as enforced by the Hessian structure) can boost HP while hurting MP*. Our approach also allows us to quantify *negative bias* when principal covariance directions are mis-aligned with the important weights. To the best of our knowledge, this is the first work that provides *exact analytical formulas for the performance of MP/HP*.

• **Understanding Hessian bias and the role of initialization:** For two-layer ReLU networks, we tackle the following question: *If both layers are very large and can interpolate the training data, who contributes more towards learning, who gets pruned eventually and at what cost?* We study these questions via a simple, yet insightful, network initialization model and show that the answers depend crucially on the Hessian structure which governs the training dynamics. Our empirical study reveals that: (i) HI is invariant to Hessian bias and (ii) as MI decreases, NI (which captures the training/test accuracy) increases. To explain this, we first show that magnitudes of the weights and magnitudes of their Hessians move in opposing directions and then establish a *"larger Hessian wins more"* theorem which accurately quantifies the relative contribution of different weight groups (e.g. layers) during training in terms of the Hessian bias.

1.1 Related work

Our work relates to the literature on neural net pruning, implicit regularization and over-parameterization. For analysis, we also use tools related to high-dimensional statistics [63, 51, 62, 29].

Implicit bias and over-parameterization: Contemporary deep networks often contain many more parameters than the dataset size and there is a growing literature dedicated to understanding their optimization/generalization properties and how over-parameterization can act as a catalyst. A key observation is that gradient-based algorithms are implicitly guided by the problem structure towards certain favorable solutions [3, 47]. For linear models, implicit bias phenomena is studied for various loss functions and algorithms (e.g. logistic loss converging to max-margin solution on separable data) [34, 58, 45]. Recent works show that such results continue to hold for nonlinear problems [23, 49, 5]. This line of works led to the more recent generalization/optimization guarantees for deep networks and their connections to random features [15, 2, 10, 8, 9, 40, 42]. A related line of work connects the benefits of over-parameterization to the double descent phenomena [46, 7, 6, 29].

Neural network pruning: The large model sizes in deep learning led to a substantial interest in model pruning/quantization [25, 27, 37]. The network pruning literature is diverse and involves various architectural, algorithmic, and hardware considerations [60, 26]. Recent works [26, 20, 19] use magnitude-based pruning criteria and achieve stellar performance. Related to over-parameterizarion, lottery ticket hypothesis [18] shows that large neural networks contain a small subset of favorable weights (for pruning) which can achieve similar performance as the original network when trained from same initialization. [66, 41] demonstrates that these subsets may achieve good test performance even without any training. [64] theoretically connects lottery tickets to over-parameterization. Various saliency-based approaches are proposed for neural net pruning [27, 28, 37, 12]. [38, 65] prune the network before training by the connection sensitivity or preserving the

gradient flow. [57] uses Jacobian-based pruning for recurrent networks. Furthermore, [1, 48, 35] uses ℓ_1 penalization for pruning and provides certain provable guarantees.

The rest of the paper is organized as follows. Section 2 sets the notation and introduces definitions on importance and pruning. Section 3 studies pruning for linear models, characterizes covariance bias, and introduces analytical performance formulas. Section 4 explores pruning for neural network and introduces results on optimization and pruning dynamics and Section 5 provides a discussion.

2 Problem Setup

We first set the notation. For a vector $\boldsymbol{v}$, we denote by $\|\boldsymbol{v}\|_{\ell_2}$ its Euclidean norm. $\odot$ returns the Hadamard (entrywise) product of two vectors. The (i,j)-th element of a matrix $\boldsymbol{M}$ is denoted by $\boldsymbol{M}_{i,j}$. The minimum singular value, spectral norm, and Frobenius norm of $\boldsymbol{M}$ is denoted by $\sigma_{\min}(\boldsymbol{M}), \|\boldsymbol{M}\|, \|\boldsymbol{M}\|_F$ respectively. $\boldsymbol{I}_k$ is the identity matrix of size k. The set $\{1,\dots,p\}$ is denoted by $[p]$. Given $\Delta \subset [p]$, $\bar{\Delta} = [p] - \Delta$ and $\boldsymbol{\theta}_\Delta$ denotes the vector obtained by setting the entries of $\boldsymbol{\theta}$ over $\bar{\Delta}$ to zero. $\mathbb{1}_p$ denotes the all ones vector in $\mathbb{R}^p$.

To proceed, we review definitions that will be discussed throughout. Our discussion will stem from the following definition which captures the impact of a set of weights on the loss function.

Definition 2.1 (Natural importance (NI)) *Given a loss function $\mathcal{L}(\boldsymbol{\theta})$, a reference vector $\boldsymbol{\theta}^R$ and set of indices $\Delta \subseteq [p]$, note that $\boldsymbol{\theta}^R_\Delta + \boldsymbol{\theta}_{\bar{\Delta}}$ is the vector obtained by replacing the entries of $\boldsymbol{\theta}$ at indices Δ by the corresponding entries of $\boldsymbol{\theta}^R$. The NI of the weights of $\boldsymbol{\theta}$ over Δ with respect to (w.r.t) $\mathcal{L}$ is defined as*

$$\mathcal{I}^N_\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}^R) = \mathcal{L}(\boldsymbol{\theta}^R_\Delta + \boldsymbol{\theta}_{\bar{\Delta}}) - \mathcal{L}(\boldsymbol{\theta}).$$

When $\boldsymbol{\theta}^R = 0$, we will use the notation $\mathcal{I}^N_\Delta(\boldsymbol{\theta})$. NI quantifies the *exact change in the loss and captures the problem-dependent nature of pruning.* The loss function in practice can be training (or test) loss or classification error. Here, the vector $\boldsymbol{\theta}^R$ aims to quantify the relative benefit of the change of weights of $\boldsymbol{\theta}$ with respect to a reference. For our purposes, we discuss two choices for the reference vector, which we call pruning and init-pruning, respectively.

- **(Regular) Pruning:** This is the standard pruning where the goal is to obtain a sparse model, thus the reference vector is $\boldsymbol{\theta}^R = 0$.
- **Init-Pruning:** Deep network training is often initialized from nonzero weights $\boldsymbol{\theta}_0$ such as random initialization or pre-trained weights. In this case, the contribution of different weights throughout the optimization can be assessed with respect to the point of initialization by choosing $\boldsymbol{\theta}^R = \boldsymbol{\theta}_0$.

We remark that, our characterization of the weight importance is similar to the saliency criterion which is widely used in literature on model pruning/trimming [37, 38, 44, 59]. Besides Definition 2.1, we also consider two other commonly-accepted importance criteria, which can be viewed as proxies for NI. To keep the discussion focused, the next two definitions only consider regular pruning i.e. $\boldsymbol{\theta}^R = 0$.

Definition 2.2 (Magnitude- and Hessian-based Importance) *Recall Def. 2.1. Suppose $\mathcal{L}$ is twice differentiable with Hessian $\mathcal{H}(\boldsymbol{\theta}) = \nabla^2\mathcal{L}(\boldsymbol{\theta})$. The MI $\mathcal{I}^M_\Delta(\boldsymbol{\theta})$ and HI $\mathcal{I}^H_\Delta(\boldsymbol{\theta})$ are defined as*

$$\mathcal{I}^M_\Delta(\boldsymbol{\theta}) = \sum_{i\in\Delta} \boldsymbol{\theta}_i^2 \quad \text{and} \quad \mathcal{I}^H_\Delta(\boldsymbol{\theta}) = \sum_{i\in\Delta} \mathcal{H}(\boldsymbol{\theta})_{i,i}\boldsymbol{\theta}_i^2. \tag{2.1}$$

Observe that our definition of HI is based on Optimal Brain Damage (OBD) [37]. Next, we define pruning based on a given importance criteria. A pruning algorithm identifies a set of weights with the smallest importance and sets them to zero.

Definition 2.3 (Pruning) *Given an importance criteria $\mathcal{I}$ (e.g. $\mathcal{I}^N, \mathcal{I}^M, \mathcal{I}^H$), a vector $\boldsymbol{\theta}$, and a target sparsity s, the pruning algorithm returns an s-sparse model $\Pi_s(\boldsymbol{\theta})$ (e.g. Π^N_s, Π^M_s, Π^H_s) where*

$$\Pi_s(\boldsymbol{\theta}) = \boldsymbol{\theta}_{\bar{\Delta}}, \quad \text{for} \quad \Delta = \arg\min_{|\Delta|=p-s} \mathcal{I}_\Delta(\boldsymbol{\theta}).$$

We will study and compare three different methods of pruning, namely, magnitude-based (MP), Hessian-based (HP) and natural pruning (NP). While NI captures the "true importance", NP is a combinatorially challenging subset selection problem and HP and MP provides computationally-efficient alternatives. For MP, this definition reduces to the hard-thresholding operation. Furthermore, MP and HP coincide when the Hessian has equal diagonal entries. We will focus our attention on pruning the trained model. Thus, typically we are interested in pruning the minimizer of the empirical (or population) loss. The following sections will relate these pruning methods, compare their performances, and explore the role of implicit regularization in pruning.

3 Importance and Covariance Bias for Linear Models

This section provides our results on pruning linear models and the role of feature covariance. Given a data distribution $\mathcal{D}$, we obtain a dataset $\mathcal{S}$ containing n i.i.d. samples $\mathcal{S} = (\boldsymbol{x}_i, y_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}$. Let $(\boldsymbol{x}, y) \sim \mathcal{D}$ be a generic sample. We assume $(\boldsymbol{x}, y) \in (\mathbb{R}^p, \mathbb{R})$ has finite second moments.

Covariance/Hessian structure: To understand the role of feature covariance (i.e. Hessian) on pruning, we introduce a class of datasets where the input features are shaped by an invertible diagonal scaling matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times p}$ while output label y is preserved. Here, a key motivation is modeling the properties of neural networks where the Hessian spectrum is not flat and often low-rank [29, 54, 53, 42, 4]. The intuition is that the importance of a weight captures the contribution of the corresponding input feature and should be invariant to how the feature is scaled. Perhaps surprisingly, we will also show this intuition fails for over-parameterized problems. To proceed, given $\boldsymbol{\Lambda}$, we consider a distribution $\mathcal{D}_{\boldsymbol{\Lambda}}$, with samples $(\boldsymbol{x}^{\boldsymbol{\Lambda}}, y) \sim \mathcal{D}_{\boldsymbol{\Lambda}}$ distributed as $(\boldsymbol{\Lambda}\boldsymbol{x}, y)$. Similarly, given $\mathcal{S}$, we generate a dataset $\mathcal{S}_{\boldsymbol{\Lambda}} = (\boldsymbol{x}_i^{\boldsymbol{\Lambda}}, y_i)_{i=1}^n$ where $\boldsymbol{x}_i^{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{x}_i$. We gather the data in matrix notation via

$$\boldsymbol{X}_{\boldsymbol{\Lambda}} = [\boldsymbol{x}_1^{\boldsymbol{\Lambda}}\ \boldsymbol{x}_2^{\boldsymbol{\Lambda}}\ \dots\ \boldsymbol{x}_n^{\boldsymbol{\Lambda}}]^T \in \mathbb{R}^{n\times p} \quad \text{and} \quad \boldsymbol{y} = [y_1\ y_2\ \dots\ y_n]^T \in \mathbb{R}^n.$$

To proceed, using quadratic loss, we define the empirical (training) and population (test) losses,

$$\hat{\mathcal{L}}_{\boldsymbol{\Lambda}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^n (y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i^{\boldsymbol{\Lambda}})^2 = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\Lambda}}\boldsymbol{\theta}\|_{\ell_2}^2 \quad \text{and} \quad \mathcal{L}_{\boldsymbol{\Lambda}}(\boldsymbol{\theta}) := \mathbb{E}[(y - \boldsymbol{\theta}^T\boldsymbol{x}^{\boldsymbol{\Lambda}})^2]. \tag{3.1}$$

Let $\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$ be the global minima of $\hat{\mathcal{L}}_{\boldsymbol{\Lambda}}$ and $\mathcal{L}_{\boldsymbol{\Lambda}}$ respectively. Let $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$ be the population covariance and $\boldsymbol{b} = \mathbb{E}[\boldsymbol{x}y]$ be the cross-correlation. For simplicity, we assume $\boldsymbol{\Sigma}$ is full-rank. *We will drop the subscript* $\boldsymbol{\Lambda}$ *when* $\boldsymbol{\Lambda} = \boldsymbol{I}_p$. The solutions $\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$ are given by

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}} = \boldsymbol{X}_{\boldsymbol{\Lambda}}^{\dagger}\boldsymbol{y} \quad \text{and} \quad \bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{b},$$

respectively, where $\dagger$ denotes the pseudo-inverse. The following lemma is instructive in understanding the weight importance and invariance to feature scaling for the least-squares problem above (3.1).

Lemma 3.1 (Pruning with Population) *Let* $\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$ *be the minimizer of population loss and fix* $\Delta \subseteq [p]$. *NI* $\mathcal{I}_\Delta^N(\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}})$ *and HI* $\mathcal{I}_\Delta^H(\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}})$ *w.r.t. population loss* $\mathcal{L}_{\boldsymbol{\Lambda}}$ *are invariant under invertible diagonal* $\boldsymbol{\Lambda}$. *If the covariance* $\boldsymbol{\Sigma}$ *is also diagonal, then NI and HI are equal. In contrast, MI is* $\boldsymbol{\Lambda}$ *dependent via* $\mathcal{I}_\Delta^M(\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}) = \sum_{i\in\Delta} \boldsymbol{\Lambda}_{i,i}^{-2}\bar{\boldsymbol{\theta}}_i^2$ *where* $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}^{\boldsymbol{I}_p}$ *is the original model.*

This lemma states that NI and HI are invariant to scaling and coincide when features are uncorrelated. On the other hand, MI suffers from feature scaling. As the features get larger, the corresponding weight decreases which results in an artificial decrease in importance. This highlights a fundamental shortcoming of MP and necessity of feature normalization, which was previously discussed in the literature [56, 31, 33, 17]. In Sections 3.1 and 4, we will see that MP fails as soon as the problem is not well-conditioned either in terms of covariance spectrum or neural network initialization.

Invariance to feature scaling is a property of over-determined problems ($n > p$) which admit unique solution (population loss is a special case with $n = \infty$). Focusing on training loss, suppose $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is not rank deficient. Then, the minimum-norm solution $\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$ has the form

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}} = \begin{cases} \boldsymbol{\Lambda}^{-1}\hat{\boldsymbol{\theta}}, & \text{when} \quad n \geq p, \\ \boldsymbol{\Lambda}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\Lambda}^2\boldsymbol{X}^T)^{-1}\boldsymbol{y}, & \text{otherwise.} \end{cases} \tag{3.2}$$

When $n \geq p$, we trivially have $(\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}})^T\boldsymbol{x}^{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\theta}}^T\boldsymbol{x}$, thus $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$ achieve the exact same test/training loss. On the other hand, for over-parameterized problems ($n < p$), which is the regime of interest for neural network pruning, this is no longer the case, and we will see that $\boldsymbol{\Lambda}$ plays a critical role in the eventual test performance as it dictates which solution the optimization problem selects.

3.1 Characterizing Pruning Performance and Covariance Bias

In this section, we provide analytical formulas which enable us to compare different pruning methods and assess implicit covariance bias when $n < p$ under a realizable dataset model. Suppose $(\boldsymbol{x}_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ so that $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}} := \mathbb{E}[\boldsymbol{x}^{\boldsymbol{\Lambda}}(\boldsymbol{x}^{\boldsymbol{\Lambda}})^T] = \boldsymbol{\Lambda}^2$. Given a ground-truth vector $\bar{\boldsymbol{\theta}} \in \mathbb{R}^p$ (which corresponds to the population minima), we generate the labels via $y = \boldsymbol{x}^T\bar{\boldsymbol{\theta}} + z$ and

$$y_i = \boldsymbol{x}_i^T\bar{\boldsymbol{\theta}} + z_i \quad \text{for} \quad 1 \leq i \leq n,$$

where $z, (z_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ are the additive noise. We will study the minimum norm least-squares solution (3.2) which is also the solution gradient descent converges when initialized from zero. To assess pruning performance, we need to quantify the test loss of the pruned solution $\Pi_s(\hat{\boldsymbol{\theta}})$.

Connection to denoising: We accomplish this by *relating the test loss of the pruned model to the risk of a simple denoising problem.* In essence, this denoising question is as follows: Given noisy measurements $\boldsymbol{\theta}_{\text{nsy}} = \bar{\boldsymbol{\theta}} + \boldsymbol{g}$ of a ground-truth vector $\bar{\boldsymbol{\theta}}$ with $\boldsymbol{g} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_p)$, what is the pruning error $\mathbb{E}[\|\Pi_s(\boldsymbol{\theta}_{\text{nsy}}) - \bar{\boldsymbol{\theta}}\|_{\ell_2}^2]$? Note that this error typically doesn't have a closed form answer as hard-thresholding is not a continuous function, however, it greatly simplifies the original problem of solving least-squares. We also note that if one uses soft-thresholding (i.e. shrinkage) operator for pruning, closed form solution is available. The fundamental connection between denoising and linear inverse problems are studied for under-parameterized least-squares and lasso regression [13, 14]. Our connection to denoising is established by connecting the distribution of the $\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$ to an auxiliary distribution described below.

Definition 3.2 (Auxiliary distribution) *Fix $p > n \geq 1$ and set $\kappa = p/n > 1$. Given $\sigma > 0$, positive definite diagonal matrix $\boldsymbol{\Lambda}$ and ground-truth vector $\bar{\boldsymbol{\theta}}$, define the unique non-negative terms $\Xi, \Gamma, \boldsymbol{\zeta} \in \mathbb{R}^p$ and $\boldsymbol{\gamma} \in \mathbb{R}^p$ as follows*

$$\Xi > 0 \quad \textit{is the unique solution of} \quad 1 = \frac{\kappa}{p}\sum_{i=1}^p \frac{1}{1 + (\Xi\boldsymbol{\Lambda}_{i,i}^2)^{-1}}, \tag{3.3}$$

$$\Gamma = \frac{\sigma^2 + \sum_{i=1}^p \boldsymbol{\zeta}_i^2\bar{\boldsymbol{\theta}}_i^2}{\kappa(1 - \frac{\kappa}{p}\sum_{i=1}^p (1 + (\Xi\boldsymbol{\Lambda}_{i,i}^2)^{-1})^{-2})},$$

$$\boldsymbol{\zeta}_i = \frac{1}{1 + \Xi\boldsymbol{\Lambda}_{i,i}^2} \quad \textit{and} \quad \boldsymbol{\gamma}_i = \frac{\kappa\sqrt{\Gamma}}{1 + (\Xi\boldsymbol{\Lambda}_{i,i}^2)^{-1}} \quad \textit{for} \quad 1 \leq i \leq p.$$

Let $\boldsymbol{h} \sim \mathcal{N}(0, \frac{1}{p}\boldsymbol{I}_p)$. Define the auxiliary vector $\boldsymbol{\theta}_{aux}^{\boldsymbol{\Lambda}}$ of the ground-truth $\bar{\boldsymbol{\theta}}$ as

$$\boldsymbol{\theta}_{aux}^{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}^{-1}[(\mathbb{1}_p - \boldsymbol{\zeta}) \odot \bar{\boldsymbol{\theta}} + \boldsymbol{\gamma} \odot \boldsymbol{h}]. \tag{3.4}$$

We remark that this definition can be adapted to asymptotic setup $p \to \infty$ assuming covariance spectrum converges (e.g. discrete sum over entries converges to an integral). In the special case of identity covariance

($\boldsymbol{\Sigma} = \boldsymbol{I}_p$), $\boldsymbol{\theta}_{\text{aux}}$ reduces to $\boldsymbol{\theta}_{\text{aux}} = \frac{1}{\kappa}\bar{\boldsymbol{\theta}} + \sqrt{\frac{\sigma^2}{\kappa-1} + \frac{(\kappa-1)\|\bar{\boldsymbol{\theta}}\|_{\ell_2}^2}{\kappa^2}}\boldsymbol{h}$. This distribution arises from applying Convex Gaussian Min-Max Theorem (CGMT) [22, 21, 63, 51, 62] to over-parameterized least-squares. CGMT provides a framework for predicting the asymptotic properties of optimization problems involving random matrices by connecting them to simpler auxiliary optimizations involving random vectors (some example applications [43, 11, 55, 52]). Thus, based on CGMT, $\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$ and the auxiliary vector $\boldsymbol{\theta}^{\boldsymbol{\Lambda}}_{\text{aux}}$ are expected to have similar distributional properties and $\boldsymbol{\theta}^{\boldsymbol{\Lambda}}_{\text{aux}}$ can be used as a proxy to capture the properties of $\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$. In supplementary, we discuss to what extent this distributional similarity can be formalized (e.g. for Lipschitz functions). Note that, after solving for $\boldsymbol{\zeta}, \boldsymbol{\gamma}$ in (3.3), we can sample from the auxiliary distribution which is a noisy version of $\bar{\boldsymbol{\theta}}$ which connects us to denoising. To proceed, our analytic formulas for the test error of an s-sparse model via MP and HP takes the following form:

$$\text{MP loss: } \mathbb{E}_{\boldsymbol{h}}[\|\boldsymbol{\Lambda}\Pi_s^M(\boldsymbol{\theta}^{\boldsymbol{\Lambda}}_{\text{aux}}) - \bar{\boldsymbol{\theta}}\|_{\ell_2}^2] + \sigma^2, \quad \text{HP loss: } \mathbb{E}_{\boldsymbol{h}}[\|\Pi_s^M(\boldsymbol{\Lambda}\boldsymbol{\theta}^{\boldsymbol{\Lambda}}_{\text{aux}}) - \bar{\boldsymbol{\theta}}\|_{\ell_2}^2] + \sigma^2.$$

Next, we verify our performance prediction and study the role of covariance structure $\boldsymbol{\Lambda}$. We generate $\bar{\boldsymbol{\theta}}$ with polynomially decaying entries, specifically $\bar{\boldsymbol{\theta}}_i = 1/(1+4i/p)^2$, and then scale it to unit Euclidian norm. Recall that original covariance is identity, thus initial larger entries of $\bar{\boldsymbol{\theta}}$ are more important for population risk. In our experiments, we parameterize $\boldsymbol{\Lambda}$ by a scalar λ and set it as

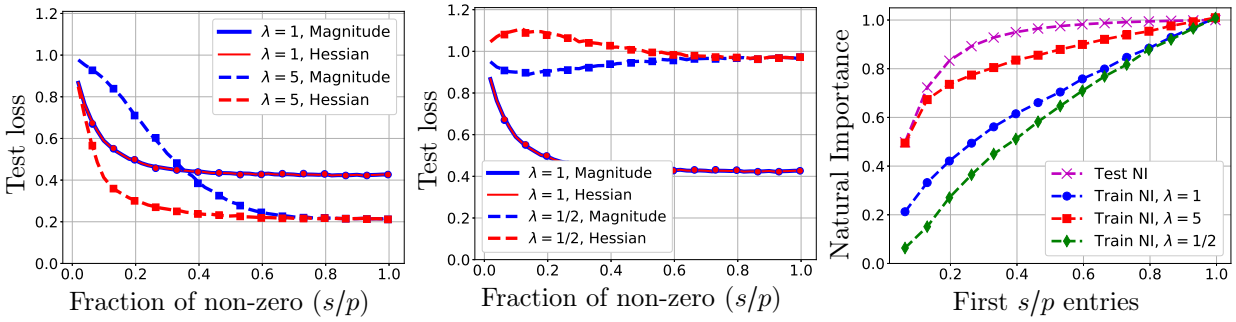
$$\boldsymbol{\Lambda}_{i,i} = \begin{cases} \lambda & \text{if} \quad 1 \le i \le p/10, \\ 1 & \text{if} \quad i > p/10. \end{cases} \tag{3.5}$$

This choice modifies the most important 10% weights of the problem. We consider $\lambda \in \{1/2, 1, 5\}$. As formalized in Thm. 4.3, when $\lambda > 1$, we expect a positive covariance bias since important weights are aligned with the principal directions of the covariance and are easier to learn. In Figures 1a and 1b, the lines are the analytical predictions based on Definition 3.2 and the markers are performance of the actual least-squares solution which nicely match for all pruning methods and λ. Figure 1a contrasts $\lambda = 1$ and $\lambda = 5$. For $\lambda = 1$, MP and HP coincide as the Hessian is identity. However when $\lambda = 5$, HP performs much better than $\lambda = 1$ for all sparsity levels. MP drastically fails for small sparsity levels as the initial weights of $\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$ are important but small due to the λ-scaling thus MP inaccurately ignores them. Decreasing magnitudes with increasing λ is more clear for under-parameterized case (via (3.2)) however $\boldsymbol{\Lambda}^{-1}$ dependence is also visible in (3.4). Fig 1b additionally highlights $\lambda = 1/2$ which reduces the covariance and scales up the coefficients of the important weights. This leads to a negative bias because covariance structure guides the solution away from important weights. While both MP and HP performs worse than $\lambda = 1$ case, HP performs worse due to additional penalization of the initial important weights. Finally, covariance bias is visualized in Figure 1c which displays the test NI (for $\bar{\boldsymbol{\theta}}$) and the training NI's (for $\hat{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}$) of the first s weights. When $\lambda = 5$, initial weights, which are important for test, have a larger training NI. As λ gets smaller, remaining weights, which are not as important for test, have larger say during training and pruning performance degrades. Our Theorem 4.3 formalizes these by quantifying the contributions of different weights during training.

4 Hessian Bias and the Role of Initialization for Neural Nets

This section extends our discussion of importance and pruning to another fundamental model class: neural networks with one-hidden layer. Suppose input dimension is d, output dimension is K and the network has m hidden units. Such a network with ReLU activation is given by $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{V}\text{ReLU}(\boldsymbol{W}\boldsymbol{x})$, where $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{K\times m}$ are the input and output layers respectively and $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{V}) \in \mathbb{R}^{p=(d+K)m}$ is the vector composed of the entries of $\boldsymbol{W}, \boldsymbol{V}$. Let $\Delta_{\boldsymbol{W}}$ and $\Delta_{\boldsymbol{V}}$ denote the index of the entries of $\boldsymbol{W}, \boldsymbol{V}$ in $\boldsymbol{\theta}$. Given a dataset $\mathcal{S} = (\boldsymbol{x}_i, y_i)_{i=1}^n$ and loss ℓ, we minimize

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)). \tag{4.1}$$



(a) Pruning with $\lambda \in \{1, 5\}$ (b) Pruning with $\lambda \in \{1, 1/2\}$ (c) NI of the first s entries

Figure 1: In (a) and (b), the lines are the analytical prediction from Def 3.2 and markers are the actual min-norm solution. $p = 1000, \kappa = p/n = 5/3$ and $\sigma = 0.1$. (c) Natural importance associated with the first s weights. For $\lambda = 5$, (a) HP achieves better performance and (c) training and test NI have a better match.

Equivalent networks: To study neural net pruning and initialization, we shall consider a class of networks $\boldsymbol{\theta}^\lambda = (\lambda \boldsymbol{W}, \lambda^{-1}\boldsymbol{V})$ generated from a base network $\boldsymbol{\theta}^1 = (\boldsymbol{W}, \boldsymbol{V})$. Observe that all vectors $\boldsymbol{\theta}^\lambda$ implement the same function due to the linearity of ReLU however magnitudes of layers are varying. The following lemma shows how the parameter λ affects MI, HI, and Hessian.

Lemma 4.1 *Consider the loss (4.1) and class of networks $(\boldsymbol{\theta}^\lambda)_{\lambda>0}$. For all $\lambda > 0$, MI, HI and partial Hessians w.r.t. input/output layer weights $\boldsymbol{W}, \boldsymbol{V}$ obey*

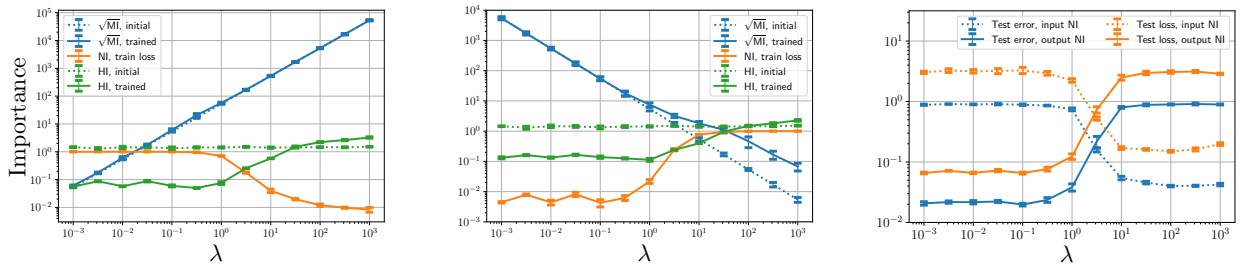
$$\mathcal{I}^M_{\Delta_{\boldsymbol{W}}}(\boldsymbol{\theta}^\lambda) = \lambda^2 \mathcal{I}^M_{\Delta_{\boldsymbol{W}}}(\boldsymbol{\theta}^1) \quad \text{and} \quad \mathcal{I}^M_{\Delta_{\boldsymbol{V}}}(\boldsymbol{\theta}^\lambda) = \lambda^{-2} \mathcal{I}^M_{\Delta_{\boldsymbol{V}}}(\boldsymbol{\theta}^1),$$
$$\mathcal{I}^H_{\Delta_{\boldsymbol{W}}}(\boldsymbol{\theta}^\lambda) = \mathcal{I}^H_{\Delta_{\boldsymbol{W}}}(\boldsymbol{\theta}^1) \quad \text{and} \quad \mathcal{I}^H_{\Delta_{\boldsymbol{V}}}(\boldsymbol{\theta}^\lambda) = \mathcal{I}^H_{\Delta_{\boldsymbol{V}}}(\boldsymbol{\theta}^1), \tag{4.2}$$
$$\frac{\partial^2}{\partial^2 \boldsymbol{W}} \mathcal{L}(\boldsymbol{\theta}^\lambda) = \lambda^{-2} \frac{\partial^2}{\partial^2 \boldsymbol{W}} \mathcal{L}(\boldsymbol{\theta}^1) \quad \text{and} \quad \frac{\partial^2}{\partial^2 \boldsymbol{V}} \mathcal{L}(\boldsymbol{\theta}^\lambda) = \lambda^2 \frac{\partial^2}{\partial^2 \boldsymbol{V}} \mathcal{L}(\boldsymbol{\theta}^1). \tag{4.3}$$

In words, increasing λ increases MI, preserves HI, and decreases the Hessian magnitude for the input layer and has the reverse effect on the output layer. Suppose we train the network from initializations $\boldsymbol{\theta}^\lambda$ on (4.1). What happens at the end of the training as a function of λ? Does eventual MI and HI exhibit similar behavior to initialization? What about NI?

To answer these, in Figure 2, we conduct an empirical study on MNIST by training a one-hidden layer network with cross-entropy loss. Here $m = 1024$, $d = 784$ and $K = 10$. We set $\boldsymbol{\theta}^1_{\text{init}} = (\boldsymbol{W}_{\text{init}}, \boldsymbol{V}_{\text{init}})$ with *He normal* initialization [30]. We then train networks with λ-scaled initializations $\boldsymbol{\theta}^\lambda_{\text{init}} = (\lambda \boldsymbol{W}_{\text{init}}, \lambda^{-1}\boldsymbol{V}_{\text{init}})$. Let $\boldsymbol{\theta}^{(\lambda)}_{\text{final}} = (\boldsymbol{W}^{(\lambda)}_{\text{final}}, \boldsymbol{V}^{(\lambda)}_{\text{final}})$ be the final model obtained by training until interpolation to training data (or maximum 150 epochs). Figures 2a and 2b display MI, HI, and NI for input and output layers respectively. Here, for NI, we use *Init-Pruning* and quantify importance of a layer (e.g. $\boldsymbol{W}^{(\lambda)}_{\text{final}}$) with respect to its initial weights (e.g. $\boldsymbol{W}^\lambda_{\text{init}} = \lambda \boldsymbol{W}_{\text{init}}$). Observe that, regular pruning is not informative as setting a layer to zero kills the network output.

Understanding MI and HI: Figures 2a and 2b show that initial and final MI exhibit a near perfect match. The initial HI stays constant as predicted by Lemma 4.1. Final HI increases with λ for both layers, however it can be verified that the ratio of HI between input and output layers is approximately preserved. Perhaps surprisingly, Lemma 4.1 seems to predict not only the initial importance but also the MI/HI of the final network. Fortunately, this can be mostly explained by the optimization dynamics of wide and large networks where gradient descent finds a global optima close to initialization and *final weights (and Hessian) do not deviate much from initial ones* [10, 4, 50, 16, 32, 2, 39].

In Figures 3a and 3b, we first prune $\boldsymbol{\theta}^{(\lambda)}_{\text{final}}$ to a fixed nonzero fraction and then retrain the pruned weights from the same initialization (i.e. [20]). MP is only competitive with HP when $\lambda = 1$ where input and output layer entries have similar magnitude due to He initialization. In Fig. 3a, as λ grows output layer becomes small and gets fully pruned. As λ gets smaller, eventually input layer is fully pruned. Here, what is rather



(a) Importance of input layer $\boldsymbol{W}^{\lambda}_{\text{final}}$ (b) Importance of output layer $\boldsymbol{V}^{\lambda}_{\text{final}}$ (c) Natural importances w.r.t. test

Figure 2: (a) and (b) show the comparison of importance criterias for input and output layers when training with a shallow network with cross entropy and with initializations $\boldsymbol{\theta}^{\lambda}_{\text{init}} = (\lambda \boldsymbol{W}_{\text{init}}, \lambda^{-1}\boldsymbol{V}_{\text{init}})$. (c) shows the NI w.r.t. test classification error and loss obtained by setting one of the layers to its initialization.

remarkable is the robustness of HP for full range of λ choices which arises from (4.2). Arguably, HI being invariant to λ makes it more attractive than NI as it avoids the issue of *layer death* i.e. all of the weights in a layer getting pruned. Figure 3c visualizes the fraction of unpruned weights in input and output layers for various λ. HP (solid) curves are stable whereas MP (dotted) curves are highly volatile and easily hit zero except a narrow region. We note that, an alternative way of avoiding layer death is pruning layers individually. Supplementary provides further experiments on this for completeness.

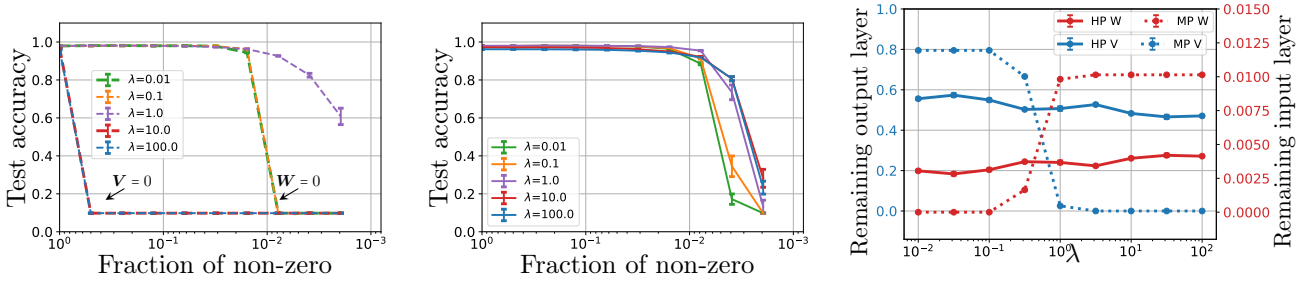
Understanding NI and optimization dynamics: If our shallow network is sufficiently wide, each layer (or large groups of weights) can individually fit the training dataset. This can be viewed as a competition between the layers and a natural question is how much a layer contributes to the learning. This question is answered by NI. In Figure 2a orange line displays the change in input layer NI (with $\mathcal{L}$ of Def. 2.1 is training loss) which demonstrates that NI is decreasing function of λ and moves in the opposite direction to MI. Figure 2c verifies the same NI behavior for test loss and test error. Specifically, for large λ, input layer is responsible for most of the test accuracy and for small λ, it is the output layer. Our key technical contribution in this section is *providing a theoretical explanation to this NI behavior and relating it to optimization dynamics*. In essence, we will connect NI to the only feature in Lemma 4.1 that exhibit similar behavior, the Hessian. Below we state our result on the Hessian and NI relation in terms of Polyak-Lojasiewicz (PL) condition [36].

Definition 4.2 (Partial PL and Smoothness (PPLS)) *Let $\mathcal{L}(\boldsymbol{\theta})$ be a loss function satisfying $\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = 0$. Given an index set $\Delta \subset [p]$, we say that PPLS holds with parameter $L \geq \mu \geq 0$ if partial derivative $\frac{\partial}{\partial \boldsymbol{\theta}_\Delta}\mathcal{L}(\boldsymbol{\theta})$ is L-Lipschitz function of $\boldsymbol{\theta}_\Delta$ and obeys $\|\frac{\partial}{\partial \boldsymbol{\theta}_\Delta}\mathcal{L}(\boldsymbol{\theta})\|_{\ell_2}^2 \geq 2\mu\mathcal{L}(\boldsymbol{\theta})$.*

While PL allows for non-convex optimization, when specialized to strong convexity, Partial PL condition provides a lower bound on the submatrix of Hessian induced by the set Δ. Regular PL condition guarantees global convergence of gradient descent, thus if PPLS holds over Δ, training only over Δ is sufficient to achieve zero loss. A good example of PPLS is linear regression with two feature sets $\boldsymbol{X}_1 \in \mathbb{R}^{n\times p_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n\times p_2}$ with $p_1, p_2 \geq n$ where we fit

$$\mathcal{L}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}=[\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2]} 0.5\|\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\theta}_1 - \boldsymbol{X}_2\boldsymbol{\theta}_2\|_{\ell_2}^2. \tag{4.4}$$

$\mathcal{L}$ satisfies PPLS over $[p_1] = \{1, \ldots, p_1\}$ with parameters $L_1 = \|\boldsymbol{X}_1\|^2$ and $\mu_1 = \sigma_{\min}(\boldsymbol{X}_1)^2$. For randomly initialized over-parameterized networks, each layer solves a kernel regression and would satisfy PPLS under mild conditions on the dataset [10, 16, 32, 2, 50] . Specifically, linearized neural network dynamics on $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{V})$ connects to the regression task (4.4) via the Taylor expansion around initialization where input and output layers have linearized features arising from the Jacobian map given by $\boldsymbol{X}_{\boldsymbol{W}} = [\frac{\partial f(\boldsymbol{x}_1)}{\partial \boldsymbol{W}} \ \ldots \ \frac{\partial f(\boldsymbol{x}_n)}{\partial \boldsymbol{W}}]^T \in \mathbb{R}^{n\times md}$ and $\boldsymbol{X}_{\boldsymbol{V}} = [\frac{\partial f(\boldsymbol{x}_1)}{\partial \boldsymbol{V}} \ \ldots \ \frac{\partial f(\boldsymbol{x}_n)}{\partial \boldsymbol{V}}]^T \in \mathbb{R}^{n\times Km}$. The following theorem provides a theoretical explanation of NI behavior via PPLS by quantifying relative contributions of different sets of weights.



(a) Test accuracy using MP. (b) Test accuracy using HP. (c) Remaining nonzeros for $\boldsymbol{V}, \boldsymbol{W}$

Figure 3: In (a) and (b), we first apply MP and HP on the network weights $\boldsymbol{\theta}_{\text{final}}^{(\lambda)}$ for varying pruning levels s/p where $p = (K+d)m$. We then retrain the pruned network from same initial nonzeros (lottery initialization of [20]) and display the test accuracy. HP is more stable compared to MP under λ-scaled initializations. (c) Visualization of the remaining fractions of nonzero weights in input (red) and output (blue) layers after pruning the network to 1% sparsity. Nonzero counts in both layers are stable under HP but rapidly change in MP as a function of λ.

Theorem 4.3 (Larger Hessian Wins More) *Suppose the entries of $\boldsymbol{\theta} \in \mathbb{R}^p$ are union of D non-intersecting sets $(\Delta_i)_{i=1}^D \subset [p]$ and PPLS holds over Δ_i with parameters $L_i \geq \mu_i \geq 0$ for all i. Set $\mu = \sum_{i=1}^D \mu_i$ and $L = \sum_{i=1}^D L_i$. Starting from a point $\boldsymbol{\theta}_0$, and using a learning rate $\eta \leq 1/L$, run gradient iterations $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_\tau)$. For all iterates τ, the loss obeys $\mathcal{L}(\boldsymbol{\theta}_\tau) \leq (1-\eta\mu)^\tau \mathcal{L}(\boldsymbol{\theta}_0)$. Furthermore, setting $\kappa = L_i/\mu$, the following bounds hold for Δ_i and $\bar{\Delta}_i = [p] - \Delta_i$ for all $\tau \geq 0$*

$$\|\boldsymbol{\theta}_{\Delta_i,\tau} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2}^2 \leq 8\kappa \mathcal{L}(\boldsymbol{\theta}_0)/\mu, \tag{4.5}$$

$$\mathcal{I}_{\Delta_i}^N(\boldsymbol{\theta}_\tau, \boldsymbol{\theta}_0)/\mathcal{L}(\boldsymbol{\theta}_0) \leq 8\kappa^2 + 4\kappa(1-\eta\mu)^{\tau/2}, \tag{4.6}$$

$$\mathcal{I}_{\bar{\Delta}_i}^N(\boldsymbol{\theta}_\tau, \boldsymbol{\theta}_0)/\mathcal{L}(\boldsymbol{\theta}_0) \geq 1 - 8\kappa^2 - 4\kappa - (1-\eta\mu)^\tau. \tag{4.7}$$

In words, this theorem captures the NI of a subset of weight throughout the training via the upper and lower bounds (4.6) and (4.7). For the experiments in Fig. 2, based on (4.3) of Lemma 4.1, PPLS parameters $(\mu_{\boldsymbol{W}}, L_{\boldsymbol{W}})$ of the input layer decay as λ^{-2} and output layer parameters grow as λ^2. Thus, assuming $\lambda \geq 1$, for output layer we have $\kappa = L_{\boldsymbol{W}}/(\mu_{\boldsymbol{W}} + \mu_{\boldsymbol{V}}) \sim \lambda^{-4}$ and, using (4.6) with $\tau = \infty$, NI is expected to decay as $\kappa^2 \sim \lambda^{-8}$ (e.g. for quadratic loss). Similarly, NI of the input layer is lower bounded via (4.7) which grows as $1 - \mathcal{O}(\lambda^{-4})$. Finally, for small λ, we have the reversed upper/lower bounds. In summary, our Theorem 4.3 successfully explains the empirical NI behavior in Fig. 2.

(4.5) generalizes the "short distance from initialization" results of [49, 24] by controlling individual subsets of weights and also provides a bound on MI when $\boldsymbol{\theta}_0 = 0$. As explained in supplementary, *this theorem is tight up to local (L_i/μ_i) and global (L/μ) condition numbers and accurately captures the relative contributions of the weights $(\boldsymbol{\theta}_{\Delta_i})_{i=1}^D$*. Observe that this theorem considers the Init-Pruning (w.r.t. $\boldsymbol{\theta}_0$) which is better suited for assessing optimization dynamics.

Note that the bounds of Thm 4.3 greatly simplify at the global minima ($\tau \to \infty$). As mentioned earlier, training NI of Figure 1c can be explained by Thm 4.3. In essence, scaling up a set of features increase their covariance (and PPLS parameter μ) increasing the NI w.r.t. training loss.

5 Conclusion

We provided a principled exploration of model pruning for linear models and shallow networks. Our work reveals and formalizes the importance of Hessian/covariance structure for pruning over-parameterized models. We found that magnitude-based pruning is very brittle and requires good normalization whereas Hessian-based pruning is robust to problem structure. We also derived the first analytical performance formulas exactly capturing pruning for linear models which enabled us to do a thorough comparison between different methods. There are several interesting open directions. Can we derive similar sharp performance bounds for pruning

random features or neural networks? What are the optimal initialization strategies for deep nets to enable ideal pruning performance?

References

[1] Aghasi, A., Abdi, A., Nguyen, N., and Romberg, J. Net-trim: Convex pruning of deep neural networks with performance guarantee. In *Advances in Neural Information Processing Systems* (2017), pp. 3177–3186.

[2] Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning* (2019), pp. 242–252.

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

[4] Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584* (2019).

[5] Azizan, N., and Hassibi, B. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. In *International Conference on Learning Representations* (2019).

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

[7] Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571* (2019).

[8] Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning* (2018), pp. 541–549.

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

[10] Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems* (2019), pp. 2933–2943.

[11] Deng, Z., Kammoun, A., and Thrampoulidis, C. A model of double descent for high-dimensional logistic regression. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), IEEE, pp. 4267–4271.

[12] Dong, X., Chen, S., and Pan, S. Learning to prune deep neural networks via layer-wise optimal brain surgeon. In *Advances in Neural Information Processing Systems* (2017), pp. 4857–4867.

[13] Donoho, D. L., Johnstone, I., and Montanari, A. Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE transactions on information theory 59*, 6 (2013), 3396–3433.

[14] Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences 106*, 45 (2009), 18914–18919.

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

[16] Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054* (2018).

[17] Ekenel, H. K., and Stiefelhagen, R. Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)* (2006), pp. 34–34.

[18] Frankle, J., and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations* (2019).

[19] Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Stabilizing the lottery ticket hypothesis. *arXiv, page*.

[20] Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. The lottery ticket hypothesis at scale. *arXiv preprint arXiv:1903.01611* (2019).

[21] Gordon, Y. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics 50*, 4 (1985), 265–289.

[22] Gordon, Y. *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer, 1988.

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

[24] Gupta, C., Balakrishnan, S., and Ramdas, A. Path length bounds for gradient descent and flow. *arXiv preprint arXiv:1908.01089* (2019).

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

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

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

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

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

[30] He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 1026–1034.

[31] Ioffe, S., and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

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

[33] Jayalakshmi, T., and Santhakumaran, A. Statistical normalization and back propagation for classification. *International Journal of Computer Theory and Engineering 3*, 1 (2011), 1793–8201.

[34] Ji, Z., and Telgarsky, M. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300* (2018).

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

[36] Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2016), Springer, pp. 795–811.

[37] LeCun, Y., Denker, J. S., and Solla, S. A. Optimal brain damage. In *Advances in neural information processing systems* (1990), pp. 598–605.

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

[39] Li, M., Soltanolkotabi, M., and Oymak, S. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *arXiv preprint arXiv:1903.11680* (2019).

[40] Liang, T., and Rakhlin, A. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387* (2018).

[41] Malach, E., Yehudai, G., Shalev-Shwartz, S., and Shamir, O. Proving the lottery ticket hypothesis: Pruning is all you need. *arXiv preprint arXiv:2002.00585* (2020).

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

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

[44] Mozer, M. C., and Smolensky, P. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In *Advances in neural information processing systems* (1989), pp. 107–115.

[45] Nacson, M. S., Srebro, N., and Soudry, D. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics* (2019), pp. 3051–3059.

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

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

[48] Oymak, S. Learning compact neural networks with regularization. *International Conference on Machine Learning* (2018).

[49] Oymak, S., and Soltanolkotabi, M. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning* (2019), pp. 4951–4960.

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

[51] Oymak, S., Thrampoulidis, C., and Hassibi, B. The squared-error of generalized lasso: A precise analysis. *arXiv preprint arXiv:1311.0830* (2013).

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

[53] Papyan, V. The full spectrum of deep net hessians at scale: Dynamics with sample size. *arXiv preprint arXiv:1811.07062* (2018).

[54] Sagun, L., Evci, U., Guney, V. U., Dauphin, Y., and Bottou, L. Empirical analysis of the hessian of over-parametrized neural networks. In *International Conference on Learning Representations* (2018).

[55] Salehi, F., Abbasi, E., and Hassibi, B. A precise analysis of phasemax in phase retrieval. In *2018 IEEE International Symposium on Information Theory (ISIT)* (2018), IEEE, pp. 976–980.

[56] Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems* (2018), pp. 2483–2493.

[57] Shunshi Zhang, M., and Stadie, B. One-shot pruning of recurrent neural networks by jacobian spectrum evaluation. *arXiv* (2019), arXiv–1912.

[58] Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research 19*, 1 (2018), 2822–2878.

[59] Sum, J., Leung, C.-S., Young, G. H., and Kan, W.-K. On the kalman filtering method in neural network training and pruning. *IEEE Transactions on Neural Networks 10*, 1 (1999), 161–166.

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

[61] Thrampoulidis, C., Abbasi, E., and Hassibi, B. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems* (2015), pp. 3420–3428.

[62] Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory 64*, 8 (2018), 5592–5628.

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

[64] Tian, Y., Jiang, T., Gong, Q., and Morcos, A. Luck matters: Understanding training dynamics of deep relu networks. *arXiv preprint arXiv:1905.13405* (2019).

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

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

Organization of the Supplementary Material

Supplementary material is organized as follows.

1. Appendix A derives Auxiliary Distribution (Definition 3.2). We also provide the relevant background and supporting results on Convex Gaussian Min-Max Theorem (CGMT) and discuss how distributional similarity based on Def. 3.2 can be formalized.
2. Appendix B proves Theorem 4.3. In Appendix B.2 (see Proposition B.2), we also provide theoretical results proving the tightness of the bounds provided in Theorem 4.3.
3. Appendix C proves Lemmas 3.1 and Lemma 4.1.
4. Appendix D provides further numerical results on Section 4. Appendix D provides results on layer-wise pruning, where pruning is done on each layer individually, and compares to Section 4 which uses standard pruning.
5. Appendix E provides further technical results supporting Appendix A.

A Auxiliary Distribution for Pruning Linear Models

A.1 Technical Background on Convex Gaussian Min-Max Theorem

CGMT framework is proposed by [63] and allows for accurate analysis of a large class of optimization problems involving random matrices. The key idea is relating the original problem (Primary Optimization PO) to an Auxiliary Optimization (AO) problem. Given compact convex set $\mathcal{S} \in \mathbb{R}^p$, regularization parameter $\lambda > 0$ and continuous convex function $\psi(\cdot) : \mathbb{R}^p \to \mathbb{R}$, define the functions

$$\Phi_\lambda(\boldsymbol{X}) = \min_{\boldsymbol{w}\in\mathcal{S}} \max_{\|\boldsymbol{a}\|_{\ell_2}\leq\lambda} \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{w} + \psi(\boldsymbol{w}) = \min_{\boldsymbol{w}\in\mathcal{S}} \lambda\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2} + \psi(\boldsymbol{w}) \tag{A.1}$$

$$\phi_\lambda(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}} \max_{\|\boldsymbol{a}\|_{\ell_2}\leq\lambda} \|\boldsymbol{w}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{a} - \|\boldsymbol{a}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{w} + \psi(\boldsymbol{w}) \tag{A.2}$$

$$= \min_{\boldsymbol{w}\in\mathcal{S}} \lambda(\|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2} - \boldsymbol{h}^T\boldsymbol{w})_+ + \psi(\boldsymbol{w}) \tag{A.3}$$

Suppose $\boldsymbol{X} \in \mathbb{R}^{n\times p}, \boldsymbol{g} \in \mathbb{R}^n, \boldsymbol{h} \in \mathbb{R}^p \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then, CGMT yields the following inequality for any $\mu \in \mathbb{R}, t > 0$,

$$\mathbb{P}(|\Phi_\lambda(\boldsymbol{X}) - \mu| > t) \leq 2\mathbb{P}(|\phi_\lambda(\boldsymbol{g},\boldsymbol{h}) - \mu| > t). \tag{A.4}$$

In words, the right and left-hand side objectives are probabilistically equal.
Relation to ridge regression: Observe that (A.1) can easily be related to ridge regression which solves

$$\min_{\boldsymbol{\theta}} \mathcal{L}_\lambda(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \lambda\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_{\ell_2} + \|\boldsymbol{\theta}\|_{\ell_2}. \tag{A.5}$$

Recalling $\boldsymbol{y} = \boldsymbol{X}\bar{\boldsymbol{\theta}} + \sigma\boldsymbol{z}$ with $\boldsymbol{z} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and applying the change of variable $\boldsymbol{w} = \bar{\boldsymbol{\theta}} - \boldsymbol{\theta}$, we find

$$\mathcal{L}_\lambda(\boldsymbol{w}) = \lambda\|[\boldsymbol{X}\ \boldsymbol{z}]\begin{bmatrix}\boldsymbol{w}\\ \sigma\end{bmatrix}\|_{\ell_2} + \|\bar{\boldsymbol{\theta}} - \boldsymbol{w}\|_{\ell_2}.$$

Observe that $\boldsymbol{X}' = [\boldsymbol{X}\ \boldsymbol{z}] \in \mathbb{R}^{n\times(p+1)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ thus setting $\psi(\boldsymbol{w}) = \|\bar{\boldsymbol{\theta}} - \boldsymbol{w}\|_{\ell_2}$, minimization over $\mathcal{L}(\boldsymbol{w})$ has the exact same form as (A.1) and CGMT is applicable with

$$\Phi_\lambda(\boldsymbol{X}') = \min_{\boldsymbol{w}} \lambda\|\boldsymbol{X}'\begin{bmatrix}\boldsymbol{w}\\ \sigma\end{bmatrix}\|_{\ell_2} + \|\bar{\boldsymbol{\theta}} - \boldsymbol{w}\|_{\ell_2}$$

Covariance on the design matrix can be handled as well as described in Appendix A.3.

Over-parameterized Least-Squares: In Section 3.1 we study over-parameterized least-squares which interpolates the training labels perfectly rather than using ridge regularization. Specifically, we solve the min Euclidian norm problem

$$\arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_{\ell_2} \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}.$$

Note that this corresponds to solving (A.5) with $\lambda \to \infty$. Using the same change of variable, we end up with the primary optimization

$$\Phi_\infty(\boldsymbol{X}') = \min_{\boldsymbol{w}} \|\bar{\boldsymbol{\theta}} - \boldsymbol{w}\|_{\ell_2} \quad \text{subject to} \quad \boldsymbol{X}' \begin{bmatrix} \boldsymbol{w} \\ \sigma \end{bmatrix} = 0.$$

Unfortunately, CGMT framework for our scenario has two drawbacks due to technical issues. First, it only handles the regularization term and doesn't allow for random matrix constraints. Secondly, as mentioned earlierin (A.1), $\boldsymbol{w}$ has to lie on a compact set $\mathcal{S}$. Even $\boldsymbol{w} \in \mathbb{R}^p$ has to be addressed with care. We first have the following theorem which circumvents these issues. The following result is a corollary of Theorem E.1 and allows for equality constraints on $\boldsymbol{X}$ and replaces compactness on $\mathcal{S}$ with closedness.

Theorem A.1 (CGMT with constraints) *Given a closed $\mathcal{S}$ and a continuous function ψ satisfying $\lim_{\|\boldsymbol{v}\|_{\ell_2}\to\infty} \psi(\boldsymbol{v}) = \infty$, define the PO and AO problems*

$$\Phi_\infty(\boldsymbol{X}) = \min_{\boldsymbol{w}\in\mathcal{S}, \boldsymbol{X}\boldsymbol{w}=0} \psi(\boldsymbol{w}) \tag{A.6}$$

$$\phi_\infty(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}, \|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2} \le \boldsymbol{h}^T\boldsymbol{w}} \psi(\boldsymbol{w}). \tag{A.7}$$

Suppose $\boldsymbol{X} \in \mathbb{R}^{n\times p}, \boldsymbol{g} \in \mathbb{R}^n, \boldsymbol{h} \in \mathbb{R}^p \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Then, for any $t > 0$ and $\mu \in \mathbb{R}$, we have that

- $\mathbb{P}(\Phi_\infty(\boldsymbol{X}) < t) \le 2\mathbb{P}(\phi_\infty(\boldsymbol{g},\boldsymbol{h}) \le t)$.
- *If $\mathcal{S}$ is convex, we additionally have* $\mathbb{P}(\Phi_\infty(\boldsymbol{X}) > t) \le 2\mathbb{P}(\phi_\infty(\boldsymbol{g},\boldsymbol{h}) \ge t)$.

A.2 Using CGMT to Infer the Properties of the Solution

In this section, we provide a discussion of how CGMT can be used to infer the properties of the solution of (A.1) by studying the solution of (A.3). This is already the topic of several interesting papers on random matrix theory and high-dimensional statistics [61, 63, 62]. Below, we formalize the distributional similarity of the solution of the primary problem (A.1) and auxiliary problem (A.3) in terms of subsets of $\mathbb{R}^p$ for which auxiliary solution concentrates on.

Lemma A.2 (AO solution to PO solution) *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}, \boldsymbol{g} \in \mathbb{R}^n, \boldsymbol{h} \in \mathbb{R}^p \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Suppose we have two loss functions $\mathcal{L}_{PO}(\boldsymbol{w};\boldsymbol{X})$ and $\mathcal{L}_{AO}(\boldsymbol{w};\boldsymbol{g},\boldsymbol{h})$ as a function of $\boldsymbol{w}$*[1]*. Given a set $\mathcal{S}$, define the objectives*

$$\Phi_{\mathcal{S}}(\boldsymbol{X}) = \min_{\boldsymbol{w}\in\mathcal{S}} \mathcal{L}_{PO}(\boldsymbol{w};\boldsymbol{X}) \quad \text{and} \quad \phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}} \mathcal{L}_{AO}(\boldsymbol{w};\boldsymbol{g},\boldsymbol{h}). \tag{A.8}$$

Suppose Φ and ϕ satisfies the following conditions for any closed set $\mathcal{S}$ and t

- $\mathbb{P}(\Phi_{\mathcal{S}}(\boldsymbol{X}) < t) \le 2\mathbb{P}(\phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) \le t)$.
- *Furthermore, if $\mathcal{S}$ is convex,* $\mathbb{P}(\Phi_{\mathcal{S}}(\boldsymbol{X}) > t) \le 2\mathbb{P}(\phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) \ge t)$.

Define the set of global minima $\mathcal{M} = \{\boldsymbol{w} \mid \mathcal{L}(\boldsymbol{w};\boldsymbol{X}) = \Phi(\boldsymbol{X})\}$. For any closed set $\mathcal{S}$, we have that

$$\mathbb{P}(\mathcal{M} \in \mathcal{S}^c) \ge 1 - 2\min_t (\mathbb{P}(\phi_{\mathbb{R}^p}(\boldsymbol{g},\boldsymbol{h}) \ge t) + \mathbb{P}(\phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) \le t)). \tag{A.9}$$

[1] $\mathcal{L}(\boldsymbol{w},\boldsymbol{a})$ can account for additional set constraints of type $\boldsymbol{w} \in \mathcal{C}$ by adding the indicator penalty $\max_{\lambda\ge 0} \lambda 1_{\boldsymbol{w}\notin\mathcal{C}}$.

Proof Let $\boldsymbol{w}^* \in \mathcal{M}$. Suppose the events $\Phi_{\mathbb{R}^p}(\boldsymbol{g},\boldsymbol{h}) \le t$ and $\Phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) > t$ hold. These two imply that $\boldsymbol{w}^* \notin \mathcal{S}$ hence $\mathcal{M} \subseteq \mathcal{S}^c$. To proceed, for any choice of t

$$\begin{aligned}
\mathbb{P}(\mathcal{M} \in \mathcal{S}^c) &\ge \mathbb{P}(\{\Phi_{\mathbb{R}^p}(\boldsymbol{g},\boldsymbol{h}) \le t\} \cap \{\Phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) > t\}) && \text{(A.10)}\\
&\ge 1 - \mathbb{P}(\Phi_{\mathbb{R}^p}(\boldsymbol{g},\boldsymbol{h}) > t) - \mathbb{P}(\Phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) \le t) && \text{(A.11)}\\
&\ge 1 - \mathbb{P}(\Phi_{\mathbb{R}^p}(\boldsymbol{g},\boldsymbol{h}) > t) - \lim_{t' \to t^+} \mathbb{P}(\Phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) < t') && \text{(A.12)}\\
&\ge 1 - 2\mathbb{P}(\phi_{\mathbb{R}^p}(\boldsymbol{g},\boldsymbol{h}) \ge t) - 2\lim_{t' \to t^+} \mathbb{P}(\phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) \le t'). && \text{(A.13)}
\end{aligned}$$

Since this holds for all t and cumulative distribution function is continuous, we get the advertised bound (A.9). ∎

Note that assumptions of this lemma on the loss functions (A.8) holds for over-parameterized least-squares based on Theorem A.1. In words, this lemma states that, if we can identify a set $\mathcal{S}$ such that $\mathcal{S}$-constrained auxiliary cost $\phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h})$ is larger than the unconstrained cost $\phi_{\mathbb{R}^p}(\boldsymbol{g},\boldsymbol{h})$, then, the solution of the primary problem provably lies on the complement $\mathcal{S}^c$.

Then, if we wish to prove the global minima $\mathcal{M}$ of the primary problem satisfies some property $\mathcal{P}$, the line of attack is as follows.

- Let $\mathcal{S}$ be the set of vectors not satisfying $\mathcal{P}$.
- Show that $\phi_{\mathcal{S}}(\boldsymbol{g},\boldsymbol{h}) > \phi_{\mathbb{R}^p}(\boldsymbol{g},\boldsymbol{h})$ with high probability.

In our application, we wish to argue that pruned auxiliary distribution $\Pi_s^M(\boldsymbol{\theta}_{\text{aux}})$ achieves the same test error as the pruned primary solution $\Pi_s^M(\hat{\boldsymbol{\theta}})$. Thus, the undesired set $\mathcal{S}$ can be defined as the set of vectors whose test error after pruning does not deviate much from the expected test error of pruned auxiliary solution $\boldsymbol{\theta}_{\text{aux}} = \bar{\boldsymbol{\theta}} - \boldsymbol{w}_{\text{aux}}$ i.e. (assuming $\boldsymbol{\Sigma} = \boldsymbol{I}$, the test error simplifies to Euclidian distance to the ground-truth $\bar{\boldsymbol{\theta}}$)

$$\mathcal{S} = \{\boldsymbol{w} \;\big|\; |\|\Pi_s^M(\bar{\boldsymbol{\theta}} - \boldsymbol{w}) - \bar{\boldsymbol{\theta}}\|_{\ell_2} - \mathbb{E}\,\|\Pi_s^M(\boldsymbol{\theta}_{\text{aux}}) - \bar{\boldsymbol{\theta}}\|_{\ell_2}| \le \varepsilon\},$$

where $\varepsilon > 0$ is a knob which can approach 0 asymptotically. Setting $\gamma = \mathbb{E}\,\|\Pi_s^M(\boldsymbol{\theta}_{\text{aux}}) - \bar{\boldsymbol{\theta}}\|_{\ell_2}$ and $f(\boldsymbol{w}) = \|\Pi_s^M(\bar{\boldsymbol{\theta}} - \boldsymbol{w}) - \bar{\boldsymbol{\theta}}\|_{\ell_2}$, this can be simplified to

$$\mathcal{S} = \{\boldsymbol{w} \;\big|\; |f(\boldsymbol{w}) - \gamma| \le \varepsilon\},$$

Technical Challenge in Pruning Analysis: Here, the technical challenge is analyzing the auxiliary problem over $\mathcal{S}$ which is a highly non-convex set due to the hard-thresholding operator. Even showing the concentration of the auxiliary error $\|\Pi_s^M(\boldsymbol{\theta}_{\text{aux}}) - \bar{\boldsymbol{\theta}}\|_{\ell_2}$ around its expectation γ is not trivial. If $f(\boldsymbol{w})$ is a Lipschitz function of $\boldsymbol{w}$, $\mathcal{S}$ is a more manageable set and it is typically relatively easy to show that its elements are bounded away from zero (in Euclidian norm). Once $\mathcal{S}$ is bounded away from zero, what remains is showing optimization over $\mathcal{S}$ leads to a strictly larger loss since the set doesn't include global minima in it with high probability. We again remark that using soft-thresholding based pruning would be an easier path to theoretical guarantees and fully formalizing the pruning formulas as the soft-thresholding operator $\text{shrink}_T(x) = \max(x - T, 0)$ is Lipschitz.

Finally, the next subsection derives the auxiliary distribution of Definition 3.2 by solving the auxiliary problem associated with the over-parameterized least-squares.

A.3 Deriving the Auxiliary Distribution (Definition 3.2)

A.3.1 Over-parameterized Least-Squares with Diagonal Covariance

Let us first set the exact problem we are analyzing. 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 ground-truth vector $\boldsymbol{\theta}$ and labels $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \sigma\boldsymbol{z}$,

we consider the least-squares problem subject to the minimum Euclidian norm constraint (as $\kappa = p/n > 1$) given by

$$\min_{\boldsymbol{\theta}'} \|\boldsymbol{\theta}'\|_{\ell_2} \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}'. \tag{A.14}$$

Next subsection A.3.2 will adapt the analysis of this subsection to obtain Def. 3.2. Using change of variable $\boldsymbol{\theta}' = \boldsymbol{\theta} - \boldsymbol{w}$, optimization problem (A.14) leads to

$$\Phi(\boldsymbol{X}) = \min_{\boldsymbol{w}} \|\boldsymbol{\theta} - \boldsymbol{w}\|_{\ell_2} \quad \text{subject to} \quad \boldsymbol{X}\boldsymbol{w} + \sigma\boldsymbol{z} = 0. \tag{A.15}$$

Write $\boldsymbol{X} = \bar{\boldsymbol{X}}\sqrt{\boldsymbol{\Sigma}}$ where $\bar{\boldsymbol{X}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Noticing $\|\boldsymbol{X}\boldsymbol{w} + \sigma\boldsymbol{z}\|_{\ell_2} = \|\bar{\boldsymbol{X}}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{w} + \sigma\boldsymbol{z}\|_{\ell_2}$, and recalling the constrained CGMT forms (A.6) and (A.7), the auxiliary problem takes the form

$$\phi(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}} \|\boldsymbol{\theta} - \boldsymbol{w}\|_{\ell_2} \quad \text{subject to} \quad \|\boldsymbol{g}\|_{\ell_2}\|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{w}\ \sigma\|_{\ell_2} \le \boldsymbol{h}^T\sqrt{\boldsymbol{\Sigma}}\boldsymbol{w} + \sigma h. \tag{A.16}$$

where $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_n)$, $\boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{I}_p)$, $h \sim \mathcal{N}(0,1)$. Set $\bar{\boldsymbol{h}} = \boldsymbol{h}/\sqrt{p}$. Letting $p \to \infty$ and setting $\kappa = p/n$ a constant, observe that $h/\|\boldsymbol{g}\|_{\ell_2} \to 0$, $h/\|\boldsymbol{g}\|_{\ell_2} = \sqrt{\kappa}\bar{h}$, and we have pointwise convergence (over $\boldsymbol{w}$) to the problem

$$\phi(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}} \|\boldsymbol{\theta} - \boldsymbol{w}\|_{\ell_2} \quad \text{subject to} \quad \|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{w}\ \sigma\|_{\ell_2} \le \sqrt{\kappa}\bar{\boldsymbol{h}}^T\sqrt{\boldsymbol{\Sigma}}\boldsymbol{w}. \tag{A.17}$$

Taking the squares of both sides, we find the equivalent optimization (which preserves the minima)

$$\phi(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}} \|\boldsymbol{\theta} - \boldsymbol{w}\|_{\ell_2}^2 \quad \text{subject to} \quad \|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{w}\ \sigma\|_{\ell_2}^2 \le \kappa(\bar{\boldsymbol{h}}^T\sqrt{\boldsymbol{\Sigma}}\boldsymbol{w})^2, \tag{A.18}$$

Set $S(\boldsymbol{w}) = \bar{\boldsymbol{h}}^T\sqrt{\boldsymbol{\Sigma}}\boldsymbol{w} = \sum_{i=1}^p \bar{\boldsymbol{h}}_i\boldsymbol{w}_i\sqrt{\boldsymbol{\Sigma}_{i,i}}$. The optimization above can alternatively be written in the entrywise decomposed form

$$\phi(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}} \sum_{i=1}^p (\boldsymbol{\theta}_i - \boldsymbol{w}_i)^2 \quad \text{subject to} \quad \sigma^2 + \sum_{i=1}^p \boldsymbol{\Sigma}_{i,i}\boldsymbol{w}_i^2 \le \kappa S(\boldsymbol{w})^2. \tag{A.19}$$

Considering the Lagrangian form, we find

$$\phi(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}} \max_{\Xi \ge 0} \sum_{i=1}^p (\boldsymbol{\theta}_i - \boldsymbol{w}_i)^2 + \Xi[\sigma^2 + \sum_{i=1}^p \boldsymbol{\Sigma}_{i,i}\boldsymbol{w}_i^2 - \kappa S(\boldsymbol{w})^2]. \tag{A.20}$$

We will decompose entries of $\boldsymbol{w}_i$ as a term dependent on $\bar{\boldsymbol{h}}_i$ and an independent bias term via

$$\boldsymbol{w}_i = \frac{\gamma_i}{\sqrt{\boldsymbol{\Sigma}_{i,i}}}\bar{\boldsymbol{h}}_i + \zeta_i\boldsymbol{\theta}_i. \tag{A.21}$$

Also set the variable

$$\Gamma = (\frac{1}{p}\sum_{i=1}^p \gamma_i)^2.$$

Using Law of Large Numbers, we have

$$\lim_{p\to\infty} S(\boldsymbol{w}) = \mathbb{E}[\bar{\boldsymbol{h}}^T\sqrt{\boldsymbol{\Sigma}}\boldsymbol{w}] = \mathbb{E}[\sum_{i=1}^p \gamma_i\bar{\boldsymbol{h}}_i^2] = \sqrt{\Gamma},$$

and

$$\lim_{p\to\infty} \sum_{i=1}^p (\boldsymbol{\theta}_i - \boldsymbol{w}_i)^2 = \mathbb{E}[\sum_{i=1}^p (\boldsymbol{\theta}_i - \boldsymbol{w}_i)^2] = \sum_{i=1}^p (1-\zeta_i)^2\boldsymbol{\theta}_i^2 + \frac{\gamma_i^2}{p\boldsymbol{\Sigma}_{i,i}},$$

and

$$\lim_{p\to\infty} \sum_{i=1}^p \boldsymbol{\Sigma}_{i,i}\boldsymbol{w}_i^2 = \mathbb{E}[\sum_{i=1}^p \boldsymbol{\Sigma}_{i,i}\boldsymbol{w}_i^2] = \sum_{i=1}^p \boldsymbol{\Sigma}_{i,i}\zeta_i^2\boldsymbol{\theta}_i^2 + \frac{\gamma_i^2}{p}.$$

Thus, we rewrite the problem (A.20) as

$$\phi(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{\zeta},\boldsymbol{\gamma}} \max_{\Xi\geq 0} \sum_{i=1}^{p} (1-\zeta_i)^2\boldsymbol{\theta}_i^2 + \frac{\gamma_i^2}{p\boldsymbol{\Sigma}_{i,i}} + \Xi[\sigma^2 + \sum_{i=1}^{p} \boldsymbol{\Sigma}_{i,i}\zeta_i^2\boldsymbol{\theta}_i^2 + \frac{\gamma_i^2}{p} - \kappa\Gamma]. \tag{A.22}$$

Differentiating with respect to γ_i and ζ_i, and recalling the definition of Γ, we obtain the equations

$$\frac{\gamma_i}{p\boldsymbol{\Sigma}_{i,i}} + \Xi(\frac{\gamma_i}{p} - \frac{\kappa\sqrt{\Gamma}}{p}) = 0 \iff \gamma_i = \frac{\kappa\sqrt{\Gamma}}{1+(\Xi\boldsymbol{\Sigma}_{i,i})^{-1}} \tag{A.23}$$

$$(\zeta_i - 1)\boldsymbol{\theta}_i^2 + \Xi\boldsymbol{\Sigma}_{i,i}\boldsymbol{\theta}_i^2\zeta_i = 0 \iff \zeta_i = \frac{1}{1+\Xi\boldsymbol{\Sigma}_{i,i}}. \tag{A.24}$$

Using the definition of Γ, we find that, $\Xi > 0$ has to satisfy

$$\sqrt{\Gamma} = \frac{1}{p}\sum_{i=1}^{p}\gamma_i = \frac{1}{p}\sum_{i=1}^{p}\frac{\kappa\sqrt{\Gamma}}{1+(\Xi\boldsymbol{\Sigma}_{i,i})^{-1}} \iff \tag{A.25}$$

$$1 = \frac{\kappa}{p}\sum_{i=1}^{p}\frac{1}{1+(\Xi\boldsymbol{\Sigma}_{i,i})^{-1}}. \tag{A.26}$$

Finally, since $\Xi > 0$, we need to satisfy the complementary slackness i.e. the term multiplying Ξ has to be zero. This implies the equality

$$\sigma^2 + \sum_{i=1}^{p}\frac{\gamma_i^2}{p} + \boldsymbol{\Sigma}_{i,i}\zeta_i^2\boldsymbol{\theta}_i^2 = \kappa\Gamma. \tag{A.27}$$

In summary, following (A.21), we found that the solution to auxiliary problem (A.16) has the form

$$\boldsymbol{w}(\boldsymbol{g},\boldsymbol{h}) = \boldsymbol{\zeta}\odot\boldsymbol{\theta} + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\gamma}\odot\bar{\boldsymbol{h}},$$

where $\boldsymbol{\gamma},\boldsymbol{\zeta}\in\mathbb{R}^p$ are given by solving the following equations.

- Ξ satisfies (A.26). Note that there is a unique positive Ξ solving this equation because when $\Xi = 0$ right side is p/n which is larger than one and the right side is strictly decreasing function of Ξ thus mean-value theorem implies unique solution,
- ζ_i satisfies (A.24),
- γ_i satisfies (A.23),
- Finally Γ satisfies (A.27) which leads to (after substituting γ_i definition)

  $$\sigma^2 + \sum_{i=1}^{p}\frac{\kappa^2\Gamma}{p(1+(\Xi\boldsymbol{\Sigma}_{i,i})^{-1})^2} + \boldsymbol{\Sigma}_{i,i}\zeta_i^2\boldsymbol{\theta}_i^2 = \kappa\Gamma \iff \sigma^2 + \sum_{i=1}^{p}\boldsymbol{\Sigma}_{i,i}\zeta_i^2\boldsymbol{\theta}_i^2 = \kappa\Gamma(1-\frac{\kappa}{p}\sum_{i=1}^{p}(1+(\Xi\boldsymbol{\Sigma}_{i,i})^{-1})^{-2}),$$

  which yields

  $$\Gamma = \frac{\sigma^2 + \sum_{i=1}^{p}\boldsymbol{\Sigma}_{i,i}\zeta_i^2\boldsymbol{\theta}_i^2}{\kappa(1-\frac{\kappa}{p}\sum_{i=1}^{p}(1+(\Xi\boldsymbol{\Sigma}_{i,i})^{-1})^{-2})}. \tag{A.28}$$

  Finally, the parameter distribution of the axuiliary problem is given by reversing the change of variable i.e.

  $$\boldsymbol{\theta}_{\text{aux}} = \boldsymbol{\theta} - \boldsymbol{w}(\boldsymbol{g},\boldsymbol{h}) = (\mathbb{1}_p - \boldsymbol{\zeta})\odot\boldsymbol{\theta} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\gamma}\odot\bar{\boldsymbol{h}}, \tag{A.29}$$

  where $\bar{\boldsymbol{h}}\sim\mathcal{N}(0,\boldsymbol{I}_p/p)$.

A.3.2 Obtaining the Auxiliary Distribution of Definition 3.2

The setup in Section 3 can be mapped to the previous section as follows.

- Feature covariance is $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^2$ for some diagonal matrix $\boldsymbol{\Lambda}$,
- The ground-truth vector is $\boldsymbol{\theta} = \boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{\theta}}$ (as $\boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{\theta}}$ is the population minima of $\mathcal{L}_{\boldsymbol{\Lambda}}$).

Plugging these into (A.24), (A.23), (A.26), (A.28) and finally the equation of the auxiliary solution (A.29) leads to Definition 3.2. Specifically, the terms are stated in terms of $\boldsymbol{\Lambda}$ rather than $\boldsymbol{\Sigma}$ and we also remark that $\Gamma, \boldsymbol{\theta}_{\text{aux}}$ terms slightly differ due to the ground-truth vector mapping $\boldsymbol{\theta} \leftrightarrow \boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{\theta}}$.

B Larger Hessian Wins More

This section proves Theorem 4.3 and explains the tightness of its bounds. The following lemma is a standard result under smoothness (Lipschitz gradient) condition.

Lemma B.1 *Suppose $\mathcal{L}$ has L-Lipschitz gradients and $\min_{\boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}') \geq 0$. Then, we have that*

$$\|\nabla \mathcal{L}(\boldsymbol{\theta})\|_{\ell_2} \leq \sqrt{2L\mathcal{L}(\boldsymbol{\theta})}.$$

Proof L-smoothness of the function implies

$$\mathcal{L}(\boldsymbol{a}) \leq \mathcal{L}(\boldsymbol{b}) + \langle \nabla \mathcal{L}(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + \frac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_{\ell_2}^2.$$

Setting $\boldsymbol{a} = \boldsymbol{b} - \nabla \mathcal{L}(\boldsymbol{b})/L$, we find the desired result via

$$\frac{\|\nabla \mathcal{L}(\boldsymbol{b})\|_{\ell_2}^2}{2L} \leq \mathcal{L}(\boldsymbol{b}) - \mathcal{L}(\boldsymbol{a}) \leq \mathcal{L}(\boldsymbol{b}) - \min_{\boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}') \leq \mathcal{L}(\boldsymbol{b}).$$

■

B.1 Proof of Theorem 4.3

Proof Step 1: Proving (4.5): Our proof will be accomplished by carefully keeping track of the gradient descent dynamics for both parameters. Observe that if PPLS holds, then the full gradient satisfies PL condition with parameter $\mu = \sum_{i=1}^{D} \mu_i$ since

$$\|\nabla \mathcal{L}(\boldsymbol{\theta})\|_{\ell_2}^2 \sum_{i=1}^{D} \|\frac{\partial}{\partial \boldsymbol{\theta}_{\Delta_i}} \mathcal{L}(\boldsymbol{\theta})\|_{\ell_2}^2 \geq 2 \sum_{i=1}^{D} \mu_i \mathcal{L}(\boldsymbol{\theta}) = 2\mu \mathcal{L}(\boldsymbol{\theta}).$$

With this observation, the statement

$$\mathcal{L}(\boldsymbol{\theta}_\tau) \leq (1 - \eta\mu)^\tau \mathcal{L}(\boldsymbol{\theta}_0) \tag{B.1}$$

on linear convergence is standard knowledge on PL inequality. Denote the ith partial derivative via $\nabla_i \mathcal{L}(\boldsymbol{\theta}_\tau)$. Using properties of Hessian and L_i-Lipschitzness of partial gradient with respect to $\boldsymbol{\theta}_{\Delta_i}$, note that overall function is $L = \sum_{i=1}^{D} L_i$-smooth using positive-semidefiniteness of Hessian and upper bounds on its block diagonals. Secondly using PL condition and Lemma B.1, we have that

$$\frac{\|\nabla_i \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}}{\|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}} \leq \frac{\sqrt{L_i \mathcal{L}(\boldsymbol{\theta}_\tau)}}{\sqrt{\mu \mathcal{L}(\boldsymbol{\theta}_\tau)}} \leq \sqrt{L_i/\mu}.$$

Thus, we can write

$$\|\boldsymbol{\theta}_{\Delta_i,\tau+1} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2} \le \|\boldsymbol{\theta}_{\Delta_i,\tau} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2} + \eta\|\nabla_i \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2} \tag{B.2}$$

$$\le \|\boldsymbol{\theta}_{\Delta_i,\tau} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2} + \eta\sqrt{L_i/\mu}\|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}. \tag{B.3}$$

For any $\eta \le 1/L$, L-smoothness and PL condition guarantees

$$\mathcal{L}(\boldsymbol{\theta}_{\tau+1}) \le \mathcal{L}(\boldsymbol{\theta}_\tau) - \frac{\eta}{2}\|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2 \implies \tag{B.4}$$

$$\sqrt{\mathcal{L}(\boldsymbol{\theta}_{\tau+1})} \le \sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} - \frac{\eta}{4\sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)}}\|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2 \tag{B.5}$$

$$\sqrt{\mathcal{L}(\boldsymbol{\theta}_{\tau+1})} \le \sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} - \eta\sqrt{\mu/8}\|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}. \tag{B.6}$$

Define the Lyapunov function

$$\mathcal{V}_\tau = \sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} + \max_{1\le i\le D} C_i\|\boldsymbol{\theta}_{\Delta_i,\tau} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2}.$$

We will find proper C_i's such that $\mathcal{V}_\tau$ is non-increasing. Observe that

$$\mathcal{V}_{\tau+1} - \mathcal{V}_\tau \le C_i\eta\sqrt{L_i/\mu}\|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2} - \eta\sqrt{\mu/8}\|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2} \le 0,$$

when $C_i = \mu/\sqrt{8L_i}$. Thus we pick

$$\mathcal{V}_\tau = \sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} + \max_{1\le i\le D} \frac{\mu}{\sqrt{8L_i}}\|\boldsymbol{\theta}_{\Delta_i,\tau} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2}.$$

Since Lyapunov function is non-increasing, for all $\tau \ge 0$, we are guaranteed to have

$$\|\boldsymbol{\theta}_{\Delta_i,\tau} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2}^2 \le \frac{8L_i}{\mu^2}\mathcal{L}(\boldsymbol{\theta}_0).$$

What remains is upper bounding the contribution of $\boldsymbol{\theta}_i$ to the objective function which is addressed next.

Step 2: Proving (4.6): Using the bound on $\mathcal{L}(\boldsymbol{\theta}_\tau)$ and L_i-smoothness of the partial derivative with respect to $\boldsymbol{\theta}_{\Delta_i}$ and Lemma B.1, we find

$$\|\frac{\partial}{\partial \boldsymbol{\theta}_{\Delta_i}}\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2} \le \sqrt{2L_i(1-\eta\mu)^\tau \mathcal{L}(\boldsymbol{\theta}_0)}. \tag{B.7}$$

At iteration τ, define $\boldsymbol{\theta}(t) = t\boldsymbol{\theta}_\tau + (1-t)\bar{\boldsymbol{\theta}}_\tau$ for $0 \le t \le 1$. Observe that, via line integration, we can bound

$$|\mathcal{L}(\boldsymbol{\theta}_\tau) - \mathcal{L}(\bar{\boldsymbol{\theta}}_\tau)| \le \sup_{0\le t\le 1}\|\frac{\partial}{\partial \boldsymbol{\theta}_{\Delta_i}}\mathcal{L}(\boldsymbol{\theta}(t))\|_{\ell_2}\|\boldsymbol{\theta}_{\Delta_i,\tau} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2}. \tag{B.8}$$

For the right-hand side, we use the earlier upper bound

$$\|\boldsymbol{\theta}_{\Delta_i,\tau} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2} \le R.$$

Next, using (B.7) and L_i-smoothness again, we also bound the gradient norm via

$$\|\frac{\partial}{\partial \boldsymbol{\theta}_{\Delta_i}}\mathcal{L}(\boldsymbol{\theta}(t))\|_{\ell_2} \le \|\frac{\partial}{\partial \boldsymbol{\theta}_{\Delta_i}}\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2} + L_i\|\boldsymbol{\theta}_\tau - \bar{\boldsymbol{\theta}}_\tau\|_{\ell_2} \tag{B.9}$$

$$\le RL_i + \sqrt{2L_i(1-\eta\mu)^\tau \mathcal{L}(\boldsymbol{\theta}_0)}. \tag{B.10}$$

Recalling (B.8) and substituting R, we find

$$|\mathcal{L}(\boldsymbol{\theta}_\tau) - \mathcal{L}(\bar{\boldsymbol{\theta}}_\tau)| \le R^2 L_i + R\sqrt{2L_i(1-\eta\mu)^\tau \mathcal{L}(\boldsymbol{\theta}_0)} \tag{B.11}$$

$$\le \mathcal{L}(\boldsymbol{\theta}_0)(8L_i^2/\mu^2 + 4(L_i/\mu)(1-\eta\mu)^{\tau/2}) \tag{B.12}$$

$$\le \mathcal{L}(\boldsymbol{\theta}_0)(8\kappa^2 + 4\kappa(1-\eta\mu)^{\tau/2}). \tag{B.13}$$

This yields our bound (4.6). ■

Step 3: Proving (4.7): Throughout the remaining proof, let $\tilde{\boldsymbol{\theta}}_\tau = [\boldsymbol{\theta}_{\bar{\Delta}_i,0}\ \boldsymbol{\theta}_{\Delta_i,\tau}]$ be the $\boldsymbol{\theta}_{\bar{\Delta}_i}$-ablated vector which sets the entries $\boldsymbol{\theta}_{\bar{\Delta}_i,\tau}$ of the τ'th iterate to their initial state $\boldsymbol{\theta}_{\bar{\Delta}_i,0}$. Similarly, let $\bar{\boldsymbol{\theta}}_\tau = [\boldsymbol{\theta}_{\bar{\Delta}_i,\tau}\ \boldsymbol{\theta}_{\Delta_i,0}]$ be the $\boldsymbol{\theta}_{\Delta_i}$ ablated vector. By construction

$$\mathcal{I}^N_{\bar{\Delta}_i}(\boldsymbol{\theta}_\tau, \boldsymbol{\theta}_0) = \mathcal{L}(\tilde{\boldsymbol{\theta}}_\tau) - \mathcal{L}(\boldsymbol{\theta}_\tau), \quad \mathcal{I}^N_{\Delta_i}(\boldsymbol{\theta}_\tau, \boldsymbol{\theta}_0) = \mathcal{L}(\bar{\boldsymbol{\theta}}_\tau) - \mathcal{L}(\boldsymbol{\theta}_\tau).$$

Set the distance parameter $R = \sqrt{8L_i\mathcal{L}(\boldsymbol{\theta}_0)}/\mu \ge \|\boldsymbol{\theta}_{\Delta_i,\tau}\|_{\ell_2}$ as a short hand notation.

Applying Lemma B.1 on $\boldsymbol{\theta}_{\Delta_i}$, for any $\boldsymbol{\theta}(t) = t\boldsymbol{\theta}_0 + (1-t)\tilde{\boldsymbol{\theta}}_\tau$, we have that

$$\|\frac{\partial}{\partial \boldsymbol{\theta}_{\Delta_i}}\mathcal{L}(\boldsymbol{\theta}(t))\|_{\ell_2} \le \sqrt{2L_i\mathcal{L}(\boldsymbol{\theta}_0)} + RL_i.$$

Consequently, using line integration bound and $\|\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}_\tau\|_{\ell_2} = \|\boldsymbol{\theta}_{\Delta_i,\tau}\|_{\ell_2} \le R$, we get

$$\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}(\tilde{\boldsymbol{\theta}}_\tau) \le |\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}(\tilde{\boldsymbol{\theta}}_\tau)| \le R(\sqrt{2L_i\mathcal{L}(\boldsymbol{\theta}_0)} + RL_i) \tag{B.14}$$

$$\le \mathcal{L}(\boldsymbol{\theta}_0)(4L_i/\mu + 8L_i^2/\mu^2) \tag{B.15}$$

$$\le \mathcal{L}(\boldsymbol{\theta}_0)(8\kappa^2 + 4\kappa). \tag{B.16}$$

Combining this with (B.1), we obtain the second bound (4.7) via

$$\frac{\mathcal{L}(\tilde{\boldsymbol{\theta}}_\tau) - \mathcal{L}(\boldsymbol{\theta}_\tau)}{\mathcal{L}(\boldsymbol{\theta}_0)} \ge 1 - 8\kappa^2 - 4\kappa - (1-\eta\mu)^\tau.$$

B.2 Theorem 4.3 is Tight

To demonstrate the tightness of Theorem 4.3, we consider an over-parameterized linear regression setup similar to (4.4). Consider D feature sets $(\boldsymbol{X}_i)_{i=1}^D \in \mathbb{R}^{n\times p_i}$ with $p_i \ge n$ where we fit

$$\mathcal{L}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}=(\boldsymbol{\theta}_i)_{i=1}^D} 0.5\|\boldsymbol{y} - \sum_{i=1}^D \boldsymbol{X}_i\boldsymbol{\theta}_i\|_{\ell_2}^2. \tag{B.17}$$

Let Δ_i be the set of entries corresponding to $\boldsymbol{\theta}_i$. PPLS holds over Δ_i with parameters $\mu_i = \sigma_{\min}(\boldsymbol{X}_i)^2$ and $L_i = \|\boldsymbol{X}_i\|^2$. The overall problem is a regression with the design matrix $\boldsymbol{X} = [\boldsymbol{X}_1\ \ldots\ \boldsymbol{X}_D] \in \mathbb{R}^{n\times p}$ where $p = \sum_{i=1}^D p_i$ and $\boldsymbol{X}$ satisfies the PL and smoothness bounds with $\mu = \sum_{i=1}^D \mu_i$ and $L = \sum_{i=1}^D L_i$. To proceed, we have the following proposition that proves the tightness of Theorem 4.3 up to condition numbers L_i/μ_i and L/μ. Specifically, this proposition provides bounds sharply complementing Theorem 4.3 by using the properties of the minimum ℓ_2 norm solution to (B.17) which is the solution gradient descent converges to starting from zero initialization.

Proposition B.2 *Let $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}_i^\star)_{i=1}^D$ be the solution found by gradient descent on the loss (B.17) starting from an initialization $\boldsymbol{\theta}_0$ (with learning rate $\eta \le 1/L$). Set $\tilde{\kappa} = \mu_i/L$. Then, $\boldsymbol{\theta}_i^\star$ satisfies the following bounds*

$$\|\boldsymbol{\theta}_{\Delta_i,\tau} - \boldsymbol{\theta}_{\Delta_i,0}\|_{\ell_2}^2 \ge 2\tilde{\kappa}\mathcal{L}(\boldsymbol{\theta}_0)/L, \tag{B.18}$$

$$\mathcal{I}^N_{\Delta_i}(\boldsymbol{\theta}_\tau, \boldsymbol{\theta}_0)/\mathcal{L}(\boldsymbol{\theta}_0) \ge \tilde{\kappa}^2, \tag{B.19}$$

$$\mathcal{I}^N_{\bar{\Delta}_i}(\boldsymbol{\theta}_\tau, \boldsymbol{\theta}_0)/\mathcal{L}(\boldsymbol{\theta}_0) \le 1 - \tilde{\kappa}^2 - 2\tilde{\kappa} \quad \textit{when} \quad n = 1. \tag{B.20}$$

In short, the bounds of this proposition perfectly complements the bounds of Theorem 4.3 after accounting for the local condition number L_i/μ_i and global condition number L/μ associated with PL condition and smoothness. Specifically, we simply replace $\kappa = L_i/\mu$ with $\tilde{\kappa} = \mu_i/L$ and the converse bounds hold on $\tilde{\kappa}$ up to very small constants. We remark that (B.18) and (B.19) holds generally whereas we show (B.20) for the special case of $n = 1$. Note that $\frac{\kappa}{\tilde{\kappa}} = \frac{L_i}{\mu_i}\frac{L}{\mu}$ which is the multiplication of the local and global condition numbers. Thus, Theorem 4.3 is tight up to these condition numbers and very small constants as claimed in the main body.

Proof Let $\boldsymbol{\theta}^\dagger$ be the pseudo-inverse solution given by

$$\boldsymbol{\theta}^\dagger = \boldsymbol{X}^\dagger \boldsymbol{y} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\tilde{\boldsymbol{y}}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{y}_0$ and $\boldsymbol{y}_0 = \boldsymbol{X}\boldsymbol{\theta}_0$. Gradient descent solution on linear least-squares converges to minimum Euclidian distance solution given by $\boldsymbol{\theta}^\star = \boldsymbol{\theta}_0 + \boldsymbol{\theta}^\dagger$. Observe that $\mathcal{L}(\boldsymbol{\theta}_0) = \|\tilde{\boldsymbol{y}}\|_{\ell_2}^2/2$ and

$$\|\boldsymbol{\theta}^\dagger_{\Delta_{\boldsymbol{W}}}\|_{\ell_2}^2 = \|\boldsymbol{X}_i^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\tilde{\boldsymbol{y}}\|_{\ell_2}^2 \geq \frac{\sigma_{\min}(\boldsymbol{X}_i)^2}{\|\boldsymbol{X}\|^4}\|\tilde{\boldsymbol{y}}^2\|_{\ell_2} \geq \tilde{\kappa}(2\mathcal{L}(\boldsymbol{\theta}_0))/L.$$

This proves the first statement of (B.18). To show the second statement, note that at $\boldsymbol{\theta}^\star$, the loss is equal to zero thus, the Δ_i pruned vector $\boldsymbol{\theta}^p = \boldsymbol{\theta}^\star_{\bar{\Delta}_i} + \boldsymbol{\theta}_{0,\Delta_i}$ achieves a loss of

$$\mathcal{L}(\boldsymbol{\theta}^p) = 0.5\|\boldsymbol{X}_i\boldsymbol{\theta}^\dagger_{\Delta_i}\|_{\ell_2}^2 \geq 0.5\|\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\tilde{\boldsymbol{y}}\|_{\ell_2}^2 \geq \tilde{\kappa}^2\mathcal{L}(\boldsymbol{\theta}_0),$$

yielding (B.19). Finally, we look at the $\bar{\Delta}_i$ pruned vector $\boldsymbol{\theta}^p = \boldsymbol{\theta}^\star_{\Delta_i} + \boldsymbol{\theta}_{0,\bar{\Delta}_i}$. In this case, we wish to show that loss function $\mathcal{L}(\boldsymbol{\theta}^p)$ is upper bounded. We have that

$$\begin{aligned}
2\mathcal{L}(\boldsymbol{\theta}^p) &= \|\tilde{\boldsymbol{y}} - \boldsymbol{X}_i\boldsymbol{\theta}^\dagger_{\Delta_i}\|_{\ell_2}^2 && \text{(B.21)}\\
&= \|\tilde{\boldsymbol{y}} - \boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\tilde{\boldsymbol{y}}\|_{\ell_2}^2 && \text{(B.22)}\\
&= \|(\boldsymbol{I}_n - \boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1})\tilde{\boldsymbol{y}}\|_{\ell_2}^2. && \text{(B.23)}
\end{aligned}$$

To proceed, note that, when $n = 1$, $\boldsymbol{I}_n \geq \boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1} \geq (\mu_i/L)\boldsymbol{I}_n = \tilde{\kappa}\boldsymbol{I}_n$. Consequently,

$$2\mathcal{L}(\boldsymbol{\theta}^p) \leq (1-\tilde{\kappa})^2\|\tilde{\boldsymbol{y}}\|_{\ell_2}^2 = 2(1-\tilde{\kappa})^2\mathcal{L}(\boldsymbol{\theta}_0),$$

concluding the proof of (B.20). ■

C Proofs of Lemmas 3.1 and 4.1

C.1 Proof of Lemma 3.1

Proof The least-squares loss evaluated at a point $\boldsymbol{\theta}$ with design covariance $\boldsymbol{\Sigma}$ is given by

$$\mathbb{E}[(y - \boldsymbol{x}^T\boldsymbol{\theta})^2] = \mathbb{E}[y^2] - 2\boldsymbol{b}^T\boldsymbol{\theta} + \boldsymbol{\theta}^T\boldsymbol{\Sigma}\boldsymbol{\theta}.$$

We first show that HI and NI is invariant to the scaling $\boldsymbol{\Lambda}$ regardless of the covariance $\boldsymbol{\Sigma}$. Observe that the covariance of $\boldsymbol{x}^{\boldsymbol{\Lambda}}$ is $\boldsymbol{\Sigma}^{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}$ and $\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{\theta}}$. Consequently, we find that

$$\mathcal{I}^H_\Delta(\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}) = \sum_{i\in\Delta}\boldsymbol{\Sigma}^{\boldsymbol{\Lambda}}_{i,i}(\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}_i)^2 = \sum_{i\in\Delta}\boldsymbol{\Lambda}^2_{i,i}\boldsymbol{\Sigma}_{i,i}(\boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{\theta}})^2_i = \mathcal{I}^H_\Delta(\bar{\boldsymbol{\theta}}).$$

For NI, observing $\boldsymbol{b}^{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{b}$ and accounting for the $\boldsymbol{\Lambda}$ cancellations, we similarly have

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\Lambda}}(\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}_{\bar{\Delta}}) - \mathcal{L}_{\boldsymbol{\Lambda}}(\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}) &= [-2\boldsymbol{b}^{\boldsymbol{\Lambda}^T}\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}_{\bar{\Delta}} + (\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}_{\bar{\Delta}})^T\boldsymbol{\Sigma}^{\boldsymbol{\Lambda}}\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}_{\bar{\Delta}}] - [-2\boldsymbol{b}^{\boldsymbol{\Lambda}^T}\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}} + (\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}})^T\boldsymbol{\Sigma}^{\boldsymbol{\Lambda}}\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}] && \text{(C.1)}\\
&= [-2\boldsymbol{b}^T\bar{\boldsymbol{\theta}}_{\bar{\Delta}} + (\bar{\boldsymbol{\theta}}_{\bar{\Delta}})^T\boldsymbol{\Sigma}\bar{\boldsymbol{\theta}}_{\bar{\Delta}}] - [-2\boldsymbol{b}^T\bar{\boldsymbol{\theta}} + (\bar{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}\bar{\boldsymbol{\theta}}] && \text{(C.2)}\\
&= 2\boldsymbol{b}^T\bar{\boldsymbol{\theta}}_\Delta + \bar{\boldsymbol{\theta}}^T_{\bar{\Delta}}\boldsymbol{\Sigma}\bar{\boldsymbol{\theta}}_{\bar{\Delta}} - \bar{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}\bar{\boldsymbol{\theta}}, && \text{(C.3)}
\end{aligned}$$

which is independent of $\boldsymbol{\Lambda}$. To proceed, we focus on diagonal covariance matrix $\boldsymbol{\Sigma}$. For HI/NI, we only need to show the result for $\boldsymbol{\Lambda} = \boldsymbol{I}_p$ and establish $\mathcal{I}^N_\Delta(\bar{\boldsymbol{\theta}}) = \mathcal{I}^H_\Delta(\bar{\boldsymbol{\theta}})$. We can then apply the $\boldsymbol{\Lambda}$ invariance result above. The least-squares loss for diagonal covariance evaluated at a point $\boldsymbol{\theta}$ can be written as

$$\mathbb{E}[(y - \boldsymbol{x}^T\boldsymbol{\theta})^2] = \mathbb{E}[y^2] - 2\sum_{i=1}^{p} \boldsymbol{b}_i\boldsymbol{\theta}_i + \boldsymbol{\Sigma}_{i,i}\boldsymbol{\theta}_i^2.$$

Note that $\bar{\boldsymbol{\theta}}_i = \boldsymbol{b}_i/\boldsymbol{\Sigma}_{i,i}$. Thus, recalling the definition of $\bar{\boldsymbol{\theta}}_{\bar{\Delta}}$, we establish the desired HI equal to NI bound as follows

$$\mathcal{I}^N_\Delta(\bar{\boldsymbol{\theta}}) = \mathcal{L}(\bar{\boldsymbol{\theta}}_{\bar{\Delta}}) - \mathcal{L}(\bar{\boldsymbol{\theta}}) = 2\sum_{i\in\Delta} \boldsymbol{b}_i\bar{\boldsymbol{\theta}}_i - \boldsymbol{\Sigma}_{i,i}\bar{\boldsymbol{\theta}}_i^2 = \sum_{i\in\Delta} \frac{\boldsymbol{b}_i^2}{\boldsymbol{\Sigma}_{i,i}} = \sum_{i\in\Delta} \boldsymbol{\Sigma}_{i,i}\bar{\boldsymbol{\theta}}_i^2 = \mathcal{I}^H_\Delta(\bar{\boldsymbol{\theta}}).$$

Finally, magnitude-based importance with diagonal covariance is simply given by $\mathcal{I}^M_\Delta(\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}) = \sum_{i\in\Delta}(\bar{\boldsymbol{\theta}}^{\boldsymbol{\Lambda}}_i)^2 = \sum_{i\in\Delta}(\boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{\theta}})_i^2 = \sum_{i\in\Delta} \boldsymbol{\Lambda}_{i,i}^{-2}\bar{\boldsymbol{\theta}}_i^2$. ■

C.2 Proof of Lemma 4.1

Proof The first statement on MI immediately follows from the definition of MI and the construction of $\boldsymbol{\theta}^\lambda$. For the remaining statements, we analyze the gradient and Hessian as a function of λ. Since Hessian and gradient are linear, we can focus on a single example $(\boldsymbol{x}, y)$. To prevent notational confusion, let us denote the point of evaluation by $(\boldsymbol{W}_0, \boldsymbol{V}_0)$ and the input/output layer variables by $(\boldsymbol{W}, \boldsymbol{V})$. Thus, suppose $\boldsymbol{\theta}^1 = (\boldsymbol{W}_0, \boldsymbol{V}_0)$ and $\boldsymbol{\theta}^\lambda = (\lambda\boldsymbol{W}_0, \lambda^{-1}\boldsymbol{V}_0)$. Use shorthand $f = f_{\boldsymbol{\theta}^\lambda}(\boldsymbol{x})$ which is invariant to λ. Let $L_1 = \nabla_f \ell(y, f) \in \mathbb{R}^K$ and $L_2 = \nabla^2_f \ell(y, f) \in \mathbb{R}^{K\times K}$. Let input layer have $p_I = m \times d$ parameters and output layer has $p_O = K \times m$ parameters. Also denote the partial first and second order derivatives of input layer w.r.t. prediction f via $F^{\boldsymbol{W}}_{1,\lambda} = \nabla_{\boldsymbol{W}} f_{\boldsymbol{\theta}^\lambda}(\boldsymbol{x}) \in \mathbb{R}^{p_I\times K}$ and $F^{\boldsymbol{W}}_{2,\lambda} = \nabla^2_{\boldsymbol{W}} f_{\boldsymbol{\theta}^\lambda}(\boldsymbol{x}) \in \mathbb{R}^{p_I\times p_I\times K}$. Similarly denote the partial derivatives of the output layer by $F^{\boldsymbol{V}}_{1,\lambda} = \nabla_{\boldsymbol{V}} f_{\boldsymbol{\theta}^\lambda}(\boldsymbol{x}) \in \mathbb{R}^{p_O\times K}$ and $F^{\boldsymbol{V}}_{2,\lambda} = \nabla^2_{\boldsymbol{V}} f_{\boldsymbol{\theta}^\lambda}(\boldsymbol{x}) \in \mathbb{R}^{p_O\times p_O\times K}$. First, focusing on gradient (of the vectorized input/output layers), we have the size p_I, p_O partial gradients

$$\nabla_{\boldsymbol{W}} \ell(y, f_{\boldsymbol{\theta}^\lambda}(\boldsymbol{x})) = F^{\boldsymbol{W}}_{1,\lambda} L_1 \tag{C.4}$$

$$\nabla_{\boldsymbol{V}} \ell(y, f_{\boldsymbol{\theta}^\lambda}(\boldsymbol{x})) = F^{\boldsymbol{V}}_{1,\lambda} L_1. \tag{C.5}$$

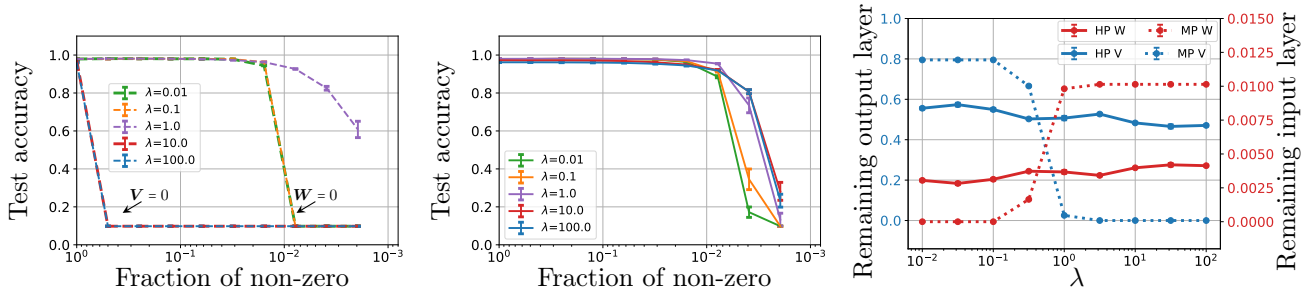
Let $\mu(\cdot)$ be the step function which will correspond to the activation pattern. To proceed, observe that $\mathrm{ReLU}(\lambda\boldsymbol{W}_0\boldsymbol{x}) = \lambda\mathrm{ReLU}(\boldsymbol{W}_0\boldsymbol{x})$ and $\mu(\lambda\boldsymbol{W}_0\boldsymbol{x}) = \mu(\boldsymbol{W}_0\boldsymbol{x})$.

$$\begin{aligned} F^{\boldsymbol{W}}_{1,\lambda} = \nabla_{\boldsymbol{W}=\lambda\boldsymbol{W}_0}(\lambda^{-1}\boldsymbol{V}_0\mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x})) &= \nabla_{\boldsymbol{W}=\boldsymbol{W}_0}(\lambda^{-1}\boldsymbol{V}_0\mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x})) && \text{(C.6)}\\ &= \lambda^{-1}\nabla_{\boldsymbol{W}=\boldsymbol{W}_0}(\boldsymbol{V}_0\mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x})) && \text{(C.7)}\\ &= \lambda^{-1}F^{\boldsymbol{W}}_{1,1} && \text{(C.8)}\\ F^{\boldsymbol{V}}_{1,\lambda} = \nabla_{\boldsymbol{V}=\lambda^{-1}\boldsymbol{V}_0}(\boldsymbol{V}\mathrm{ReLU}(\lambda\boldsymbol{W}_0\boldsymbol{x})) &= \nabla_{\boldsymbol{V}=\boldsymbol{V}_0}(\boldsymbol{V}\mathrm{ReLU}(\lambda\boldsymbol{W}_0\boldsymbol{x})) && \text{(C.9)}\\ &= \lambda\nabla_{\boldsymbol{V}=\boldsymbol{V}_0}(\boldsymbol{V}\mathrm{ReLU}(\boldsymbol{W}_0\boldsymbol{x})) && \text{(C.10)}\\ &= \lambda F^{\boldsymbol{V}}_{1,1}. && \text{(C.11)} \end{aligned}$$

which are the advertised results on gradient.

We next proceed with the Hessian analysis and show similar behavior to gradient. Let us use $\bigotimes$ to denote the tensor-vector multiplication which multiplies an $a \times b \times c$ tensor with a size c vector along the third mode to find an $a \times b$ matrix. Note that

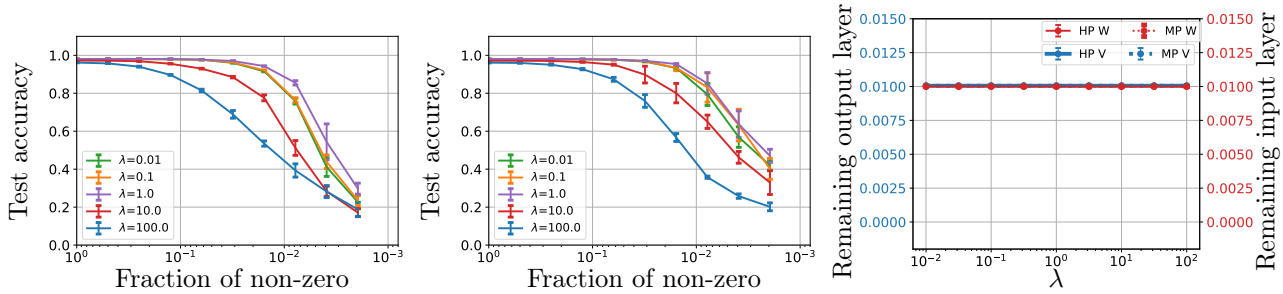
$$\begin{aligned} \nabla^2_{\boldsymbol{W}} \ell(y, f_{\boldsymbol{\theta}^\lambda}(\boldsymbol{x})) &= F^{\boldsymbol{W}}_{2,\lambda} \bigotimes L_1 + F^{\boldsymbol{W}}_{1,\lambda} L_2 F^{\boldsymbol{W}\,T}_{1,\lambda} && \text{(C.12)}\\ &= F^{\boldsymbol{W}}_{2,\lambda} \bigotimes L_1 + \lambda^{-2} F^{\boldsymbol{W}}_{1,1} L_2 F^{\boldsymbol{W}\,T}_{1,1} && \text{(C.13)}\\ \nabla^2_{\boldsymbol{V}} \ell(y, f_{\boldsymbol{\theta}^\lambda}(\boldsymbol{x})) &= F^{\boldsymbol{V}}_{2,\lambda} \bigotimes L_1 + F^{\boldsymbol{V}}_{1,\lambda} L_2 F^{\boldsymbol{V}\,T}_{1,\lambda} && \text{(C.14)}\\ &= F^{\boldsymbol{V}}_{2,\lambda} \bigotimes L_1 + \lambda^2 F^{\boldsymbol{V}}_{1,1} L_2 F^{\boldsymbol{V}\,T}_{1,1}. && \text{(C.15)} \end{aligned}$$



(a) Test accuracy using MP and standard pruning.

(b) Test accuracy using HP and standard pruning.

(c) Remaining nonzeros for $\boldsymbol{V}, \boldsymbol{W}$ with standard pruning



(d) Test accuracy using MP and layer-wise pruning.

(e) Test accuracy using HP and layer-wise pruning.

(f) Remaining nonzeros for $\boldsymbol{V}, \boldsymbol{W}$ with layer-wise pruning

Figure 4: This figure compares layer-wise pruning with standard pruning. Figures (a), (b) and (c) are the same figures in Fig. 3. We use the same setup on (d), (e) and (f) except we use layer-wise pruning instead of standard pruning. Compared to standard MP, layer-wise MP dose not suffer from a full layer dying but the performance is worse when $\lambda = 1$. Moreover, standard HP outperforms layer-wise HP except in the extremely sparse regime (nonzero $\leq 0.2\%$). In this regime, both approaches result in lackluster accuracy.

Thus, to conclude with the proof of (4.3), we will show that $F^{\boldsymbol{V}}_{2,\lambda} = 0$ and $F^{\boldsymbol{W}}_{2,\lambda} = \lambda^{-2} F^{\boldsymbol{W}}_{2,1}$. For the input layer, we use the fact that second derivative of ReLU is the Dirac δ function which satisfies $\delta(x/C) = C\delta(x)$ for $C > 0$. Thus, we find

$$F^{\boldsymbol{W}}_{2,\lambda} = \nabla^2_{\boldsymbol{W}=\lambda \boldsymbol{W}_0}(\lambda^{-1}\boldsymbol{V}_0 \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x})) \tag{C.16}$$

$$= \lambda^{-1}\nabla^2_{\boldsymbol{W}=\lambda \boldsymbol{W}_0}(\boldsymbol{V}_0 \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x})) \tag{C.17}$$

$$= \lambda^{-1}\nabla^2_{\boldsymbol{W}=\boldsymbol{W}_0}(\lambda^{-1}\boldsymbol{V}_0 \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x})) \tag{C.18}$$

$$= \lambda^{-2}\nabla^2_{\boldsymbol{W}=\boldsymbol{W}_0}(\boldsymbol{V}_0 \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x})) = \lambda^{-2} F^{\boldsymbol{W}}_{2,1}. \tag{C.19}$$

Similarly, $f_{\boldsymbol{\theta}^\lambda}$ is a linear function of the output layer thus

$$F^{\boldsymbol{V}}_{2,\lambda} = \nabla^2_{\boldsymbol{V}=\lambda^{-1}\boldsymbol{V}_0}(\boldsymbol{V}\mathrm{ReLU}(\lambda \boldsymbol{W}_0 \boldsymbol{x})) = 0.$$

This proves that Hessian exhibits the advertised behavior (4.3). Finally, (4.2) follows from the fact that the diagonal entries of the Hessian of the input layer decays with λ^2 whereas its entries grow with λ so that HI remains unchanged (and similar story for the output layer). ■

D Further Experiments and Comparison to Layer-wise Pruning

In Section 4 we used standard network pruning which prunes the whole set of weights to a certain sparsity level regardless of which layer they belong. We observed that MP can completely prune a layer when we

apply very large or small λ-scaling in Fig 3a. We also showed HP significantly mitigates this problem as it is inherently invariant to λ. Layer-wise pruning prunes the exact same fraction of the parameters in each layer individually and it is an alternative way to avoid the *layer death* problem. Thus, in this section, we compare standard pruning with layer-wise pruning and display the results in Fig. 4. Fig. 4a and 4d show that layer-wise MP mitigates the layer death problem under λ-scaling because it keeps the same fraction of nonzero parameters in each layer. However when $\lambda = 1$ the performance of layer-wise MP is worse than standard MP. Note that there is nothing special about $\lambda = 1$ except the fact that input dimension (784) and number of hidden nodes (1024) are close to each other and He initialization results in input and output weights of similar magnitudes.

Fig. 4b and 4e compare standard HP with layer-wise HP showing that standard HP outperforms layer-wise HP except when the network is extremely spares (fraction of nonzero $\leq 0.2\%$). Our explanation for this behavior is as follows: The weights of certain layers (specifically output layer) are more important, in average, than others (specifically input layer). The standard HP fully takes this into account by jointly pruning the complete set of weights based on importance. In Figure 4c it can be seen that, for 1% sparsity target, standard HP keeps around 50% of the output layer whereas layer-wise HP keeps exactly the target value 1% (Fig 4f). However the fact that standard HP favors the output layer weights results in input layer getting overly pruned in the extremely sparse regime and in this regime layer-wise pruning has a slight edge. However both methods lead to lackluster accuracy ($\sim 40\%$ accuracy on MNIST) in this regime, thus for practical purposes, it is plausible to say standard HP is better than or equal to layer-wise in all sparsities.

E Relaxing Conditions on Convex Gaussian Min-Max Theorem

The following lemma replaces the compactness constrained with the closedness in CGMT. It also applies to problems with random equality constraints (which is of interest for over-parameterized least-squares) besides regularized form.

Theorem E.1 (Flexible CGMT) *Let ψ be a function obeying $\lim_{\|\boldsymbol{w}\|_{\ell_2}\to\infty} \psi(\boldsymbol{w}) = \infty$. Given a closed set $\mathcal{S}$, define*

$$\Phi_\lambda(\boldsymbol{X}) = \min_{\boldsymbol{w}\in\mathcal{S}} \lambda\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2} + \psi(\boldsymbol{w}) \tag{E.1}$$

$$\phi_\lambda(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}} \lambda(\|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2} - \boldsymbol{h}^T\boldsymbol{w})_+ + \psi(\boldsymbol{w}), \tag{E.2}$$

and

$$\Phi_\infty(\boldsymbol{X}) = \min_{\boldsymbol{w}\in\mathcal{S},\boldsymbol{X}\boldsymbol{w}=0} \psi(\boldsymbol{w}) \tag{E.3}$$

$$\phi_\infty(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S},\|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2}\leq\boldsymbol{h}^T\boldsymbol{w}} \psi(\boldsymbol{w}). \tag{E.4}$$

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

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

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

Proof As an application of Theorem 3 of [63] and Lemma E.2 and Lemma E.3, these two statements hold for a compact $\mathcal{S}$ and compact convex $\mathcal{S}$ respectively. We remark that Theorem 3 of [63] does not explicitly state $\mathbb{P}(\Phi_\lambda(\boldsymbol{X}) > t) \leq 2\mathbb{P}(\phi_\lambda(\boldsymbol{X}) \geq t)$. However it is explicitly stated in the proof of this theorem (see Proof of Eq (13) in pg 22). Our goal is extending the proof to closed sets rather than compact. To achieve this, we consider a sequence of problems with the sets

$$\mathcal{S}_r = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_{\ell_2} \leq r\} \cap \mathcal{S}.$$

$\mathcal{S}_r$ is compact thus the advertised inequalities hold for $\mathcal{S}_r$. The remaining argument is showing pointwise convergence and applying the Dominated Convergence Theorem as in the proofs of Lemma E.2 and Lemma E.3. We will argue the result for $\lambda = \infty$. Finite λ follows essentially the identical argument. Define $\Phi_\lambda^r(\boldsymbol{X}) = \min_{\boldsymbol{w}\in\mathcal{S}_r, \boldsymbol{X}\boldsymbol{w}=0} \psi(\boldsymbol{w})$ and $\phi_\lambda^r(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}_r, \|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2}\le \boldsymbol{h}^T\boldsymbol{w}} \psi(\boldsymbol{w})$. Fix a matrix $\boldsymbol{X}$ and define the indicator $E_\lambda^r = 1_{\Phi_\lambda^r(\boldsymbol{X})<t}$. We claim that $\lim_{r\to\infty} E_\lambda^r = E_\lambda$. To see this consider the two cases: Case 1: If original problem is infeasible and $\Phi_\lambda(\boldsymbol{X}) = \infty$ then $\Phi_\lambda^r(\boldsymbol{X}) = \infty$ as well thus $\lim_{r\to\infty} E_\lambda^r = E_\lambda = 0$. Case 2: $\Phi_\lambda(\boldsymbol{X})$ is finite. By the divergence assumption on ψ, the set of optimal solutions $\boldsymbol{w}^*$ of the original problem achieving $\Phi_\lambda(\boldsymbol{X})$ lie on a bounded ℓ_2 set. Thus for sufficiently large r, $E_\lambda^r = E_\lambda$ (note that Φ_λ^r is a non-increasing function of r). To proceed, applying Dominated Convergence Theorem, this yields

$$\lim_{r\to\infty} \mathbb{E}[E_\lambda^r] = \mathbb{E}[E_\lambda] \iff \mathbb{P}(\Phi_\lambda(\boldsymbol{X}) < t) = \lim_{r\to\infty} \mathbb{P}(\Phi_\lambda^r(\boldsymbol{X}) < t).$$

Applying the same argument to ϕ_λ^r we obtain the desired bound

$$\begin{aligned}
\mathbb{P}(\Phi_\lambda(\boldsymbol{X}) < t) &= \lim_{r\to\infty} \mathbb{P}(\Phi_\lambda^r(\boldsymbol{X}) < t) && \text{(E.5)}\\
&\le 2 \lim_{r\to\infty} \mathbb{P}(\phi_\lambda^r(\boldsymbol{g},\boldsymbol{h}) \le t) && \text{(E.6)}\\
&\le 2\mathbb{P}(\phi_\lambda(\boldsymbol{g},\boldsymbol{h}) \le t). && \text{(E.7)}
\end{aligned}$$

Repeating the identical/very similar arguments for the convex case and finite λ (omitted for avoiding repetitions), we conclude the proof. Finally, the combination of upper and lower bounds yield the two sided bound by observing

$$\mathbb{P}(|\Phi_\lambda(\boldsymbol{X}) - \mu| > t) = \mathbb{P}(\Phi_\lambda(\boldsymbol{X}) > \mu + t) + \mathbb{P}(\Phi_\lambda(\boldsymbol{X}) < \mu - t).$$

■

E.1 Proof of Constrained CGMT

E.1.1 Proof for the convex case

Lemma E.2 *Given a convex and compact $\mathcal{S}$, define the PO and AO problems*

$$\begin{aligned}
\Phi_\infty(\boldsymbol{X}) &= \min_{\boldsymbol{w}\in\mathcal{S}, \boldsymbol{X}\boldsymbol{w}=0} \psi(\boldsymbol{w}) && \text{(E.8)}\\
\phi_\infty(\boldsymbol{g},\boldsymbol{h}) &= \min_{\boldsymbol{w}\in\mathcal{S}, \|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2}\le \boldsymbol{h}^T\boldsymbol{w}} \psi(\boldsymbol{w}). && \text{(E.9)}
\end{aligned}$$

Suppose $\boldsymbol{X},\boldsymbol{g},\boldsymbol{h} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Then, we have that

$$\mathbb{P}(\Phi_\infty(\boldsymbol{X}) > t) \le 2\mathbb{P}(\phi_\infty(\boldsymbol{g},\boldsymbol{h}) \ge t). \tag{E.10}$$

Proof Using convex-concavity of $\mathcal{L}_{\mathrm{PO}}(\boldsymbol{w},a) = a\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2} + \psi(\boldsymbol{w})$ we have that

$$\begin{aligned}
\Phi_\infty(\boldsymbol{X}) &= \min_{\boldsymbol{w}\in\mathcal{S}, \boldsymbol{X}\boldsymbol{w}=0} \psi(\boldsymbol{w}) && \text{(E.11)}\\
&= \max_{a\ge 0} \min_{\boldsymbol{w}\in\mathcal{S}} a\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2} + \psi(\boldsymbol{w}) && \text{(E.12)}\\
&= \lim_{\lambda\to\infty} \max_{0\le a\le\lambda} \min_{\boldsymbol{w}\in\mathcal{S}} a\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2} + \psi(\boldsymbol{w}) && \text{(E.13)}\\
&= \lim_{\lambda\to\infty} \min_{\boldsymbol{w}\in\mathcal{S}} \max_{0\le a\le\lambda} a\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2} + \psi(\boldsymbol{w}) && \text{(E.14)}\\
&= \lim_{\lambda\to\infty} \Phi_\lambda(\boldsymbol{X}). && \text{(E.15)}
\end{aligned}$$

Note that if the problem is infeasible, both sides yield ∞. Similarly using convex-concavity of $\mathcal{L}_{\mathrm{AO}}(\boldsymbol{w}, a) = a(\|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2} + \boldsymbol{h}^T\boldsymbol{w})_+ + \psi(\boldsymbol{w})$, we have

$$\Phi_\infty(\boldsymbol{g}, \boldsymbol{h}) = \lim_{\lambda\to\infty} \phi_\lambda(\boldsymbol{g}, \boldsymbol{h}).$$

Now that we connected the equality constrained problems Φ_∞ and ϕ_∞ to regularized problems, we proceed with establishing a probabilistic bound using CGMT. We remark that Theorem 3 of [63] does not explicitly state $\mathbb{P}(\Phi_\lambda(\boldsymbol{X}) > t) \le 2\mathbb{P}(\phi_\lambda(\boldsymbol{X}) \ge t)$. However it is explicitly stated in the proof of this theorem (see Proof of Eq (13) in pg 22). Define the indicator function $E_\lambda = 1_{\Phi_\lambda(\boldsymbol{X})>t}$. Observe that, for any choice of $\boldsymbol{X}$,

$$\lim_{\lambda\to\infty} E_\lambda = \lim_{\lambda\to\infty} 1_{\Phi_\lambda(\boldsymbol{X})>t} = 1_{\Phi_\infty(\boldsymbol{X})>t}.$$

Note that, if the problem is infeasible, then $\lim_{\lambda\to\infty} E_\lambda = E_\infty = 1$. To proceed, we are in a position to apply Dominated Convergence Theorem to find

$$\lim_{\lambda\to\infty} \mathbb{E}[E_\lambda] = \mathbb{E}[E_\infty] \iff \mathbb{P}(\Phi_\infty(\boldsymbol{X}) > t) = \lim_{\lambda\to\infty} \mathbb{P}(\Phi_\lambda(\boldsymbol{X}) > t). \tag{E.16}$$

Applying the identical argument on $\phi_{\boldsymbol{g},\boldsymbol{h}}$ to find $\mathbb{P}(\phi_\infty(\boldsymbol{g}, \boldsymbol{h}) \ge t) = \lim_{\lambda\to\infty} \mathbb{P}(\phi_\lambda(\boldsymbol{g}, \boldsymbol{h}) \ge t)$, we obtain the desired relation

$$\begin{aligned}
\mathbb{P}(\Phi_\infty(\boldsymbol{X}) > t) &= \lim_{\lambda\to\infty} \mathbb{P}(\Phi_\lambda(\boldsymbol{X}) > t) && \text{(E.17)}\\
&\le 2 \lim_{\lambda\to\infty} \mathbb{P}(\phi_\lambda(\boldsymbol{g}, \boldsymbol{h}) \ge t) && \text{(E.18)}\\
&= 2\mathbb{P}(\phi_\infty(\boldsymbol{g}, \boldsymbol{h}) \ge t). && \text{(E.19)}
\end{aligned}$$

■

E.1.2 Proof for the general case

Lemma E.3 *Given a compact set $\mathcal{S}$, define the PO and AO problems as in Lemma E.2. We have that*

$$\mathbb{P}(\Phi_\infty(\boldsymbol{X}) < t) \le 2\mathbb{P}(\phi_\infty(\boldsymbol{g}, \boldsymbol{h}) < t). \tag{E.20}$$

Proof The proof is similar to that of Lemma E.2. For a general compact set $\mathcal{S}$, application of Gordon's theorem yields the one-sided bound

$$\mathbb{P}(\Phi_\lambda(\boldsymbol{X}) < t) \le 2\mathbb{P}(\phi_\lambda(\boldsymbol{g}, \boldsymbol{h}) \le t). \tag{E.21}$$

To move from finite λ to infinite, we make use of Lemma E.4. Define the indicator function $E_\lambda = 1_{\Phi_\lambda(\boldsymbol{X})\le t}$. Using Lemma E.4, for any choice of $\boldsymbol{X}$, $\lim_{\lambda\to\infty} E_\lambda = \lim_{\lambda\to\infty} 1_{\Phi_\lambda(\boldsymbol{X})<t} = 1_{\Phi_\infty(\boldsymbol{X})<t}$. Note again that, if the problem is infeasible, then $\lim_{\lambda\to\infty} E_\lambda = E_\infty = 0$. To proceed, we are in a position to apply Dominated Convergence Theorem to find

$$\lim_{\lambda\to\infty} \mathbb{E}[E_\lambda] = \mathbb{E}[E_\infty] \iff \mathbb{P}(\Phi_\infty(\boldsymbol{X}) < t) = \lim_{\lambda\to\infty} \mathbb{P}(\Phi_\lambda(\boldsymbol{X}) < t). \tag{E.22}$$

Applying the identical argument on $\phi_{\boldsymbol{g},\boldsymbol{h}}$ to find $\mathbb{P}(\phi_\infty(\boldsymbol{g}, \boldsymbol{h}) \le t) = \lim_{\lambda\to\infty} \mathbb{P}(\phi_\lambda(\boldsymbol{g}, \boldsymbol{h}) \le t)$, we obtain the desired relation

$$\begin{aligned}
\mathbb{P}(\Phi_\infty(\boldsymbol{X}) < t) &= \lim_{\lambda\to\infty} \mathbb{P}(\Phi_\lambda(\boldsymbol{X}) < t) && \text{(E.23)}\\
&\le 2 \lim_{\lambda\to\infty} \mathbb{P}(\phi_\lambda(\boldsymbol{g}, \boldsymbol{h}) \le t) && \text{(E.24)}\\
&= 2\mathbb{P}(\phi_\infty(\boldsymbol{g}, \boldsymbol{h}) \le t). && \text{(E.25)}
\end{aligned}$$

■

Lemma E.4 *Let $\mathcal{S}$ be a compact set and $\psi(\cdot)$ be a continuous function and $f(\boldsymbol{w})$ be a non-negative continuous function. Then*

$$\lim_{\lambda\to\infty}\min_{\boldsymbol{w}\in\mathcal{S}}\lambda f(\boldsymbol{w})+\psi(\boldsymbol{w})=\min_{\boldsymbol{w}\in\mathcal{S},f(\boldsymbol{w})=0}\psi(\boldsymbol{w})$$

Thus, setting $f(\boldsymbol{w})=\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2}$ and $f(\boldsymbol{w})=\|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2}-\boldsymbol{h}^T\boldsymbol{w}$, we have that

$$\lim_{\lambda\to\infty}\Phi_\lambda(\boldsymbol{X})=\Phi_\infty(\boldsymbol{X})$$
$$\lim_{\lambda\to\infty}\phi_\lambda(\boldsymbol{g},\boldsymbol{h})=\phi_\infty(\boldsymbol{g},\boldsymbol{h}).$$

Proof Since f is continuous, it has closed sub-level sets. Suppose $\{\boldsymbol{w}\in\mathcal{S}\;\big|\;f(\boldsymbol{w})=0\}=\varnothing$. Since $\mathcal{S}$ is compact, both sides are infinity and the equality holds. To proceed, we assume the problem is feasible. If $\min_{\boldsymbol{w}\in\mathcal{S}}\psi(\boldsymbol{w})=\min_{\boldsymbol{w}\in\mathcal{S},f(\boldsymbol{w})=0}\psi(\boldsymbol{w})$ again both sides are equal to $\min_{\boldsymbol{w}\in\mathcal{S}}\psi(\boldsymbol{w})$ thus we assume the right-hand side objective is strictly larger than $\min_{\boldsymbol{w}\in\mathcal{S}}\psi(\boldsymbol{w})$. Define the sublevel sets $\mathcal{C}_\alpha=\mathcal{S}\cap\{\boldsymbol{w}\;\big|\;f(\boldsymbol{w})\le\alpha\}$.

Let $c_\lambda=\min_{\boldsymbol{w}\in\mathcal{S}}\lambda f(\boldsymbol{w})+\psi(\boldsymbol{w})$ and $c_\infty=\min_{\boldsymbol{w}\in\mathcal{S},f(\boldsymbol{w})=0}\psi(\boldsymbol{w})$. Let $\boldsymbol{w}_\lambda=\arg\min_{\boldsymbol{w}\in\mathcal{S}}\lambda f(\boldsymbol{w})+\psi(\boldsymbol{w})$ and $\boldsymbol{w}_\infty=\arg\min_{\boldsymbol{w}\in\mathcal{S},f(\boldsymbol{w})=0}\psi(\boldsymbol{w})$ be optimal solutions of regularized and constrained problems achieving c_λ,c_∞ respectively. If the claim is wrong, then for some $\varepsilon>0$ and all $\lambda>0$, $c_\lambda\le c_\infty-\varepsilon$. Since f is nonnegative, this also implies that $\psi(\boldsymbol{w}_\lambda)\le\psi(\boldsymbol{w}_\infty)-\varepsilon$.

Since ψ is a continuous function, ψ uniformly converges on $\mathcal{S}$. Uniform convergence implies that for any $\varepsilon>0$, there exists $\delta>0$ such that for all pairs $\|\boldsymbol{w}-\boldsymbol{v}\|_{\ell_2}<\delta$, we have $|\psi(\boldsymbol{w})-\psi(\boldsymbol{v})|<\varepsilon$. Conversely, if $|\psi(\boldsymbol{w})-\psi(\boldsymbol{v})|\ge\varepsilon$, we have that $\|\boldsymbol{w}-\boldsymbol{v}\|_{\ell_2}\ge\delta$. In our context, this means that, for all $\lambda\ge0$

$$\operatorname{dist}(\boldsymbol{w}_\lambda,\mathcal{C}_0)\ge\delta.$$

Set $\Gamma=\psi(\boldsymbol{w}_\infty)-\min_{\boldsymbol{w}\in\mathcal{S}}\psi(\boldsymbol{w})>0$. For any $\lambda\ge0$, $\lambda f(\boldsymbol{w}_\lambda)\le\Gamma\implies f(\boldsymbol{w}_\lambda)\le\Gamma/\lambda\implies\boldsymbol{w}_\lambda\in\mathcal{C}_{\Gamma/\lambda}$. This implies that for any choice of $\alpha>0$ (via $\alpha\leftrightarrow\Gamma/\lambda$), $\mathcal{C}_\alpha$ contains points that are δ away from $\mathcal{C}_0$. Note that $\mathcal{C}_\alpha$ is a non-decreasing sequence of sets (i.e. $\mathcal{C}_{\alpha_1}\subseteq\mathcal{C}_{\alpha_2}$ whenever $\alpha_1\le\alpha_2$). Via Bolzano–Weierstrass theorem $(\boldsymbol{w}_\lambda)_{\lambda\ge\Gamma}$ contains a convergent subsequence. Index this subsequence by $(\boldsymbol{w}_{\lambda_i})_{i=1}^\infty$ and suppose $\bar{\boldsymbol{w}}=\lim_{i\to\infty}\boldsymbol{w}_{\lambda_i}$. Clearly $\operatorname{dist}(\bar{\boldsymbol{w}},\mathcal{C}_0)\ge\delta$ as distance is a continuous function. Note that $\bar{\boldsymbol{w}}\in\mathcal{C}_\alpha$ for any $\alpha>0$ since $\mathcal{C}_\alpha$ is non-decreasing and compact thus $\mathcal{C}_\alpha$ contains all the elements of $(\boldsymbol{w}_{\lambda_i})_{i=1}^\infty$ after a certain point including its limit. Finally, define $\bar{\mathcal{C}}=\lim_{\alpha\to0^+}\mathcal{C}_\alpha=\bigcap_{\alpha>0}\mathcal{C}_\alpha$. Clearly $\bar{\boldsymbol{w}}\in\bar{\mathcal{C}}$. This means that $\bar{\mathcal{C}}$ contains the element $\bar{\boldsymbol{w}}$ which is not inside $\mathcal{C}_0$. Finally, this leads to contradiction since $\bar{\mathcal{C}}\subseteq\mathcal{C}_0$. Specifically, if $\bar{\boldsymbol{w}}\in\bar{\mathcal{C}}$, then this implies

$$f(\bar{\boldsymbol{w}})\le\alpha\text{ for all }\alpha>0\implies f(\bar{\boldsymbol{w}})=0\implies\bar{\boldsymbol{w}}\in\mathcal{C}_0.$$

This concludes the proof. ∎