Safety vs. Performance: How Multi-Objective Learning Reduces Barriers to Market Entry

Meena Jagadeesan, Michael I. Jordan, and Jacob Steinhardt

University of California, Berkeley

September 27, 2024

Abstract

Emerging marketplaces for large language models and other large-scale machine learning (ML) models appear to exhibit market concentration, which has raised concerns about whether there are insurmountable barriers to entry in such markets. In this work, we study this issue from both an economic and an algorithmic point of view, focusing on a phenomenon that *reduces* barriers to entry. Specifically, an incumbent company risks reputational damage unless its model is sufficiently aligned with safety objectives, whereas a new company can more easily avoid reputational damage. To study this issue formally, we define a multi-objective high-dimensional regression framework that captures reputational damage, and we characterize the number of data points that a new company needs to enter the market. Our results demonstrate how multi-objective considerations can fundamentally reduce barriers to entry—the required number of data points can be significantly smaller than the incumbent company's dataset size. En route to proving these results, we develop scaling laws for high-dimensional linear regression in multi-objective environments, showing that the scaling rate becomes slower when the dataset size is large, which could be of independent interest.

1 Introduction

Large language models and other large-scale machine learning (ML) models have led to an important shift in the information technology landscape, one which has significant economic consequences. Whereas earlier generations of ML models provided the underpinnings for platforms and services, new models—such as language models—are themselves the service. This has led to new markets where companies offer language models as their service and compete for user usage. As in other markets, it is important to reason about market competitiveness: in particular, to what extent there are barriers to entry for new companies.

A widespread concern about these markets is that new companies face insurmountable *barriers to entry* that drive market concentration [Vipra and Korinek, 2023]. The typical argument is that incumbent companies with high market share can purchase or capture significant amounts of data and compute,[1] and then invest these resources into the training of models that achieve even higher performance [Kaplan et al., 2020]. This suggests that the company's market share would further increase, and that the scale and scope of this phenomenon would place incumbent companies beyond the reach of new companies trying to enter the market. The scale is in fact massive—language

[1] Large companies can afford these resources since the marketplace is an economy of scale (i.e., fixed costs of training significantly exceed per-query inference costs). They also generate high volumes of data from user interactions.

assistants such as ChatGPT and Gemini each have hundreds of millions of users [Cook, 2024]. In light of the concerns raised by policymakers [Vipra and Korinek, 2023] and regulators [The White House, 2023, European Union, 2022b] regarding market concentration, it is important to investigate the underlying economic and algorithmic mechanisms at play.

While standard arguments assume that market share is determined by model performance, the reality is that the incumbent company risks reputational damage if their model violates safety-oriented objectives. For example, incumbent companies face public and regulatory scrutiny for their model's safety violations—such as threatening behavior [Perrigo, 2023], jailbreaks [Wei et al., 2023], and releasing dangerous information [The White House, 2023]—even when the model performs well in terms of helpfulness and usefulness to users. In contrast, new companies face less regulatory scrutiny since compliance requirements often prioritize models trained with more resources [The White House, 2023, California Legislature, 2024], and new companies also may face less public scrutiny given their smaller user bases.

In this work, we use a multi-objective learning framework to show that the threat of reputational damage faced by the incumbent company can reduce barriers to entry. For the incumbent, the possibility of reputational damage creates pressure to align with safety objectives in addition to optimizing for performance. Safety and performance are not fully aligned, so improving safety can reduce performance as a side effect. Meanwhile, the new company faces less of a risk of reputational damage from safety violations. The new company can thus enter the marketplace with significantly less data than the incumbent company, a phenomenon that our model and results formalize.

Model and results. We analyze a stylized marketplace based on multi-objective linear regression (Section 2). The performance-optimal output and the safety-optimal output are specified by two different linear functions of the input x. The marketplace consists of two companies: an incumbent company and a new company attempting to enter the market. Each company receives their own unlabelled training dataset, decides what fraction of training data points to label according to the performance-optimal vs. safety-optimal outputs, and then runs ridge regression. The new company requires a less stringent level of safety to avoid reputational damage than the incumbent company. We characterize the *market-entry threshold* N_E^* (Definition 1) which captures how much data the new company needs to outperform the incumbent company.

First, as a warmup, we characterize N_E^* when the new company faces no safety constraint and the incumbent company has infinitely many data points (Section 3). Our key finding is that the new company can enter the market with finite data, even when the incumbent company has infinite data (Theorem 1; Figure 1). Specifically, we show that the threshold N_E^* is finite; moreover, it is increasing in the correlation (i.e., the alignment) between performance and safety, and it is decreasing in a problem-specific scaling law exponent.

Next, we turn to more general environments where the incumbent has finite data $N_I < \infty$ (Section 4.2). We find that the threshold N_E^* scales sublinearly with the incumbent's dataset size N_I, as long as N_I is sufficiently large. In fact, the threshold N_E^* scales at a slower rate as N_I increases: that is, $N_E^* = \Theta(N_I^c)$ where the exponent c is decreasing in N_I (Theorem 4; Figure 3). For example, for concrete parameter settings motivated by language models [Hoffmann et al., 2022], the exponent c decreases from 1 to 0.75 to 0 as N_I increases. In general, the exponent c takes on up to three different values depending on N_I, and is strictly smaller than 1 as long as N_I is sufficiently large.

Finally, we turn to environments where the new company also faces a nontrivial safety constraint, assuming for simplicity that the incumbent company again has infinite data (Section 4.3). We find that N_E^* is finite as long as the new company faces a strictly weaker safety constraint than the incumbent. When the two safety thresholds are closer together, the new company needs more data and in fact needs to scale up their dataset size at a faster rate: that is, $N_E^* = \Theta(D^{-c})$, where D

measures the difference between the safety thresholds and where the exponent c increases as D decreases (Theorem 5; Figure 4). For the parameter settings in [Hoffmann et al., 2022], the exponent c changes from -2.94 to -3.94 to an even larger value as D decreases. In general, the exponent c takes on up to three different values.

Technical tool: Scaling laws. To prove our results, we derive scaling laws for *multi-objective* high-dimensional linear regression, which could be of independent interest (Section 4.1; Figure 2). We study optimally-regularized ridge regression where some of the training data is labelled according to the primary linear objective (capturing performance) and the rest is labelled according to an alternate linear objective (capturing safety).

We characterize data-scaling laws for both the loss along the primary objective and the excess loss along the primary objective relative to an infinite-data ridgeless regression. Our scaling laws quantify the rate at which the loss (Theorem 2; Figure 2a) and the excess loss (Theorem 3; Figure 2b) decay with the dataset size N, and how this rate is affected by the fraction of data labelled according to each objective and other problem-specific quantities. Our analysis improves upon recent works on scaling in multi-objective environments [e.g., Jain et al., 2024, Song et al., 2024] by allowing for non-identity covariances and problem-specific regularization, which leads to new insights about scaling laws as we describe below.

Our results reveal that the scaling rate becomes slower as the dataset size increases, illustrating that multi-objective scaling laws behave qualitatively differently from classical single-objective environments. While a typical scaling exponent in a single-objective environment takes on a single value across all settings of N, the scaling exponent for multi-objective environments decreases as N increases. In particular, the scaling exponent takes on *three different values* depending on the size of N relative to problem-specific parameters. The intuition is that the regularizer must be carefully tuned to N in order to avoid overfitting to training data labelled according to the alternate objective, which in turn results in the scaling exponent being dependent on N (Section 5).

Discussion. Altogether, our work highlights the importance of looking beyond model performance when evaluating market entry in machine learning marketplaces. Our results highlight a disconnect between market entry in single-objective environments versus more realistic multi-objective environments. More broadly, a company's susceptibility to reputational damage affects how they train their model to balance between different objectives. As we discuss in Section 6, these insights have nuanced implications for regulators who wish to promote both market competitiveness and safety compliance, and also generalize beyond language models to online platforms.

1.1 Related work

Our work connects to research threads on *competition between model providers* as well as *scaling laws and high-dimensional linear regression.*

Competition between model providers. Our work contributes to an emerging line of work studying how competing model providers strategically design their machine learning pipelines to attract users. Model-provider actions range from choosing a function from a model class [Ben-Porat and Tennenholtz, 2017, 2019, Jagadeesan et al., 2023b], to selecting a regularization parameter [Iyer and Ke, 2022], to choosing an error distribution over user losses [Feng et al., 2019], to making data purchase decisions [Dong et al., 2019, Kwon et al., 2022], to deciding whether to share data [Gradwohl and Tennenholtz, 2023], to selecting a bandit algorithm [Aridor et al., 2020, Jagadeesan et al., 2023a]. While these works assume that model providers win users solely by maximizing (individual-level or population-level) accuracy, our framework incorporates the role of *safety violations* in impacting user retention implicitly via reputational damage. Moreover, our focus is on quantifying the barriers to market entry, rather than analyzing user welfare or the equilibrium decisions of model providers.

Other related work includes the study of competition between algorithms [Immorlica et al., 2011, Kleinberg and Raghavan, 2021], retraining dynamics under user participation decisions [Hashimoto et al., 2018, Ginart et al., 2021, Dean et al., 2022, Shekhtman and Dean, 2024, Su and Dean, 2024], the bargaining game between a foundation model company and a specialist [Laufer et al., 2024], and the market power of an algorithmic platform to shape user populations [Perdomo et al., 2020, Hardt et al., 2022, Mendler-Dünner et al., 2024].

Our work also relates to platform competition [Jullien and Sand-Zantman, 2021, Calvano and Polo, 2021], the emerging area of competition policy and regulation of digital marketplaces [Stigler Committee, 2019, Vipra and Korinek, 2023, Cen et al., 2023, Competition and Markets Authority, 2024], the study of how antitrust policy impacts innovation in classical markets [Baker, 2007, Segal and Whinston, 2007], and industrial organization more broadly [Tirole, 1988]. For example, recent work examines how increased public scrutiny from inclusion in the S&P 500 can harm firm performance [Bennett et al., 2023], how privacy regulation impacts firm competition [Gal and Aviv, 2020, Fallah et al., 2024], how regulatory inspections affect incentives to comply with safety constraints [Harrington, 1988, Fallah and Jordan, 2023], and how data-driven network effects can reduce innovation [Prüfer and Schottmüller, 2021].

Scaling laws and high-dimensional linear regression. Our work also contributes to an emerging line of work on scaling laws which study how model performance changes with training resources. Empirical studies have demonstrated that increases to scale often reliably improve model performance [e.g., Kaplan et al., 2020, Hernandez et al., 2021, Hoffmann et al., 2022], but have also identified settings where scaling behavior is more nuanced [e.g., Muennighoff et al., 2023, Gao et al., 2023]. We build on a recent mathematical characterization of scaling laws based on high-dimensional linear regression [e.g., Hastie et al., 2019, Bordelon et al., 2020, Bahri et al., 2021, Cui et al., 2021, Wei et al., 2022, Bach, 2023, Wei, 2024, Patil et al., 2024, Bordelon et al., 2024, Mallinar et al., 2024, Lin et al., 2024, Atanasov et al., 2024]. However, while these works focus on *single-objective* environments where all of the training data is labelled with outputs from a single predictor, we consider *multi-objective* environments where some fraction of the training data is labelled according to an alternate predictor.

We note that a handful of recent works similarly move beyond single-objective environments and study scaling laws where the training data comes a mixture of different data sources. Jain et al. [2024], Song et al. [2024] study high-dimensional ridge regression in a similar multi-objective environment to our setup. However, these results assume an *identity covariance* and focus on fixed regularization or no regularization. In contrast, we allow for richer covariance matrices that satisfy natural power scaling (Section 2.3), and we analyze optimally tuned regularization. Our analysis of these problem settings yields new insights about scaling behavior: for example, the scaling rate becomes slower with dataset size (Theorems 2-3). Other related works study scaling laws under mixtures of covariate distributions [Hashimoto, 2021], under data-quality heterogeneity [Goyal et al., 2024], under data addition [Shen et al., 2024], under mixtures of AI-generated data and real data [Dohmatob et al., 2024, Gerstgrasser et al., 2024], and with respect to the contribution of individual data points [Covert et al., 2024].

More broadly, our work relates to collaborative learning [Blum et al., 2017, Mohri et al., 2019, Sagawa et al., 2020, Haghtalab et al., 2022], federated learning [see Yang et al., 2019, for a survey], optimizing data mixtures [e.g., Rolf et al., 2021, Xie et al., 2023], and adversarial robustness [e.g., Raghunathan et al., 2020]. Finally, our work relates to non-monotone scaling laws in strategic environments [Jagadeesan et al., 2023a, Handina and Mazumdar, 2024], where increases to scale can worsen equilibrium social welfare.

2 Model

We define our linear-regression-based marketplace (Section 2.1), justify the design choices of our model (Section 2.2), and then delineate our statistical assumptions (Section 2.3).

2.1 Linear regression-based marketplace

We consider a marketplace where two companies fit linear regression models in a multi-objective environment.

Linear regression model. To formalize each company's machine learning pipeline, we consider the multi-objective, high-dimensional linear regression model described below. This multi-objective environment aims to capture how ML models are often trained to balance multiple objectives which are in tension with each other, and we consider linear regression since it has often accurately predicted scaling trends of large-scale machine learning models (see Section 2.2 for additional discussion).

More concretely, given an input $x \in \mathbb{R}^P$, let $\langle \beta_1, x \rangle$ be the output that targets performance maximization, and let $\langle \beta_2, x \rangle$ be the output that targets safety maximization. Given a linear predictor β, the performance loss is evaluated via a population loss, $L_1(\beta) = \mathbb{E}_{x \sim \mathcal{D}_F}[(\langle \beta_1, x \rangle - \langle \beta, x \rangle)^2]$, and the safety violation is captured by a loss $L_2(\beta) = \mathbb{E}_{x \sim \mathcal{D}_F}[(\langle \beta_2, x \rangle - \langle \beta, x \rangle)^2]$, where $\mathcal{D}_F$ is the input distribution.

The model provider implicitly determines how to balance β_1 and β_2 when determining how to label their training dataset. In particular, each model provider is given an unlabelled training dataset $X \in \mathbb{R}^{N \times P}$ with N inputs drawn from $\mathcal{D}_F$. To generate labels, they select the fraction $\alpha \in [0, 1]$ of training data to label according to each objective. They then sample a fraction α of the training data uniformly from X and label it as $Y_i = \langle \beta_1, X_i \rangle$; the remaining $1-\alpha$ fraction is labelled as $Y_i = \langle \beta_2, X_i \rangle$. The model provider fits a ridge regression on the labelled training dataset with least-squares loss $\ell(y, y') = (y - y')^2$, and thus solves: $\hat{\beta}(\alpha, \lambda, X) = \operatorname{argmin}_\beta \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \langle \beta, X_i \rangle)^2 + \lambda ||\beta||_2^2 \right)$.

Marketplace. The marketplace contains two companies, an *incumbent company* I already in the market and a *new (entrant) company* E trying to enter the market. At a high level, each company $C \in \{I, E\}$ faces reputational damage if their safety violation exceeds their safety constraint τ_C. Each company company C is given N_C unlabelled data points sampled from $\mathcal{D}_F$, and selects a mixture parameter α_C and regularizer λ_C to maximize their performance given their safety constraint τ_C. We assume that the incumbent company I faces a stricter safety constraint, $\tau_I < \tau_E$, due to increased public or regulatory scrutiny (see Section 2.2 for additional discussion).

When formalizing how the model providers choose hyperparameters, we make the following simplications. First, rather than work directly with the performance and safety losses of the ridge regression estimator, we assume for analytic tractability that they approximate these losses by $L_1^* := L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$ and $L_2^* := L_2^*(\beta_1, \beta_2, \mathcal{D}_F, \alpha)$ defined as follows.

- *Performance:* We define L_1^* to be a *deterministic equivalent* $L_1^{\mathsf{det}}(\beta_1, \beta_2, \Sigma, \lambda, N, \alpha)$ which we derive in Lemma 6. The deterministic equivalent [cf. Hachem et al., 2007] is a tool from random matrix theory that is closely linked to the Marčenko-Pastur law [Marčenko and Pastur, 1967]. Under standard random matrix assumptions (Assumption 1), the deterministic equivalent asymptotically approximates the loss $L_1(\hat{\beta}(\alpha, \lambda, X))$ when X is constructed from N i.i.d. samples from $\mathcal{D}_F$ (see Appendix D for additional discussion).
- *Safety:* For analytic simplicity, in the main body of the paper, we define L_2^* to be the safety violation of the infinite-data ridgeless regression estimator with mixture parameter α.[2] In Appendix E, we

[2] The infinite-data ridgeless regression estimator is $\operatorname{argmin}_\beta \left(\alpha \cdot \mathbb{E}_{x \sim \mathcal{D}_F}[\langle \beta - \beta_1, x \rangle^2] + (1 - \alpha) \cdot \mathbb{E}_{x \sim \mathcal{D}_F}[\langle \beta - \beta_2, x \rangle^2] \right)$. For this specification, the dataset size N and the regularization parameter λ only affect L_1^* and not L_2^*, which

instead define L_2^* analogously to L_1^*—i.e., as a deterministic equivalent $L_2^{\mathrm{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$—and extend our model and results to this more complex setting.[3]

Second, we assume that $(\beta_1, \beta_2) \sim \mathcal{D}_W$ for some joint distribution $\mathcal{D}_W$ and that the model providers take expectations when choosing hyperparameters, since it will be easier to specify assumptions in Section 4.3 over distributions of predictors.

Within this setup, a company C faces reputational damage if the safety violation exceeds a certain threshold:

$$\mathbb{E}_{(\beta_1,\beta_2)\sim\mathcal{D}_W}[L_2^*(\beta_1, \beta_2, \mathcal{D}_F, \alpha_C)] > \tau_C.$$

We assume that the safety thresholds for the two companies satisfy the following inequalities:

$$\tau_E >_{(A)} \tau_I \geq_{(B)} \mathbb{E}_{(\beta_1,\beta_2)\sim\mathcal{D}_W}[L_2^*(\beta_1, \beta_2, \mathcal{D}_F, 0.5)]. \tag{1}$$

Here, inequality (A) captures the notion that the incumbent needs to achieve higher safety to avoid reputational damage. Inequality (B) guarantees that both companies, $C \in \{I, E\}$, can set the mixture parameter $\alpha_C \geq 0.5$ without facing reputational damage, and thus ensures that the safety constraint does not dominate the company's optimization task.[4]

The company selects $\alpha \in [0.5, 1]$ and $\lambda \in (0, 1)$ to maximize their performance subject to their safety constraint, as formalized by the following optimization program:[5]

$$(\alpha_C, \lambda_C) = \mathrm{argmin}_{\alpha\in[0.5,1],\lambda\in(0,1)}\, \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N_C, \alpha)] \text{ s.t. } \mathbb{E}_{\mathcal{D}_W}[L_2^*(\beta_1, \beta_2, \mathcal{D}_F, \alpha)] \leq \tau_C.$$

Market-entry threshold. We define the market-entry threshold to capture the minimum number of data points N_E that the new company needs to collect to achieve better performance than the incumbent company while avoiding reputational damage.

Definition 1. *The market-entry threshold $N_E^*(N_I, \tau_I, \tau_E, \mathcal{D}_W, \mathcal{D}_F)$ is the minimum value of $N_E \in \mathbb{Z}_{\geq 1}$ such that $\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_E, N_E, \alpha_E)] \leq \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_I, N_I, \alpha_I)]$.*

The goal of our work is to analyze the function $N_E^*(N_I, \tau_I, \tau_E, \mathcal{D}_W, \mathcal{D}_F)$.

2.2 Model discussion

Now that we have formalized our statistical model, we discuss and justify our design choices in greater detail. We defer a discussion of limitations to Section 6.

Presence of competing objectives. Our multi-objective formulation is motivated by how ML models are often trained to balance multiple objectives which are in tension with each other. In some cases, the pretraining objective is in tension with the finetuning objective [Wei et al., 2023]. For example, the fine-tuning of a language model to be more aligned with user intent can degrade performance—e.g., because the model hedges too much—which creates an "alignment tax" [Ouyang et al., 2022]. In other cases, fine-tuning approaches themselves balance multiple objectives such as helpfulness (which can be mapped to performance in our model) and harmlessness (which can be mapped to safety in our model) [Bai et al., 2022]. These objectives can be in tension with one another, for example if the user asks for dangerous information.

simplifies our analysis in Sections 3-4 and enables us to obtain tight characterizations.

[3] We directly extend our results in Section 3, and we also show relaxed versions of our results in Section 4.

[4] More specifically, inequality (B) ensures that the safety constraint still allows both companies to label 50% of their training data according to the performance-optimal outputs.

[5] Technically, the optimum might be achieved at $\lambda = 0$ or $\lambda = 1$, and the min should be replaced by a inf.

High-dimensional linear regression as a statistical model. We focus on high-dimensional linear regression due to its ability to capture scaling trends observed in large-scale machine learning models such as language models, while still retaining analytic tractability. In particular, in single-objective environments, scaling trends for high-dimensional linear regression recover the empirically observed power-law scaling of the loss with respect to the dataset size [Kaplan et al., 2020, Cui et al., 2021, Wei et al., 2022]. Moreover, from an analytic perspective, the structural properties of high-dimensional linear regression make it possible to characterize the loss using random matrix machinery (see Appendix D).

Impact of market position on model provider constraint τ. Our assumption that $\tau_E > \tau_I$ (inequality (A) in (1)) is motivated by how large companies face greater reputational damage from safety violations than smaller companies. One driver of this unevenness in reputational damage is *regulation*: for example, recent regulation and policy [The White House, 2023, California Legislature, 2024] places stricter requirements on companies that use significant amounts of compute during training. In particular, these companies face more stringent compliance requirements in terms of safety assessments and post-deployment monitoring. Another driver of uneven reputational damage is *public perception*: we expect that the public is more likely to uncover safety violations for large companies, due to the large volume of user queries to the model. In contrast, for small companies, safety violations may be undetected or subject to less public scrutiny.

2.3 Assumptions on linear regression problem

To simplify our characterization of scaling trends, we follow prior work on high-dimensional linear regression [see, e.g., Cui et al., 2021, Wei et al., 2022] and make the following empirically motivated power-law assumptions. Let $\Sigma = \mathbb{E}_{x \sim \mathcal{D}_F}[xx^T]$ be the covariance matrix, and let λ_i and v_i be the eigenvalues and eigenvectors, respectively. We require the eigenvalues to decay with scaling exponent $\gamma > 0$ according to $\lambda_i = i^{-1-\gamma}$ for $1 \leq i \leq P$. For the alignment coefficients $\langle \beta_j, v_i \rangle$, it is cleaner to enforce power scaling assumptions in expectation, so that we can more easily define a correlation parameter. We require that for some $\delta > 0$, the alignment coefficients satisfy $\mathbb{E}_{\mathcal{D}_W}[\langle \beta_j, v_i \rangle^2] = i^{-\delta}$, where v_i is the ith eigenvector of Σ, for $j \in \{1, 2\}$ and $1 \leq i \leq P$. We also introduce a similar condition on the joint alignment coefficients, requiring that for some $\rho \in [0, 1)$, it holds that $\mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i \rangle \langle \beta_2, v_i \rangle] = \rho \cdot i^{-\delta}$. Finally, we assume an overparameterized limit where the number of parameters $P \to \infty$ approaches infinity. Below, we provide an example which satisfies these assumptions.

Example 1. *Suppose that the covariance Σ is a diagonal matrix with diagonal given by $\lambda_i = i^{-1-\gamma}$. Let the joint distribution over β_1 and β_2 be a multivariate Gaussian such that:*

$$\mathbb{E}_{\mathcal{D}_W}[(\beta_{j_1})_{i_1}(\beta_{j_2})_{i_2}] = \begin{cases} 0 & \text{if } 1 \leq j_1, j_2 \leq 2, 1 \leq i_1 \neq i_2 \leq P \\ i_1^{-\delta} & \text{if } 1 \leq j_1 = j_2 \leq 2, 1 \leq i_1 = i_2 \leq P \\ \rho \cdot i_1^{-\delta} & \text{if } 1 \leq j_1 \neq j_2 \leq 2, 1 \leq i_1 = i_2 \leq P. \end{cases}$$

This implies that $\mathbb{E}_{\mathcal{D}_W}[\langle \beta_j, v_i \rangle^2] = i^{-\delta}$ and $\mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i \rangle \langle \beta_2, v_i \rangle] = \rho \cdot i^{-\delta}$.

We adopt the random matrix theory assumptions on the covariance matrix and linear predictors from Bach [2023] (see Assumption 1 in Appendix D), which guarantee that the Marčenko-Pastur law holds [Marčenko and Pastur, 1967]. That is, the covariance $(\hat{\Sigma} + \lambda I)^{-1}$ of the samples can be approximated by a deterministic quantity (see Appendix D.1 for a more detailed discussion). We leverage this Marčenko-Pastur law to derive a deterministic equivalent $L_1^{\texttt{det}}$ for the performance loss $L_1(\hat{\beta}(\alpha, \lambda, X))$ of the ridge regression estimator (Lemma 6).

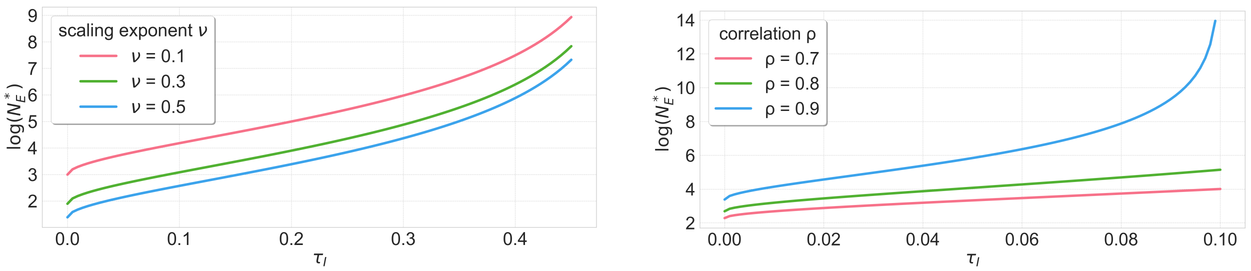


Figure 1: Market-entry threshold N_E^* as a function of the incumbent's safety constraint τ_I, when the incumbent has infinite data and entrant has no safety constraint (Theorem 1). The plots show varying values of the scaling exponent ν where the correlation parameter $\rho = 0.5$ is held fixed (left) and varying values of ρ where $\nu = 0.34$ is held fixed (right). The market-entry threshold N_E^* is finite. It is also higher when the constraint τ_I is weaker, when the correlation ρ is stronger, and when the scaling exponent ν is lower.

3 Warm Up: Infinite-Data Incumbent and Unconstrained Entrant

As a warmup, we analyze the market entry N_E^* threshold in a simplified environment where the incumbent has infinite data and the new company faces no safety constraint. In this result, we place standard power-law scaling assumptions on the covariance and alignment coefficients (Section 2.3) and we characterize the threshold N_E^* up to constants (Theorem 1; Figure 1).

Theorem 1. *Suppose that power-law scaling holds for the eigenvalues and alignment coefficients, with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Suppose that the incumbent company has infinite data (i.e., $N_I = \infty$), and that the entrant faces no constraint on their safety (i.e., $\tau_E = \infty$). Suppose that the safety constraint τ_I satisfies (1). Then, it holds that:*[6]

$$N_E^*(\infty, \tau_I, \infty, \mathcal{D}_W, \mathcal{D}_F) = \Theta\left(\left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^{-2/\nu}\right),$$

where $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma(\beta_1 - \beta_2)] = \Theta(1 - \rho)$, and where $\nu := \min(2(1 + \gamma), \delta + \gamma)$.

The intuition is as follows. The safety constraint τ_I forces the incumbent company to partially align their predictor with the safety objective β_2. Since β_1 and β_2 point in different directions, this reduces the performance of the incumbent along β_1 as a side effect, resulting in strictly positive loss with respect to performance. On the other hand, since the new company faces no safety constraint, the new company can optimize entirely for performance along β_1. This means that the new company can enter the market as long as their finite data error is bounded by the incumbent's performance loss. We formalize this intuition in the following proof sketch.

Proof sketch of Theorem 1. The incumbent chooses the *infinite-data* ridgeless estimator $\beta(\alpha, 0)$ with mixture parameter $\alpha \in [0, 1]$ tuned so the safety violation is τ_I (Lemma 11). The resulting performance loss is $\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}$. Since the new company has no safety constraint, they choose the *single-objective* ridge regression estimator where $\alpha = 1$ and where λ is chosen optimally.[7] Theorem 2 (or alternatively, existing analyses of high-dimensional linear regression [e.g., Cui et al., 2021, Wei et al., 2022]) demonstrate the loss follows a scaling law of the form $\inf_{\lambda > 0} L_1(\hat{\beta}(1, \lambda, X)) = \Theta\left(N^{-\nu}\right)$ where $\nu := \min(2(1 + \gamma), \delta + \gamma)$. The full proof is in Appendix A. □

[6] Throughout the paper, we allow $\Theta()$ and $O()$ to hide implicit constant which depends on the scaling exponents γ, δ.

[7] We formally rule out the possibility that $\alpha \neq 1$ using our multi-objective scaling law in Theorem 2.

Theorem 1 reveals that the market-entry threshold is *finite* as long as (1) the safety constraint τ_I places nontrivial restrictions on the incumbent company and (2) the safety and performance objectives are not perfectly correlated. This result captures the notion that the new company can enter the market even after the incumbent company has accumulated an infinite amount of data.

Theorem 1 further illustrates how the market-entry threshold changes with other parameters (Figure 1). When safety and performance objectives are more correlated (i.e., when ρ is higher), the market-entry threshold increases, which increases barriers to entry. When the safety constraint for the incumbent is weaker (i.e., when τ_I is higher), the market-entry threshold also increases. Finally, when the power scaling parameters of the covariance and alignment coefficients increase, which increases the scaling law exponent ν, the market-entry threshold decreases.

4 Generalized Analysis of the Market-entry Threshold

To obtain a more fine-grained characterization of the market-entry threshold, we now consider more general environments. Our key technical tool is *multi-objective scaling laws*, which capture the performance of ridge regression in high-dimensional, multi-objective environments with finite data (Section 4.1). Using these scaling laws, we characterize the market-entry threshold when the incumbent has finite data (Section 4.2) and when the new company has a safety constraint (Section 4.3).

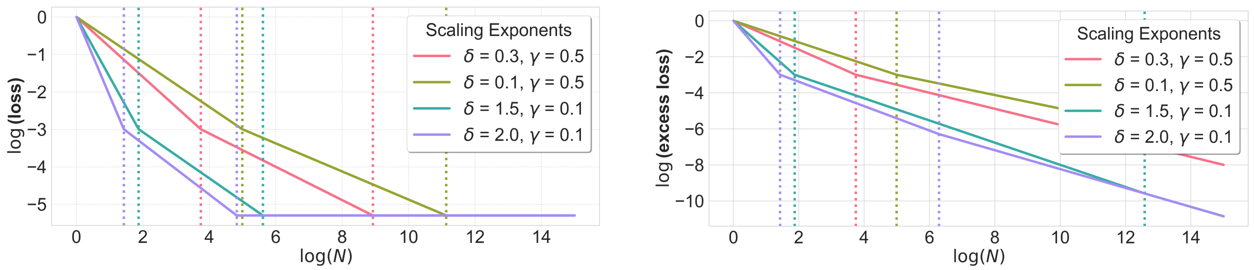
Our results in this section uncover the following conceptual insights about market entry. First, our main finding from Section 3—that the new company can enter the market with significantly less data than the incumbent—applies to these generalized environments. Moreover, our characterizations of N_E^* exhibit a *power-law-like dependence* with respect to the incumbent's dataset size (Theorem 4) and the difference in safety requirement for the two companies (Theorem 5). Interestingly, the scaling exponent c is not a constant across the full regime and instead takes on up to three different values. As a consequence, the new company can afford to scale up their dataset at a slower rate as the incumbent's dataset size increases, but needs to scale up their dataset at a faster rate as the two safety constraints become closer together. Proofs are deferred to Appendix B.

4.1 Technical tool: Scaling laws in multi-objective environments

In this section, we give an overview of multi-objective scaling laws (see Section 5 for a more formal treatment and derivations). Our scaling laws capture how the ridge regression loss $L_1(\hat{\beta}(\alpha, \lambda, X))$ along the primary objective β_1 scales with the dataset size N, when the regularizer λ is optimally tuned to both N and problem-specified parameters. We show scaling laws for both the loss $\inf_{\lambda \in (0,1)} \mathbb{E}[L_1(\hat{\beta}(\alpha, \lambda, X))]$ and the excess loss $\inf_{\lambda \in (0,1)}(\mathbb{E}[L_1(\hat{\beta}(\alpha, \lambda, X)) - L_1(\beta(\alpha, 0))])$ where $\beta(\alpha, 0)$ is the infinite-data ridgeless regression estimator.

Scaling law for the loss. We first describe the scaling law for $\inf_{\lambda \in (0,1)} \mathbb{E}[L_1(\hat{\beta}(\alpha, \lambda, X))]$ (Theorem 2; Figure 2a).

Theorem 2 (Informal Version of Corollary 8)**.** *Suppose that the power-law scaling assumptions from Section 2.3 hold with exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$. Suppose also that $P = \infty$ and $\alpha \geq 0.5$. Then, a deterministic equivalent for the expected loss under optimal regularization $\inf_{\lambda \in (0,1)} \mathbb{E}[L_1(\hat{\beta}(\alpha, \lambda, X))]$ scales according to $N^{-\nu^*}$, where the scaling exponent ν^* is*



(a) Scaling law for loss (up to constants)

(b) Scaling law for excess loss (up to constants)

Figure 2: Data scaling laws for multi-objective environments where a fraction $\alpha = 0.9$ of the data is labelled according to the primary objective and a fraction $1 - \alpha = 0.1$ is labelled according to the secondary objective. The plots show, up to constants, the loss $\Theta(\inf_{\lambda \in (0,1)} \mathbb{E}[L_1(\hat{\beta}(\alpha, \lambda, X))])$ (left, Theorem 2) and excess loss $\Theta(\inf_{\lambda \in (0,1)}(\mathbb{E}[L_1(\hat{\beta}(\alpha, \lambda, X)) - L_1(\beta(\alpha, 0))]))$ (right, Theorem 3) as a function of the total number of training data points N. The loss and excess loss both take the form N^{-c}, but where the scaling exponent c takes on *multiple (two or three) different values* depending on the size of N relative to other parameters. The scaling exponent is smaller when N is larger, thus demonstrating that the scaling rate becomes slower as the dataset size N increases.

defined to be:

$$\nu^* = \begin{cases} \nu & \textit{if } N \leq (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \\ \frac{\nu}{\nu+1} & \textit{if } (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \\ 0 & \textit{if } N \geq (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}}, \end{cases}$$

for $\nu := \min(2(1+\gamma), \delta + \gamma)$.

Theorem 2 (Figure 2a) illustrates that the scaling rate becomes slower as the dataset size N increases. In particular, while the scaling exponent in single-objective environments is captured by a *single* value, Theorem 2 illustrates that the scaling exponent ν^* in multi-objective environments takes on *three different values*, depending on the size of N relative to other parameters. When N is small (the first regime), the scaling exponent $\nu^* = \nu$ is identical to that of the single-objective environment given by β_1. When N is a bit larger (the second regime), the scaling exponent *reduces* to $\nu^* = \nu/(\nu + 1) < \nu$. To make this concrete, if we take $\nu = 0.34$ to be an empirically estimated scaling law exponent for language models [Hoffmann et al., 2022], this would mean that $\nu^* \approx 0.34$ in the first regime and $\nu^* \approx 0.25$ in the second regime. Finally, when N is sufficiently large (the third regime), the scaling exponent reduces all the way to $\nu^* = 0$ and the only benefit of additional data is to improve constants on the loss.

Scaling law for the excess loss. We next turn to the excess loss, $\inf_{\lambda \in (0,1)}(\mathbb{E}[L_1(\hat{\beta}(\alpha, \lambda, X)) - L_1(\beta(\alpha, 0))])$, which is normalized by the loss of the infinite-data ridgeless predictor $\beta(\alpha, 0)$. We show that the excess loss exhibits the same scaling behavior as the loss when N is sufficiently small, but exhibits different behavior when N is sufficiently large (Theorem 3; Figure 2b).

Theorem 3 (Informal Version of Corollary 10)**.** *Suppose that the power-law scaling assumptions from Section 2.3 hold with exponents* $\gamma, \delta > 0$ *and correlation coefficient* $\rho \in [0, 1)$*. Suppose also that* $P = \infty$ *and* $\alpha \geq 0.75$*. Then, a deterministic equivalent for the expected loss under optimal regularization* $\inf_{\lambda \in (0,1)}(\mathbb{E}[L_1(\hat{\beta}(\alpha, \lambda, X)) - L_1(\beta(\alpha, 0))])$ *scales according to* $N^{-\nu^*}$*, where the scaling* *exponent ν^* is defined to be:*

$$\nu^* = \begin{cases} \nu & \text{if } N \leq (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \\ \frac{\nu}{\nu+1} & \text{if } (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}} \\ \frac{\nu'}{\nu'+1} & \text{if } N \geq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}, \end{cases}$$

for $\nu := \min(2(1+\gamma), \delta+\gamma)$ and $\nu' := \min(1+\gamma, \delta+\gamma)$.

Theorem 3 (Figure 2b) again shows that the scaling rate can become slower as the dataset size N increases, and again reveals three regimes of scaling behavior. While the first two regimes of Theorem 3 resemble the first two regimes of Theorem 2, the third regime of Theorem 3 (where $N \geq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}$) behaves differently. In this regime, the scaling exponent for the excess loss is $\frac{\nu'}{\nu'+1}$, rather than zero—this captures the fact that additional data can nontrivially improve the excess loss even in this regime, even though it only improves the loss up to constants. In terms of the magnitude of the scaling exponent $\frac{\nu'}{\nu'+1}$, it is *strictly smaller* than the scaling exponent $\frac{\nu}{\nu+1}$ when $\delta > 1$ and *equal* to the scaling exponent $\frac{\nu}{\nu+1}$ when $\delta \leq 1$.

4.2 Finite data for the incumbent

We compute N_E^* when the incumbent has finite data and the new company has no safety constraint (Theorem 4; Figure 3). The market-entry threshold N_E^* depends on the incumbent's dataset size N_I, the incumbent's performance loss G_I if they were to have infinite data but face the same safety constraint, the scaling exponents γ, δ, and the correlation coefficient ρ.

Theorem 4. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Assume that $\tau_E = \infty$. Suppose that the safety constraint τ_I satisfies (1). Then we have that $N_E^* = N_E^*(N_I, \tau_I, \infty, \mathcal{D}_W, \mathcal{D}_F)$ satisfies:*

$$N_E^* := \begin{cases} \Theta\left(N_I\right) & \text{if } N_I \leq G_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \\ \Theta\left(N_I^{\frac{1}{\nu+1}} \cdot G_I^{-\frac{1}{2(\nu+1)}}(1-\rho)^{-\frac{1}{2(\nu+1)}}\right) & \text{if } G_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \leq N_I \leq G_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}} \\ \Theta\left(G_I^{-\frac{1}{\nu}}\right) & \text{if } N_I \geq G_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}}, \end{cases}$$

where $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1-\beta_2)^T\Sigma(\beta_1-\beta_2)] = \Theta(1-\rho)$, where $G_I := (\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))})^2$, and where $\nu = \min(2(1+\gamma), \delta+\gamma)$.

The market-entry threshold in Theorem 4 exhibits three regimes of behavior depending on N_I. In particular, the market-entry threshold takes the form $N_E^* = \Theta(N_I^c)$ where c decreases from 1 (in the first regime) to $\frac{1}{\nu+1}$ (in the second regime) to 0 (in the third regime) as N_I increases. To connect this to large-language-model marketplaces, we directly set $\nu = 0.34$ to be the empirically estimated scaling law exponent for language models [Hoffmann et al., 2022]; in this case, the scaling exponent c ranges from 1 to 0.75 to 0. The fact that there are three regimes come from the scaling law derived in Theorem 2, as the following proof sketch illustrates.

Proof sketch. The key technical tool is the scaling law for the loss $\inf_{\lambda\in(0,1)} \mathbb{E}[L_1(\hat{\beta}(\alpha, \lambda, X))]$ (Theorem 2), which has three regimes of scaling behavior for different values of N. We apply the scaling law to analyze the performance of the incumbent, who faces a safety constraint and has finite

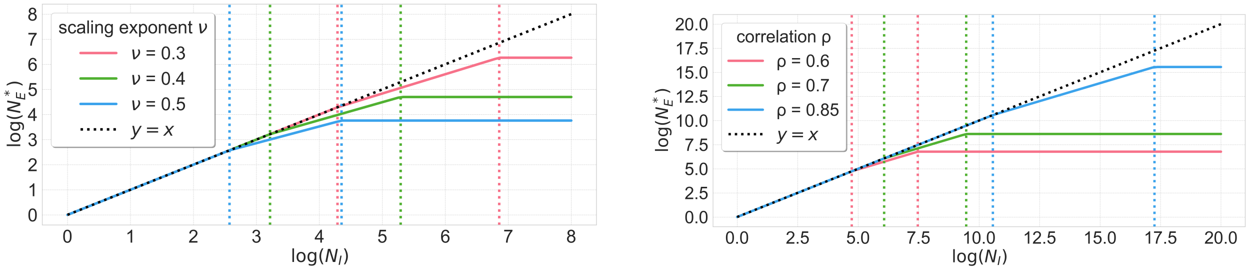


Figure 3: The market-entry threshold N_E^* as a function of the incumbent dataset size N_I, when the new company has no safety constraint (Theorem 4). The plots show varying values of the scaling exponent ν where the correlation parameter $\rho = 0.5$ is held fixed (left) and varying values of ρ where $\nu = 0.34$ is held fixed (right). When N_I is sufficiently large, the market-entry threshold N_E^* is asymptotically less than N_I (i.e., below the dotted black line). Each curve is the union of three line segments with slope decreasing in N_I, demonstrating that the new company can afford to scale up their dataset at a slower rate as N_I increases.

data. Analyzing the performance of the new company—who faces no safety constraint—is more straightforward, given that the new company can set $\alpha_E = 1$. We compute N_E^* as the number of data points needed to match the incumbent's performance level. The full proof is deferred to Appendix B.1. □

Theorem 4 reveals that the new company can enter the market with $N_E^* = o(N_I)$ data, as long as the incumbent's dataset is sufficiently large (i.e., $N_I \geq G_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}}$). The intuition is when there is sufficient data, the multi-objective scaling exponent is worse than the single-objective scaling exponent (Theorem 2). The incumbent thus faces a worse scaling exponent than the new company, so the new company can enter the market with asymptotically less data.

The three regimes in Theorem 4 further reveal that the market-entry threshold N_E^* scales at a slower rate as the incumbent's dataset size N_I increases (Figure 3). The intuition is that the multi-objective scaling exponent ν^* faced by the incumbent decreases as dataset size increases, while the single-objective scaling exponent faced by the new company is constant in dataset size (Theorem 2). The incumbent thus becomes less efficient at using additional data to improve performance, while the new company's efficiency in using additional data remains unchanged.

Theorem 4 also offers finer-grained insight into the market-entry threshold in each regime. In the first regime, where the incumbent's dataset is small, the threshold N_E^* matches the incumbent dataset size—the new company does not benefit from having a less stringent safety constraint. In the second (intermediate) regime, the new company can enter with a dataset size proportional to $N_I^{1/(\nu+1)}$. This *polynomial speedup* illustrates that the new company can more efficiently use additional data to improve performance than the incumbent company. A caveat is that this regime is somewhat restricted in that the ratio of the upper and lower boundaries is bounded. In the third regime, where the incumbent's dataset size is large, the market-entry threshold N_E^* matches the market-entry threshold from Theorem 1 where the incumbent has *infinite* data.

4.3 Safety constraint for the new company

We compute N_E^* when the new company has a nontrivial safety constraint and the incumbent has infinite data. For this result, we strengthen the conditions on τ_E and τ_I from (1), instead requiring:

$$\tau_E >_{(A)} \tau_I \geq_{(B)} \mathbb{E}_{(\beta_1,\beta_2)\sim\mathcal{D}_W}[L_2^*(\beta_1,\beta_2,\mathcal{D}_F,0.75)], \tag{2}$$

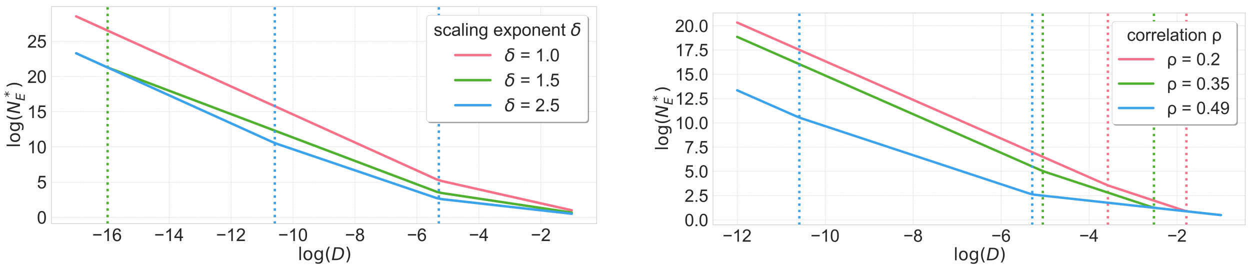


Figure 4: The market-entry threshold N_E^* as a function of the difference D between the infinite-data performance loss of the incumbent and new company, when the incumbent has infinite data (Theorem 5). The plots show varying values of the scaling exponent δ where the correlation parameter $\rho = 0.49$ is held fixed (left) and varying values of ρ where $\delta = 2.5$ is held fixed (right). The plots are shown in log space. The market-entry threshold is finite in all cases. Each curve is the union of multiple line segments with slope increasing in magnitude as $\log D$ decreases, demonstrating that the new company needs to scale up their dataset at a faster rate as D decreases.

where (2) replaces the 0.5 with a 0.75 in the right-most quantity.[8]

We state the result below (Theorem 5; Figure 4). The market-entry threshold N_E^* depends on the incumbent's safety constraint τ_I, the performance loss G_I (resp. G_E) if the incumbent (resp. new company) had infinite data and faced the same safety constraint, the difference $D = G_I - G_E$ in infinite-data performance loss achievable by the incumbent and new company, the scaling exponents γ, δ, and the correlation coefficient ρ.

Theorem 5. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Suppose that the safety constraints τ_I and τ_E satisfy (2). Then it holds that $N_E^* = N_E^*(\infty, \tau_I, \tau_E, \mathcal{D}_W, \mathcal{D}_F)$ satisfies:*

$$N_E^* := \begin{cases} \Theta(D^{-\frac{1}{\nu}}) & \text{if } D \geq G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}} \\ \Theta\left(D^{-\frac{\nu+1}{\nu}} G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}}\right) & \text{if } G_E^{\frac{\nu}{2(\nu-\nu')}}(1-\rho)^{\frac{\nu}{2(\nu-\nu')}} \leq D \leq G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}} \\ \Theta\left(\left(D \cdot G_E^{-\frac{1}{2}}(1-\rho)^{-\frac{1}{2}}\right)^{-\frac{\nu'+1}{\nu'}}\right) & \text{if } D \leq G_E^{\frac{\nu}{2(\nu-\nu')}}(1-\rho)^{\frac{\nu}{2(\nu-\nu')}}, \end{cases}$$

where $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma(\beta_1 - \beta)] = \Theta(1-\rho)$, where $\nu = \min(2(1+\gamma), \delta+\gamma)$ and $\nu' = \min(1+\gamma, \delta+\gamma)$, where $G_I := \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^2$ and $G_E := \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_E, L^*(\rho))}\right)^2$, and where $D := G_I - G_E$.

The market-entry threshold in Theorem 5 also exhibits three regimes of behavior depending on the difference D in the infinite-data performance loss achievable by the incumbent and new company. In particular, the market-entry threshold takes the form $N_E^* = \Theta(D^{-c})$ where c increases from $\frac{1}{\nu}$ to $\frac{\nu+1}{\nu}$ to $\frac{\nu'+1}{\nu'}$ as D decreases. (The third regime only exists when $\delta > 1$.) To connect this to large-language-model marketplaces, if we take $\nu = 0.34$ to be the empirically estimated scaling law exponent for language models [Hoffmann et al., 2022], then c would range from 2.94 to 3.94 to

[8] Inequality (B) in (2) requires that the safety constraint still allows both company to label 75% of their training data according to performance-optimal outputs. We make this modification, since our analysis of multi-objective scaling laws for the *excess* loss assumes $\alpha \geq 0.75$ (see Section 5.3).

potentially even larger. The fact that there are three regimes come from the scaling law derived in Theorem 3, as the following proof sketch illustrates.

Proof sketch. The key technical tool is the scaling law for the *excess loss* $\inf_{\lambda\in(0,1)}(\mathbb{E}[L_1(\hat{\beta}(\alpha,\lambda,X))-L_1(\beta(\alpha,0))])$ (Theorem 3), which has three regimes of scaling behavior for different values of N. We apply the scaling law to analyze the performance of the new company, who faces a safety constraint and has finite data. Analyzing the performance of the incumbent—who has infinite data—is more straightforward, and the incumbent's performance loss is $G_I = D + G_E$. We compute the number of data points N_E^* needed for the new company to achieve an excess loss of D. The full proof is deferred to Appendix B.2. □

Theorem 5 illustrates that the new company can enter the market with finite data N_E^*, as long as the safety constraint τ_E placed on the new company is strictly weaker than the constraint τ_I placed on the incumbent company (inequality (A) in (2)). This translates to the difference D being strictly positive. The intuition is that when the new company faces a weaker safety constraint, it can train on a greater number of data points labelled with the performance objective β_1, which improves performance.

The three regimes in Theorem 5 further reveal that the market-entry threshold N_E^* scales at a faster rate as the difference D between the two safety constraints decreases (Figure 3). The intuition is since the new company needs to achieve an excess loss of at most D, the new company faces a smaller multi-objective scaling exponent ν^* as D decreases (Theorem 3). The new company thus becomes less efficient at using additional data to improve performance.

5 Deriving Scaling Laws for Multi-Objective Environments

We formalize and derive our multi-objective scaling laws for the loss (Theorem 2) and excess loss (Theorem 3). Recall that the problem setting is high-dimensional ridge regression when a fraction α of the training data is labelled according to β_1 and the rest is labelled according to an alternate objective β_2. First, following the style of analysis of single-objective ridge regression [e.g., Cui et al., 2021, Wei et al., 2022], we first compute a *deterministic equivalent* of the loss (Section 5.1). Then we derive the scaling law under the power scaling assumptions on the eigenvalues and alignment coefficients in Section 2.3, both for the loss (Section 5.2) and for the excess loss (Section 5.3). Proofs are deferred to Appendix C.

5.1 Deterministic equivalent

We show that the loss of the ridge regression estimator can be approximated as a deterministic quantity. This analysis builds on the random matrix tools in Bach [2023] (see Appendix D). Note that our derivation of the deterministic equivalent does *not* place the power scaling assumptions on the eigenvalues or alignment coefficients; in fact, it holds for any linear regression setup which satisfies a standard random matrix theory assumption (Assumption 1).

We compute the following deterministic equivalent (proof deferred to Appendix C.5).[9]

Lemma 6. *Suppose that $N \geq 1$, $P \geq 1$, $\mathcal{D}_F$, β_1, and β_2 satisfy Assumption 1. Let Σ be the covariance matrix of $\mathcal{D}_F$, and let $\alpha \in [0,1]$ and $\lambda \in (0,1)$ be general parameters. Let $\Sigma_c = (\Sigma + cI)$*

[9] Following Bach [2023], the asymptotic equivalence notation $u \sim v$ means that u/v tends to 1 as N and P go to ∞.

for $c \geq 0$*, let* $B^{\mathtt{sn}} = \beta_1 \beta_1^T$*, let* $B^{\mathtt{df}} = (\beta_1 - \beta_2)(\beta_1 - \beta_2)^T$*, and let* $B^{\mathtt{mx}} = (\beta_1 - \beta_2)\beta_1^T$*. Let* $\kappa = \kappa(\lambda, N, \Sigma)$ *from Definition 2. Then, it holds that*

$$L_1(\hat{\beta}(\alpha, \lambda, X)) \sim L_1^{\mathtt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha) =: \frac{T_1 + T_2 + T_3 + T_4 + T_5}{Q},$$

where:

$$T_1 := \kappa^2 \cdot \mathrm{Tr}(\Sigma \Sigma_\kappa^{-2} B^{\mathtt{sn}}), \quad T_2 := (1-\alpha)^2 \left(\mathrm{Tr}\left(\Sigma_\kappa^{-2} \Sigma^3 B^{\mathtt{df}}\right)\right)$$

$$T_3 := 2(1-\alpha)\kappa \cdot \mathrm{Tr}\left(\Sigma_\kappa^{-2} \Sigma^2 B^{\mathtt{mx}}\right), \quad T_4 := -2(1-\alpha)\kappa \frac{1}{N} \mathrm{Tr}(\Sigma^2 \Sigma_\kappa^{-2}) \cdot \mathrm{Tr}\left(\Sigma_\kappa^{-1} \Sigma B^{\mathtt{mx}}\right),$$

$$T_5 := (1-\alpha)\frac{1}{N} \mathrm{Tr}(\Sigma^2 \Sigma_\kappa^{-2}) \cdot \left(\mathrm{Tr}\left(\Sigma B^{\mathtt{df}}\right) - 2(1-\alpha)\,\mathrm{Tr}\left(\Sigma_\kappa^{-1} \Sigma^2 B^{\mathtt{df}}\right)\right), \quad Q := 1 - \frac{1}{N} \mathrm{Tr}(\Sigma^2 \Sigma_\kappa^{-2}).$$

Lemma 6 shows that the loss can be approximated by a deterministic quantity $L_1^{\mathtt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$ which is sum of five terms, normalized by the standard degrees of freedom correction Q^{-1} [Bach, 2023]. The sum $T_1 + T_2 + T_3$ is the loss of infinite-data ridge regression with regularizer κ. Terms T_4 and T_5 capture additional error terms.

In more detail, term T_1/Q captures the standard single-objective environment error for N data points [Bach, 2023]: i.e., the population error of the single-objective linear regression problem with regularizer λ where all of the N training data points are labelled with β_i. Term T_2 is similar to the infinite-data ridgeless regression error but is slightly smaller due to regularization. Term T_3 is a cross term which is upper bounded by the geometric mean of term T_1 and term T_2. Term T_4 is another cross term which is subsumed by the other terms. Term T_5 captures an overfitting error which increases with the regularizer κ and decreases with the amount of data N.

From deterministic equivalents to scaling laws. In the following two subsections, using the deterministic equivalent from Lemma 6, we derive *scaling laws*. We make use of the the power scaling assumptions on the covariance and alignment coefficients described in Section 5.2, under which the deterministic equivalent takes a cleaner form (Lemma 26 in Appendix C). We note that strictly speaking, deriving scaling laws requires controlling the error of the deterministic equivalent relative to the actual loss; for simplicity, we do not control errors and instead directly analyze the deterministic equivalent.

5.2 Scaling law for the loss

We derive scaling laws for the loss $L_1^{\mathtt{det}} := L_1^{\mathtt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$. We first prove the following scaling law for a general regularizer λ (proof deferred to Appendix C.7).

Theorem 7. *Suppose that the power-law scaling assumption holds for the eigenvalues and alignment coefficients with scaling exponents* $\gamma, \delta > 0$ *and correlation coefficient* $\rho \in [0, 1)$*, suppose that* $P = \infty$*. Assume that* $\alpha \geq 0.5$ *and* $\lambda \in (0, 1)$*. Let* $L_1^{\mathtt{det}} := L_1^{\mathtt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$ *be the deterministic equivalent from Lemma 6. Let* $\nu := \min(2(1+\gamma), \delta + \gamma)$*. Then, the expected loss satisfies:*

$$\mathbb{E}_{\mathcal{D}_W}[L_1^{\mathtt{det}}] = \Theta\left(\underbrace{\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu})}_{\textit{finite data error}} + \underbrace{(1-\alpha)^2 \cdot (1-\rho)}_{\textit{mixture error}} + \underbrace{(1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)(1-\rho)}_{\textit{overfitting error}}\right).$$

Theorem 7 illustrates that the loss is the sum of a *finite data error*, an *overfitting error*, and a *mixture error*. The finite data error for $L_1^{\texttt{det}}$ matches the loss in the single-objective environment for N data points labelled with objective β_1. The mixture error equals the loss of the infinite-data ridgeless regression predictor $\beta(\alpha, 0)$. The overfitting error for $L_1^{\texttt{det}}$ equals the error incurred when the regularizer λ is too small. This term is always at most $(1-\alpha)^{-1}$ times larger than the mixture error, and it is smaller than the mixture error when λ is sufficiently large relative to N.

Due to the overfitting error, the optimal loss is *not* necessarily achieved by taking $\lambda \to 0$ for multi-objective linear regression. In fact, if the regularizer decays too quickly as a function of N (i.e., if $\lambda = O(N^{-1-\gamma})$), then the error would converge to $(1-\alpha)(1-\rho)$, which is a factor of $(1-\alpha)^{-1}$ higher than the error of the infinite-data ridgeless predictor $\beta(\alpha, 0)$. The fact that $\lambda \to 0$ is suboptimal reveals a sharp disconnect between the multi-objective setting and the single-objective setting where no explicit regularization is necessary to achieve the optimal loss [see, e.g., Cui et al., 2021, Wei et al., 2022].[10]

In the next result, we compute the optimal regularizer and derive a scaling law under optimal regularization as a corollary of Theorem 7.

Corollary 8 (Formal version of Theorem 2)**.** *Consider the setup of Theorem 7. Then, the loss under optimal regularization can be expressed as:*

$$\inf_{\lambda \in (0,1)} \mathbb{E}_{\mathcal{D}_W}[L_1^{\textit{det}}] = \begin{cases} \Theta\left(N^{-\nu}\right) & \textit{if } N \leq (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \\ \Theta\left(\left(\frac{N}{(1-\alpha)(1-\rho)}\right)^{-\frac{\nu}{\nu+1}}\right) & \textit{if } (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \\ \Theta((1-\alpha)^2(1-\rho)) & \textit{if } N \geq (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}}, \end{cases}$$

where $\nu := \min(2(1+\gamma), \delta + \gamma)$.

The scaling law exponent ν^* ranges from ν, to $\nu/(\nu+1)$, to 0 (Figure 2a). To better understand each regime, we provide intuition for when error term from Theorem 7 dominates, the form of the optimal regularizer, and the behavior of the loss.

- *Regime 1:* $N \leq (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$. Since N is small, the finite data error dominates regardless of λ. As a result, like in a single-objective environment, taking $\lambda = O(N^{-1-\gamma})$ recovers the optimal loss up to constants. Note that the loss thus behaves as if all N data points were labelled according to β_i: the learner benefits from *all* of the data, not just the data is labelled according to β_i.

- *Regime 2:* $(1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$. In this regime, the finite error term and overfitting error dominate. Taking $\lambda = \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{1+\gamma}{\nu+1}}\right)$, which equalizes the two error terms, recovers the optimal loss up to constants. The loss in this regime improves with N, but at a slower rate than in a single-objective environment.

- *Regime 3:* $N \geq (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$. Since N is large, the mixture and the overfitting error terms dominate. Taking $\lambda = \Theta((N(1-\alpha))^{-1-\gamma})$, which equalizes the two error terms, recovers the optimal loss up to constants. The loss behaves (up to the constants) as if there were *infinitely many data points* from the mixture distribution with weight α. This is the minimal possible loss and there is thus no additional benefit for data beyond improving constants.

The full proof of Corollary 8 is deferred to Appendix C.8.

[10] Tempered overfitting [Mallinar et al., 2022] can similarly occur in single-objective settings with *noisy observations*. In this sense, labelling some of the data with the alternate objective β_2 behaves qualitatively similarly to noisy observations.

5.3 Scaling law for the excess loss

Now, we turn to scaling laws for the excess loss $\mathbb{E}_{\mathcal{D}_W}[L_1^{\mathtt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha) - L_1(\beta(\alpha, 0))]$, , which is normalized by the loss of the infinite-data ridgeless predictor $\beta(\alpha, 0)$. We first prove the following scaling law for a general regularizer λ, assuming that $\alpha \geq 0.75$ (proof deferred to Appendix C.9).[11]

Theorem 9. *Suppose that the power-law scaling assumption holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, suppose that $P = \infty$. Assume that $\alpha \geq 0.75$ and $\lambda \in (0, 1)$. Let $L_1^{\boldsymbol{det}} := L_1^{\boldsymbol{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$ be the deterministic equivalent from Lemma 6. Let $\nu := \min(2(1 + \gamma), \delta + \gamma)$ and let $\nu' = \min(1 + \gamma, \delta + \gamma)$ Then, the expected loss satisfies:*

$$\mathbb{E}_{\mathcal{D}_W}[L_1^{\boldsymbol{det}} - L_1(\beta(\alpha, 0))]$$
$$= \Theta\left(\underbrace{\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu})}_{\textit{finite data error}} + \underbrace{(1-\rho)(1-\alpha)\max(\lambda^{\frac{\nu'}{1+\gamma}}, N^{-\nu'})}_{\textit{mixture finite data error}} + \underbrace{(1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)(1-\rho)}_{\textit{overfitting error}}\right).$$

Theorem 9 illustrates that the loss is the sum of a *finite data error*, an *overfitting error*, and a *mixture finite data error*. In comparison with Theorem 7, the difference is that the mixture error is replaced by the mixture finite data error. Interestingly, the mixture finite data error exhibits a different asymptotic dependence with respect to λ and N than the finite data error: the asymptotic rate of decay scales with ν' rather than ν. In fact, the rate is *slower* for the mixture finite data error than the finite data error as long as $\delta > 1$ (since this means that $\nu' < \nu$).

Since the optimal excess loss is also not necessarily achieved by taking $\lambda \to 0$, we compute the optimal regularizer for the excess loss and derive a scaling law under optimal regularization as a corollary of Theorem 9.

Corollary 10 (Formal version of Theorem 3)**.** *Consider the setup of Theorem 9. The excess loss under optimal regularization can be expressed as:*

$$\inf_{\lambda \in (0,1)} (\mathbb{E}_{\mathcal{D}_W}[L_1^{\boldsymbol{det}} - L_1(\beta(\alpha, 0))])$$
$$= \begin{cases} \Theta\left(N^{-\nu}\right) & \textit{if } N \leq (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \\ \Theta\left(\left(\frac{N}{(1-\alpha)(1-\rho)}\right)^{-\frac{\nu}{\nu+1}}\right) & \textit{if } (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}} \\ \Theta\left((1-\alpha)(1-\rho)N^{-\frac{\nu'}{\nu'+1}}\right) & \textit{if } N \geq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}, \end{cases}$$

where $\nu := \min(2(1 + \gamma), \delta + \gamma)$ and $\nu' = \min(1 + \gamma, \delta + \gamma)$.

The scaling law exponent ν^* ranges from ν, to $\nu/(\nu + 1)$, to $\nu'/(\nu' + 1)$ (Figure 2b). The first two regimes behave similarly to Corollary 8, and the key difference arises in the third regime (when N is large). In the third regime ($N \geq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}$), the mixture finite data error and the overfitting error terms dominate. Taking $\lambda = \Theta\left(N^{-\frac{1+\gamma}{\nu'+1}}\right)$—which equalizes these two error terms—recovers the optimal loss up to constants. The resulting scaling behavior captures that in this regime, additional data meaningfully improves the *excess loss*, even though additional data only improves the loss in terms of constants. The full proof of Corollary 10 is deferred to Appendix C.10.

[11] The assumption that $\alpha \geq 0.75$ simplifies the closed-form expression for the deterministic equivalent of the excess loss in Lemma 26. We defer a broader characterization of scaling laws for the excess loss to future work.

6 Discussion

We studied market entry in marketplaces for machine learning models, showing that pressure to satisfy safety constraints can reduce barriers to entry for new companies. We modelled the marketplace using a high-dimensional multi-objective linear regression model. Our key finding was that a new company can consistently enter the marketplace with significantly less data than the incumbent. En route to proving these results, we derive scaling laws for multi-objective regression, showing that the scaling rate becomes slower when the dataset size is large.

Potential implications for regulation. Our results have nuanced design consequences for regulators, who implicitly influence the level of safety that each company needs to achieve to avoid reputational damage. On one hand, our results suggest that placing greater scrutiny on dominant companies can encourage market entry and create a more competitive marketplace of model providers. On the other hand, market entry does come at a cost to the safety objective: the smaller companies exploit that they can incur more safety violations while maintaining their reputation, which leads to a race to the bottom for safety. Examining the tradeoffs between market competitiveness and safety compliance is an important direction for future work.

Barriers to market entry for online platforms. While we focused on language models, we expect that our conceptual findings about market entry also extend to recommendation and social media platforms.

In particular, our motivation and modeling assumptions capture key aspects of these online platforms. Policymakers have raised concerns have been raised about barriers to entry for social media platforms [Stigler Committee, 2019], motivated by the fact that social media platforms such as X and Facebook each have over a half billion users [Statista, 2024, Ingram, 2024]. Incumbent companies risk reputational damage if their model violates safety-oriented objectives—many recommendation platforms have faced scrutiny for promoting hate speech [European Union, 2022a], divisive content [Rathje et al., 2021], and excessive use by users [Hasan et al., 2018], even when recommendations perform well in terms of generating user engagement. This means that incumbent platforms must balance optimizing engagement with controlling negative societal impacts [Bengani et al., 2022]. Moreover, new companies face less regulatory scrutiny, given that some regulations explicitly place more stringent requirements on companies with large user bases: for example the Digital Services Act [European Union, 2022a] places a greater responsibility on Very Large Online Platforms (with over 45 million users per month) to identify and remove illegal or harmful content.

Given that incumbent platforms similarly face more pressure to satisfy safety-oriented objectives, our results suggest that multi-objective learning can also reduce barriers to entry for new online platforms.

Limitations. Our model for interactions between companies and users makes several simplifying assumptions. For example, we focused entirely whether the new company can enter the market, which leaves open the question of whether the new company can survive in the long run. Moreover, we assumed that all users choose the model with the highest overall performance. However, different users often care about performance on different queries; this could create an incentive for specialization, which could also reduce barriers to entry and market concentration. Finally, we focused on direct interactions between model providers and users, but in reality, downstream providers sometimes build services on top of a foundation model. Understanding how these market complexities affect market entry as well as long-term concentration is an interesting direction for future work.

Furthermore, our model also made the simplifying assumption that performance and safety trade off according to a multi-objective regression problem. However, not all safety objectives fit the mold of linear coefficients within linear regression. For some safety objectives such as privacy, we still expect that placing greater scrutiny on dominant companies could similarly reduce barriers to entry. Nonetheless, for other safety or societal considerations, we do expect that the implications for market entry might be fundamentally different. For example, if the safety objective is a multi-group performance criteria, and there is a single predictor that achieves zero accuracy on all distributions, then a dominant company with infinite data would be able to retain all users even if the company faces greater scrutiny. Extending our model to capture a broader scope of safety objectives is a natural direction for future work.

7 Acknowledgments

We thank Alireza Fallah, Jiahai Feng, Nika Haghtalab, Andy Haupt, Erik Jones, Jon Kleinberg, Ben Laufer, Neil Mallinar, Judy Shen, Alex Wei, Xuelin Yang, and Eric Zhao for useful feedback on this project. This work was partially supported by an Open Philanthropy AI fellowship and partially supported by the European Union (ERC-2022-SYG-OCEAN-101071601).

References

Guy Aridor, Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: The perils of exploration under competition. *CoRR*, abs/2007.10144, 2020.

Alexander B. Atanasov, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *CoRR*, abs/2405.00592, 2024.

Francis R. Bach. High-dimensional analysis of double descent for linear regression with random projections. *CoRR*, abs/2303.01372, 2023.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *CoRR*, abs/2102.06701, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022.

Jonathan B. Baker. Beyond Schumpeter vs. Arrow: How antitrust fosters innovation. *Antitrust Law Journal*, 74(3):575–602, 2007.

Omer Ben-Porat and Moshe Tennenholtz. Best response regression. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1499–1508, 2017.

Omer Ben-Porat and Moshe Tennenholtz. Regression equilibrium. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, pages 173–191. ACM, 2019.

Priyanjana Bengani, Jonathan Stray, and Luke Thorburn. What's right and what's wrong with optimizing for engagement. *Understanding Recommenders*, Apr 2022. URL https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851.

Benjamin Bennett, Rene M. Stulz, and Zexi Wang. Does greater public scrutiny hurt a firm's performance? Available at SSRN: https://ssrn.com/abstract=4321191, 2023.

Avrim Blum, Nika Haghtalab, Ariel D. Procaccia, and Mingda Qiao. Collaborative PAC learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2392–2401, 2017.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1024–1034. PMLR, 2020.

Blake Bordelon, Alexander B. Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *CoRR*, abs/2402.01092, 2024.

California Legislature. California senate bill no. 1047 (2023-2024). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047, 2024.

Emilio Calvano and Michele Polo. Market power, competition and innovation in digital markets: A survey. *Information Economics and Policy*, 54:100853, 2021.

Sarah Huiyi Cen, Aspen Hopkins, Andrew Ilyas, Aleksander Madry, Isabella Struckman, and Luis Videgaray Caso. AI supply chains. Available at SSRN: https://ssrn.com/abstract=4789403, 2023.

Competition and Markets Authority. AI foundation models: Technical update report. Technical report, UK Government, 2024. URL https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf.

Jodie Cook. ChatGPT, Claude, Gemini or another: The AI tool entrepreneurs prefer. *Forbes*, 2024. URL https://www.forbes.com/sites/jodiecook/2024/05/07/chatgpt-claude-gemini-or-another-the-ai-tool-entrepreneurs-prefer/.

Ian Covert, Wenlong Ji, Tatsunori Hashimoto, and James Zou. Scaling laws for the value of individual data points in machine learning. *CoRR*, abs/2405.20456, 2024.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 10131–10143, 2021.

Sarah Dean, Mihaela Curmei, Lillian J. Ratliff, Jamie Morgenstern, and Maryam Fazel. Multi-learner risk reduction under endogenous participation dynamics. *CoRR*, abs/2206.02667, 2022.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, François Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *CoRR*, abs/2402.07043, 2024.

Jinshuo Dong, Hadi Elzayn, Shahin Jabbari, Michael J. Kearns, and Zachary Schutzman. Equilibrium characterization for data acquisition games. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 252–258. ijcai.org, 2019.

European Union. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and Amending Directive 2000/31/EC (Digital Services Act). Official Journal of the European Union, 2022a. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065.

European Union. Regulation (EU) 2022/1925 of the European parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), 2022b. URL https://eur-lex.europa.eu/eli/reg/2022/1925/oj.

Alireza Fallah and Michael I. Jordan. Contract design with safety inspections. *CoRR*, abs/2311.02537, 2023.

Alireza Fallah, Michael I. Jordan, Ali Makhdoumi, and Azarakhsh Malekian. On three-layer data markets. *CoRR*, abs/2402.09697, 2024.

Yiding Feng, Ronen Gradwohl, Jason D. Hartline, Aleck C. Johnsen, and Denis Nekipelov. Bias-variance games. *CoRR*, abs/1909.03618, 2019.

Michal S Gal and Oshrit Aviv. The competitive effects of the gdpr. *Journal of Competition Law & Economics*, 16(3):349–391, 05 2020.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR, 2023.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *CoRR*, abs/2404.01413, 2024.

Tony Ginart, Eva Zhang, Yongchan Kwon, and James Zou. Competing AI: how does competition feedback affect machine learning? In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 1693–1701, 2021.

Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, and J. Zico Kolter. Scaling laws for data filtering—Data curation cannot be compute agnostic. *CoRR*, abs/2404.07177, 2024.

Ronen Gradwohl and Moshe Tennenholtz. Coopetition against an Amazon. *J. Artif. Intell. Res.*, 76: 1077–1116, 2023.

Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875 – 930, 2007.

Nika Haghtalab, Michael I. Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. In *Advances in Neural Information Processing Systems 35: Annual*

Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022.

Tinashe Handina and Eric Mazumdar. Rethinking scaling laws for learning in strategic environments. *CoRR*, abs/2402.07588, 2024.

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Winston Harrington. Enforcement leverage when penalties are restricted. *Journal of Public Economics*, 37(1):29–53, 1988. ISSN 0047-2727.

Md Rajibul Hasan, Ashish Kumar Jha, and Yi Liu. Excessive use of online video streaming services: Impact of recommender system use, psychological factors, and motives. *Computers in Human Behavior*, 80:220–228, 2018.

Tatsunori Hashimoto. Model performance scaling with multiple data sources. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4107–4116. PMLR, 2021.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 10–15 Jul 2018.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *CoRR*, abs/1903.08560, 2019.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022.

Nicole Immorlica, Adam Tauman Kalai, Brendan Lucier, Ankur Moitra, Andrew Postlewaite, and Moshe Tennenholtz. Dueling algorithms. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 215–224, 2011.

David Ingram. Fewer people using Elon Musk's X as platform struggles to keep users. *NBC News*, 2024. URL https://www.nbcnews.com/tech/tech-news/fewer-people-using-elon-musks-x-struggles-keep-users-rcna144115.

Ganesh Iyer and T. Tony Ke. Competitive algorithmic targeting and model selection. Available at SSRN: https://ssrn.com/abstract=4214973, 2022.

Meena Jagadeesan, Michael I. Jordan, and Nika Haghtalab. Competition, alignment, and equilibria in digital marketplaces. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023,* *Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 5689–5696. AAAI Press, 2023a.

Meena Jagadeesan, Michael I. Jordan, Jacob Steinhardt, and Nika Haghtalab. Improved Bayes risk can yield reduced social welfare under competition. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b.

Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. *CoRR*, abs/2402.04376, 2024.

Bruno Jullien and Wilfried Sand-Zantman. The economics of platforms: A theory guide for competition policy. *Information Economics and Policy*, 54:100880, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

Jon M. Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proc. Natl. Acad. Sci. USA*, 118(22):e2018340118, 2021.

Yongchan Kwon, Tony Ginart, and James Zou. Competition over data: how does data purchase affect users? *Trans. Mach. Learn. Res.*, 2022.

Benjamin Laufer, Jon M. Kleinberg, and Hoda Heidari. Fine-tuning games: Bargaining and adaptation for general-purpose models. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 66–76. ACM, 2024.

Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data. *CoRR*, abs/2406.08466, 2024.

Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Neil Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. *CoRR*, abs/2404.00522, 2024.

V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, apr 1967.

Celestine Mendler-Dünner, Gabriele Carovano, and Moritz Hardt. An engine not a camera: Measuring performative power of online search. *CoRR*, abs/2405.19073, 2024.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR, 2019.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Pratik Patil, Jin-Hong Du, and Ryan J. Tibshirani. Optimal ridge regularization for out-of-distribution prediction. *CoRR*, abs/2404.01233, 2024.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 2020.

Billy Perrigo. The new AI-powered Bing is threatening users. that's no laughing matter. *Time Magazine*, 2023. URL https://time.com/6256529/bing-openai-chatgpt-danger-alignment/.

Jens Prüfer and Christoph Schottmüller. Competing with big data. *The Journal of Industrial Economics*, 69(4):967–1008, 2021.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7909–7919. PMLR, 2020.

Steve Rathje, Jay J. Van Bavel, and Sander van der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26):e2024292118, 2021.

Esther Rolf, Theodora T. Worledge, Benjamin Recht, and Michael I. Jordan. Representation matters: Assessing the importance of subgroup allocations in training data. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9040–9051. PMLR, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Ilya Segal and Michael D. Whinston. Antitrust in innovative industries. *American Economic Review*, 97(5):1703–1730, December 2007.

Eliot Shekhtman and Sarah Dean. Strategic usage in a multi-learner setting. In *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pages 2665–2673. PMLR, 2024.

Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y Chen. The data addition dilemma. *arXiv preprint arXiv:2408.04154*, 2024.

Yanke Song, Sohom Bhattacharya, and Pragya Sur. Generalization error of min-norm interpolators in transfer learning. *CoRR*, abs/2406.13944, 2024.

Statista. Leading countries based on facebook audience size as of january 2024, 2024. URL https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/#:~:text=With%20around%202.9%20billion%20monthly,most%20popular%20social%20media%20worldwide.

Stigler Committee. Final report: Stigler committee on digital platforms. available at https://www.chicagobooth.edu/-/media/research/stigler/pdfs/digital-platforms---committee-report---stigler-center.pdf,, September 2019.

Jinyan Su and Sarah Dean. Learning from streaming data when users choose. *CoRR*, abs/2406.01481, 2024.

The White House. Executive order on the safe, secure, and trustworthy development and use of Artificial Intelligence, 2023.

Jean Tirole. *The Theory of Industrial Organization*, volume 1 of *MIT Press Books*. The MIT Press, December 1988.

Jai Vipra and Anton Korinek. Market concentration implications of foundation models. *CoRR*, abs/2311.01550, 2023.

Alexander Wei. *Learning and Decision-Making in Complex Environments*. PhD thesis, EECS Department, University of California, Berkeley, May 2024.

Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23549–23588. PMLR, 2022.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2):12:1–12:19, 2019.

A Proofs for Section 3

In this section, we prove Theorem 1. First, we state relevant facts (Appendix A.1) and prove intermediate lemmas (Appendix A.2), and then we use these ingredients to prove Theorem 1 (Appendix A.3). Throughout this section, we let

$$L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)\Sigma(\beta_1 - \beta_2)^T].$$

Moreover, let

$$\beta(\alpha, \lambda) = \text{argmin}_\beta \left(\alpha \cdot \mathbb{E}_{X \sim \mathcal{D}_F}[(\langle \beta - \beta_1, X \rangle)^2] + (1-\alpha) \cdot \mathbb{E}_{X \sim \mathcal{D}_F}[(\langle \beta - \beta_2, X \rangle)^2] + \lambda \|\beta\|_2^2\right)$$

be the infinite-data ridge regression predictor.

A.1 Facts

We can explicitly solve for the infinite-data ridge regression predictor

$$\begin{aligned}
\beta(\alpha, \lambda) &= \text{argmin}_\beta \left(\alpha \cdot \mathbb{E}_{x \sim \mathcal{D}_F}[\langle \beta - \beta_1, x \rangle^2] + (1-\alpha) \cdot \mathbb{E}_{x \sim \mathcal{D}_F}[\langle \beta - \beta_2, x \rangle^2] + \lambda ||\beta||_2^2\right) \\
&= \Sigma(\Sigma + \lambda I)^{-1}(\alpha\beta_1 + (1-\alpha)\beta_2).
\end{aligned}$$

A simple calculation shows that $\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, 0))] = (1-\alpha)^2 L^*(\rho)$ and $\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha, 0))] = \alpha^2 L^*(\rho)$. Thus, it holds that:

$$\alpha \mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, 0))] + (1-\alpha)\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha, 0))] = \alpha(1-\alpha)L^*(\rho).$$

Moreover, by the definition of the ridge regression objective, we see that:

$$\alpha \mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, \lambda))] + (1-\alpha)\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha, \lambda))] \geq \alpha \mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, 0))] + (1-\alpha)\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha, 0))].$$

A.2 Lemmas

The first lemma upper bounds the performance loss when there is regularization.

Lemma 11. *Suppose that power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$ and suppose that $P = \infty$. Let $L^*(\rho) = (\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)^T$. Let*

$$\beta(\alpha, \lambda) = \text{argmin}_\beta \left(\alpha \cdot \mathbb{E}_{X \sim \mathcal{D}_F}[(\langle \beta - \beta_1, X \rangle)^2] + (1-\alpha) \cdot \mathbb{E}_{X \sim \mathcal{D}_F}[(\langle \beta - \beta_2, X \rangle)^2] + \lambda \|\beta\|_2^2\right)$$

be the infinite-data ridge regression predictor. Assume that $\alpha \geq 1/2$. Then it holds that

$$\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, \lambda))] \geq (1-\alpha)^2 L^*(\rho)$$

and

$$\frac{\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, \lambda))]}{\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha, \lambda))]} \geq \frac{(1-\alpha)^2}{\alpha^2}.$$

Proof. We define the quantities:

$$\begin{aligned}
A &:= \lambda^2 \sum_{i=1}^{P} \frac{\lambda_i}{(\lambda_i+\lambda)^2} i^{-\delta} \\
B &:= (1-\alpha)^2(1-\rho)^2 \sum_{i=1}^{P} \frac{\lambda_i^3}{(\lambda_i+\lambda)^2} i^{-\delta} \\
C &:= \lambda(1-\rho) \sum_{i=1}^{P} \frac{\lambda_i^2}{(\lambda_i+\lambda)^2} i^{-\delta}.
\end{aligned}$$

We compute the performance loss as follows:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha,\lambda))] \\
&= \mathbb{E}_{\mathcal{D}_W}[\mathrm{Tr}(\Sigma(\beta_1-\beta(\alpha,\lambda))(\beta_1-\beta(\alpha,\lambda))^T)] \\
&= \mathbb{E}_{\mathcal{D}_W}\left[\mathrm{Tr}\left((\Sigma+\lambda I)^{-2}\Sigma\left(\lambda\beta_1+\Sigma\cdot(1-\alpha)(\beta_1-\beta_2)\right)\left(\lambda\beta_1+\Sigma\cdot(1-\alpha)(\beta_1-\beta_2)\right)^T\right)\right] \\
&= \mathbb{E}_{\mathcal{D}_W}\left[\mathrm{Tr}\left((\Sigma+\lambda I)^{-2}\Sigma\cdot\left(\lambda\beta_1+\Sigma\cdot(1-\alpha)(\beta_1-\beta_2)\right)\left(\lambda\beta_1+\Sigma\cdot(1-\alpha)(\beta_1-\beta_2)\right)^T\right)\right] \\
&= \lambda^2\mathbb{E}_{\mathcal{D}_W}\left[\mathrm{Tr}\left((\Sigma+\lambda I)^{-2}\Sigma\cdot\beta_1\beta_1^T\right)\right] + (1-\alpha)^2\mathbb{E}_{\mathcal{D}_W}\left[\mathrm{Tr}\left((\Sigma+\lambda I)^{-2}\Sigma^3\cdot(\beta_1-\beta_2)(\beta_1-\beta_2)^T\right)\right] \\
&\quad + \lambda(1-\alpha)\mathbb{E}_{\mathcal{D}_W}\left[\mathrm{Tr}\left((\Sigma+\lambda I)^{-2}\Sigma^2\cdot\beta_1(\beta_1-\beta_2)^T\right)\right] \\
&= \lambda^2\sum_{i=1}^{P}\frac{\lambda_i}{(\lambda_i+\lambda)^2}\mathbb{E}_{\mathcal{D}_W}[\langle\beta_1,v_i\rangle^2] + (1-\alpha)^2\sum_{i=1}^{P}\frac{\lambda_i^3}{(\lambda_i+\lambda)^2}\mathbb{E}_{\mathcal{D}_W}[\langle\beta_1-\beta_2,v_i\rangle^2] \\
&\quad + \lambda(1-\alpha)\sum_{i=1}^{P}\frac{\lambda_i^2}{(\lambda_i+\lambda)^2}\mathbb{E}_{\mathcal{D}_W}\left[\langle\beta_1,v_i\rangle\langle\beta_1-\beta_2,v_i\rangle\right] \\
&= \lambda^2\sum_{i=1}^{P}\frac{\lambda_i}{(\lambda_i+\lambda)^2}i^{-\delta} + (1-\alpha)^2(1-\rho)^2\sum_{i=1}^{P}\frac{\lambda_i^3}{(\lambda_i+\lambda)^2}i^{-\delta} + \lambda(1-\alpha)(1-\rho)\sum_{i=1}^{P}\frac{\lambda_i^2}{(\lambda_i+\lambda)^2}i^{-\delta} \\
&= A + (1-\alpha)^2 B + (1-\alpha)C.
\end{aligned}$$

An analogous calculation shows that the safety violation can be written as:

$$\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha,\lambda))] = A + \alpha^2 B + \alpha C$$

Since $\alpha \geq 1/2$, then it holds that:

$$\frac{\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha,\lambda))]}{\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha,\lambda))]} = \frac{A+(1-\alpha)^2B+(1-\alpha)C}{A+\alpha B+\alpha C} \geq \frac{(1-\alpha)^2}{\alpha^2}.$$

Combining this with the facts from Appendix A.1—which imply that $\alpha\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha,\lambda))] + (1-\alpha)\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha,\lambda))] \geq \alpha(1-\alpha)L^*(\rho)$—we have that $\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha,\lambda))] \geq (1-\alpha)^2 L^*(\rho)$ as desired. □

The following lemma computes the optimal values of α and λ for the incumbent.

Lemma 12. *Suppose that power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0,1)$ and suppose that $P = \infty$. Let $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1-\beta_2)^T\Sigma(\beta_1-\beta_2)^T]$. Suppose that $N_I = \infty$, and suppose that the safety constraint τ_I satisfies (1). Then it holds that $\alpha_I = \sqrt{\frac{\min(\tau_I, L^*(\rho))}{L^*(\rho)}}$, and $\lambda_I = 0$ is optimal for the incumbent. Moreover, it holds that:*

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1,\beta_2,\mathcal{D}_F,\lambda_I,\infty,\alpha_O)] = (\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))})^2.$$

Proof. First, we apply Lemma 26 with $N = \infty$ to see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, \infty, \alpha)] = \mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, \lambda))]$$

and apply the definition of L_2^* to see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_2^*(\beta_1, \beta_2, \mathcal{D}_F, \alpha)] = \mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha, 0))].$$

Let $\alpha^* = \sqrt{\frac{\min(\tau_I, L^*(\rho))}{L^*(\rho)}}$. By the assumption in the lemma statement, we know that:

$$\alpha^* \geq \sqrt{\frac{\mathbb{E}_{\mathcal{D}_W}[L_2^*(\beta_1, \beta_2, \mathcal{D}_F, 0.5)]}{L^*(\rho)}} = 0.5.$$

We show that $(\alpha_I, \lambda_I) = (\alpha^*, 0)$. Assume for sake of contradiction that $(\alpha, \lambda) \neq (\alpha^*, 0)$ satisfies the safety constraint $\mathbb{E}_{\mathcal{D}_W}[L_2^*(\beta_1, \beta_2, \mathcal{D}_F, \alpha)] \leq \tau_I$ and achieves strictly better performance loss:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, \infty, \alpha)] < \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, 0, \infty, \alpha^*)].$$

We split into two cases: $\alpha^* = \alpha, \lambda \neq 0$ and $\alpha^* \neq \alpha$.

Case 1: $\alpha^* = \alpha$, $\lambda \neq 0$. By Lemma 11, we know that

$$\mathbb{E}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, \infty, \alpha^*)] = \mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha^*, \lambda))] \geq (1 - \alpha^*)^2 L^*(\rho).$$

Equality is obtained at $\lambda = 0$, which is a contradiction.

Case 2: $\alpha \neq \alpha^*$. By Lemma 11, it must hold that $\alpha > \alpha^*$ in order for the performance to beat that of $(\alpha^*, 0)$. However, this means that the safety constraint

$$\mathbb{E}_{\mathcal{D}_W}[L_2^*(\beta_1, \beta_2, \mathcal{D}_F, \alpha)] = \alpha^2 L^*(\rho) > (\alpha^*)^2 L^*(\rho) = \tau_I$$

is violated, which is a contradiction.

Concluding the statement. This means that$(\alpha_I, \lambda_I) = (\alpha^*, 0)$, which also means that:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_I, \infty, \alpha_I)] &= \mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha_I, \lambda_I))] \\ &= (1 - \alpha_I)^2 \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)] \\ &= \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho)}\right)^2.\end{aligned}$$

□

The following claim calculates $\mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)]$.

Claim 13. *Suppose that the power-law scaling assumption holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, suppose that $P = \infty$. Then it holds that:*

$$\mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)] = 2(1 - \rho) \left(\sum_{i=1}^{P} i^{-\delta - 1 - \gamma}\right) = \Theta(1 - \rho).$$

Proof. Let $\Sigma = V \Lambda V^T$ be the eigendecomposition of Σ, where Λ is a diagonal matrix consisting of the eigenvalues. We observe that

$$\mathbb{E}_{\mathcal{D}_W}[\langle \beta_1 - \beta_2, v_i \rangle^2] = \mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i \rangle^2] + \mathbb{E}_{\mathcal{D}_W}[\langle \beta_2, v_i \rangle^2] - 2\mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i \rangle \langle \beta_2, v_i \rangle] = i^{-\delta} + i^{-\delta} - 2\rho i^{-\delta} = 2(1-\rho) i^{-\delta}.$$

This means that:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)] &= \mathrm{Tr}(\Sigma \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)(\beta_1 - \beta_2)^T]) \\
&= \mathrm{Tr}(\Lambda \mathbb{E}_{\mathcal{D}_W}[V^T (\beta_1 - \beta_2)(\beta_1 - \beta_2)^T V]) \\
&= \sum_{i=1}^{P} i^{-1-\gamma} \mathbb{E}_{\mathcal{D}_W}[\langle \beta_1 - \beta_2, v_i \rangle^2] \\
&= 2(1-\rho) \sum_{i=1}^{P} i^{-\delta-1-\gamma} \\
&= \Theta(1-\rho).
\end{aligned}$$

□

A.3 Proof of Theorem 1

We prove Theorem 1 using the above lemmas along with Corollary 8 (the proof of which we defer to Appendix C).

Proof of Theorem 1. We analyze (α_C, λ_C) first for the incumbent $C = I$ and then for the entrant $C = E$.

Analysis of the incumbent $C = I$. To compute α_I and λ_I, we apply Lemma 12. By Lemma 12, we see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_I, \infty, \alpha_I)] = \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))} \right)^2.$$

Analysis of the entrant $C = E$. Since the entrant faces no safety constraint, the entrant can choose any $\alpha \in [0.5, 1]$. We apply Corollary 8 to see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_E, N, \alpha_E)] = \inf_{\alpha \in [0.5,1]} \inf_{\lambda > 0} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)] = \Theta\left(N^{-\nu}\right),$$

which means that:

$$N_E^*(\infty, \tau_I, \infty, \mathcal{D}_W, \mathcal{D}_F) = \Theta\left(\left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))} \right)^{-2/\nu} \right)$$

as desired. We can further apply Claim 13 to see that $L^*(\rho) = \Theta(1-\rho)$. □

B Proofs for Section 4

B.1 Proofs for Section 4.2

We prove Theorem 4. The main technical tool is Theorem 7, the proof of which we defer to Appendix C.

Proof of Theorem 4. We analyze (α_C, λ_C) first for the incumbent $C = I$ and then for the entrant $C = E$. Like in the theorem statement, let $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma(\beta_1 - \beta_2)] = \Theta(1 - \rho)$ (Claim 13) and $G_I := (\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))})^2$, and $\nu = \min(2(1+\gamma), \delta + \gamma)$.

Analysis of the incumbent $C = I$. Recall from the facts in Appendix A.1 that:

$$L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \alpha) = \alpha^2 L^*(\rho).$$

This means that the safety constraint is satisfied if and only if $\alpha_I \leq \sqrt{\frac{\min(\tau_I, L^*(\rho))}{L^*(\rho)}} =: \alpha^*$. The bound in Corollary 8 implies that:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_I, N_I, \alpha_I)] \\
&= \inf_{\alpha \in [0.5, \alpha^*]} \inf_{\lambda > 0} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N_I, \alpha)] \\
&= \Theta\left(\inf_{\lambda > 0} \mathbb{E}_{\mathcal{D}_W}\left[L_1^*\left(\beta_1, \beta_2, \Sigma, \lambda, N_I, \alpha^*\right)\right]\right) \\
&= \begin{cases}
\Theta\left(N_I^{-\nu}\right) & \text{if } N_I \leq (1-\alpha^*)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \\
\Theta\left(\left(\frac{N_I}{(1-\alpha^*)(1-\rho)}\right)^{-\frac{\nu}{\nu+1}}\right) & \text{if } (1-\alpha^*)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \leq N_I \leq (1-\alpha^*)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \\
\Theta((1-\alpha^*)^2(1-\rho)) & \text{if } N_I \geq (1-\alpha^*)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}},
\end{cases} \\
&= \begin{cases}
\Theta\left(N_I^{-\nu}\right) & \text{if } N_I \leq G_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \\
\Theta\left(N_I^{-\frac{\nu}{\nu+1}} \cdot G_I^{\frac{\nu}{2(\nu+1)}}(1-\rho)^{\frac{\nu}{2(\nu+1)}}\right) & \text{if } G_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \leq N_I \leq G_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}} \\
\Theta\left(G_I\right) & \text{if } N_I \geq G_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}}
\end{cases}.
\end{aligned}$$

Analysis of the entrant $C = E$. Since the entrant faces no safety constraint, the entrant can choose any $\alpha \in [0.5, 1]$. We apply Corollary 7 to see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_E, N, \alpha_E)] = \inf_{\alpha \in [0.5, 1]} \inf_{\lambda > 0} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)] = \Theta\left(N^{-\nu}\right),$$

which means that:

$$N_E^*(N_I, \tau_I, \infty, \mathcal{D}_W, \mathcal{D}_F) = \begin{cases}
\Theta\left(N_I\right) & \text{if } N_I \leq G_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \\
\Theta\left(N_I^{\frac{1}{\nu+1}} \cdot G_I^{-\frac{1}{2(\nu+1)}}(1-\rho)^{-\frac{1}{2(\nu+1)}}\right) & \text{if } G_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \leq N_I \leq G_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}} \\
\Theta\left(G_I^{-\frac{1}{\nu}}\right) & \text{if } N_I \geq G_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}}.
\end{cases}$$

as desired.

□

B.2 Proofs for Section 4.3

We prove Theorem 5. When the the safety constraints of the two firms are sufficiently close, it no longer suffices to analyze the loss up to constants for the entrant, and we require a more fine-grained analysis of the error terms than is provided in the scaling laws in Corollary 8. In this case, we turn to scaling laws for the *excess loss* as given by Corollary 10.

Proof of Theorem 5. We analyze (α_C, λ_C) first for the incumbent $C = I$ and then for the entrant $C = E$. Like in the theorem statement, let $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)] = \Theta(1-\rho)$ (Claim 13), $G_I = (\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))})^2$, $G_E = (\sqrt{L^*(\rho)} - \sqrt{\min(\tau_E, L^*(\rho))})^2$, $D = G_I - G_E$, and $\nu = \min(2(1+\gamma), \delta + \gamma)$.

Analysis of the incumbent $C = I$. Since the incumbent has infinite data, we apply Lemma 12 to see that:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_I, \infty, \alpha_I)] &= \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^2 \\ &= D + G_E.\end{aligned}$$

Analysis of the entrant $C = E$. Recall from the facts in Appendix A.1 that:

$$L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \alpha) = \alpha^2 L^*(\rho).$$

This means that the safety constraint is satisfied if and only if $\alpha_E \leq \sqrt{\frac{\min(\tau_E, L^*(\rho))}{L^*(\rho)}} =: \alpha^*$. The bound in Corollary 10 implies that:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_E, N, \alpha_E)] \\
&= \inf_{\alpha \in [0.5, \alpha^*]} \inf_{\lambda > 0} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)] \\
&= \inf_{\alpha \in [0.5, \alpha^*]} \left(\inf_{\lambda > 0} \left(\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha) - L_1(\beta(\alpha, 0))] \right) + \mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, 0))] \right) \\
&= \inf_{\alpha \in [0.5, \alpha^*]} \left(\inf_{\lambda > 0} \left(\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha) - L_1(\beta(\alpha, 0))] \right) + (1-\alpha)^2 L^*(\rho) \right) \\
&= \Theta\left(\inf_{\lambda > 0} \left(\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha) - L_1(\beta(\alpha^*, 0))] \right) \right) + (1-\alpha^*)^2 L^*(\rho) \\
&= \begin{cases} (1-\alpha^*)^2 L^*(\rho) + \Theta\left(N^{-\nu}\right) & \text{if } N \leq (1-\alpha^*)^{-\frac{1}{\nu}} (1-\rho)^{-\frac{1}{\nu}} \\ (1-\alpha^*)^2 L^*(\rho) + \Theta\left(\left(\frac{N}{(1-\alpha^*)(1-\rho)} \right)^{-\frac{\nu}{\nu+1}} \right) & \text{if } (1-\alpha^*)^{-\frac{1}{\nu}} (1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha^*)^{-\frac{\nu'+1}{\nu-\nu'}} (1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}} \\ (1-\alpha^*)^2 L^*(\rho) + \Theta\left((1-\alpha^*)(1-\rho) N^{-\frac{\nu'}{\nu'+1}} \right) & \text{if } N \geq (1-\alpha^*)^{-\frac{\nu'+1}{\nu-\nu'}} (1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}, \end{cases} \\
&= \begin{cases} G_E + \Theta\left(N^{-\nu}\right) & \text{if } N \leq (1-\alpha^*)^{-\frac{1}{\nu}} (1-\rho)^{-\frac{1}{\nu}} \\ G_E + \Theta\left(\left(\frac{N}{(1-\alpha^*)(1-\rho)} \right)^{-\frac{\nu}{\nu+1}} \right) & \text{if } (1-\alpha^*)^{-\frac{1}{\nu}} (1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha^*)^{-\frac{\nu'+1}{\nu-\nu'}} (1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}} \\ G_E + \Theta\left((1-\alpha^*)(1-\rho) N^{-\frac{\nu'}{\nu'+1}} \right) & \text{if } N \geq (1-\alpha^*)^{-\frac{\nu'+1}{\nu-\nu'}} (1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}, \end{cases} .
\end{aligned}$$

Using this, we can compute the market-entry threshold as follows:

$$
\begin{aligned}
& N_E^*(\infty, \tau_I, \tau_E, \mathcal{D}_W, \mathcal{D}_F) \\
&= \begin{cases} \Theta(D^{-\frac{1}{\nu}}) & \text{if } D \geq (1-\alpha^*)(1-\rho) \\ \Theta\left(D^{-\frac{\nu+1}{\nu}}(1-\alpha^*)(1-\rho)\right) & \text{if } (1-\alpha^*)^{\frac{\nu}{\nu-\nu'}}(1-\rho)^{\frac{\nu}{\nu-\nu'}} \leq D \leq (1-\alpha^*)(1-\rho) \\ \Theta\left(\left(\frac{D}{(1-\alpha^*)(1-\rho)}\right)^{-\frac{\nu'+1}{\nu'}}\right) & \text{if } D \leq (1-\alpha^*)^{\frac{\nu}{\nu-\nu'}}(1-\rho)^{\frac{\nu}{\nu-\nu'}} \end{cases} \\
&= \begin{cases} \Theta(D^{-\frac{1}{\nu}}) & \text{if } D \geq G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}} \\ \Theta\left(D^{-\frac{\nu+1}{\nu}} G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}}\right) & \text{if } G_E^{\frac{\nu}{2(\nu-\nu')}}(1-\rho)^{\frac{\nu}{2(\nu-\nu')}} \leq D \leq G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}} \\ \Theta\left(\left(\frac{D}{G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}}}\right)^{-\frac{\nu'+1}{\nu'}}\right) & \text{if } D \leq G_E^{\frac{\nu}{2(\nu-\nu')}}(1-\rho)^{\frac{\nu}{2(\nu-\nu')}} \end{cases}
\end{aligned}
$$

□

C Proofs for Section 5

In this section, we derive a deterministic equivalent and scaling laws for high-dimensional multi-objective linear regression. Before diving into this, we introduce notation, derive a basic decomposition, and give an outline for the remainder of the section.

Notation. Recall that (X_i, Y_i) denotes the labelled training dataset. Let the sample covariance be:

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} X_i X_i^T.$$

We also consider the following reparameterization where we group together inputs according to how they are labelled. For $j \in \{1, 2\}$, we let $X_{1,j}, \ldots, X_{N_j,j}$ be the inputs labelled by β_j. We let

$$\hat{\Sigma}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} X_{i,1} X_{i,1}^T$$

$$\hat{\Sigma}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} X_{i,2} X_{i,2}^T.$$

It is easy to see that $\Sigma = \alpha\hat{\Sigma}_1 + (1-\alpha)\hat{\Sigma}_2$. Moreover, $\mathbb{E}[\hat{\Sigma}] = \mathbb{E}[\hat{\Sigma}_1] = \mathbb{E}[\hat{\Sigma}_2] = \Sigma$. Furthermore, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are fully independent. We let $\sim$ denote asymptotic equivalence following Bach [2023].

Basic decomposition. A simple calculation shows that the solution and population-level loss of ridge regression takes the following form.

Claim 14. *Assume the notation above. Let $B^{\texttt{sn}} = \beta_1\beta_1^T$, let $B^{\texttt{df}} = (\beta_1 - \beta_2)(\beta_1 - \beta_2)^T$, and let $B^{\texttt{mx}} = (\beta_1 - \beta_2)\beta_1^T$. The learned predictor takes the form:*

$$\hat{\beta}(\alpha, \lambda, X) = (\hat{\Sigma} + \lambda I)^{-1}(\alpha\hat{\Sigma}_1\beta_1 + (1-\alpha)\hat{\Sigma}_2\beta_2).$$

Moreover, it holds that:

$$L_1(\hat{\beta}(\alpha,\lambda,X)) = \underbrace{\lambda^2 \operatorname{Tr}((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}B^{sn})}_{(T1)} + \underbrace{(1-\alpha)^2 \operatorname{Tr}(\hat{\Sigma}_2(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{df})}_{(T2)}$$
$$+ \underbrace{2\lambda(1-\alpha)\cdot \operatorname{Tr}((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{mx})}_{(T3)}.$$

Proof. For $1 \leq i \leq N$, let Y_i be the label for input X_i in the training dataset. For $i \in \{1,2\}$ and $1 \leq i \leq N_i$, let $Y_{i,j} := \langle \beta_i, X_{i,j}\rangle$ be the label for the input $X_{i,j}$ according to β_i.

For the first part, it follows from standard analyses of ridge regression that the learned predictor takes the form:

$$\begin{aligned}
\hat{\beta}(\alpha,\lambda,X) &= (\hat{\Sigma}+\lambda I)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right) \\
&= (\hat{\Sigma}+\lambda I)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_{i,1} Y_{i,1} + \frac{1}{N}\sum_{i=1}^{N} X_{i,2} Y_{i,2}\right) \\
&= (\hat{\Sigma}+\lambda I)^{-1}\left(\alpha\hat{\Sigma}_1\beta_1 + (1-\alpha)\hat{\Sigma}_2\beta_2\right)
\end{aligned}$$

as desired.

For the second part, we first observe that the difference $\beta_1 - \hat{\beta}(\alpha,\lambda,X)$ takes the form:

$$\begin{aligned}
\beta_1 - \hat{\beta}(\alpha,\lambda,X) &= \beta_1 - (\hat{\Sigma}+\lambda I)^{-1}\left(\alpha\hat{\Sigma}_1\beta_1 + (1-\alpha)\hat{\Sigma}_2\beta_2\right) \\
&= (\hat{\Sigma}+\lambda I)^{-1}\left(\lambda\beta_1 + (1-\alpha)\hat{\Sigma}_2(\beta_1-\beta_2)\right).
\end{aligned}$$

This means that:

$$\begin{aligned}
&L_1(\hat{\beta}(\alpha,\lambda,X)) \\
&= (\beta_1 - \hat{\beta}(\alpha,\lambda,X))^T\Sigma(\beta_1 - \hat{\beta}(\alpha,\lambda,X)) \\
&= \left(\lambda\beta_1 + (1-\alpha)\hat{\Sigma}_2(\beta_1-\beta_2)\right)^T(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\left(\lambda\beta_1 + (1-\alpha)\hat{\Sigma}_2(\beta_1-\beta_2)\right) \\
&= \lambda^2\cdot\beta_1^T\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\beta_1 + (1-\alpha)^2\cdot(\beta_1-\beta_2)^T\hat{\Sigma}_2\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2(\beta_1-\beta_2) \\
&\quad + 2\lambda(1-\alpha)\cdot\beta_1^T\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2(\beta_1-\beta_2) \\
&= \lambda^2\operatorname{Tr}((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}B^{\mathtt{sn}}) + (1-\alpha)^2\operatorname{Tr}(\hat{\Sigma}_2(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2B^{\mathtt{df}}) \\
&\quad + 2\lambda(1-\alpha)\cdot\operatorname{Tr}((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2B^{\mathtt{mx}}).
\end{aligned}$$

as desired. □

Outline for the rest of this Appendix. The bulk of our analysis in this section boils down to analyzing Term 1 (T1), Term 2 (T2), and Term 3 (T3) in Claim 14. Our main technical tool is the random matrix machinery from Appendix D. In Appendix C.1, we provide useful sublemmas about intermediate deterministic equivalents that we apply to analyze Terms 2 and 3. We then analyze Term 1 (Appendix C.2), Term 2 (Appendix C.3), and Term 3 (Appendix C.4), and use this to prove Lemma 6 (Appendix C.5).

We apply the power scaling assumptions to derive a simpler expression for the deterministic equivalent (Lemma 26 in Appendix C.6). We then apply Lemma 26 to prove Theorem 7 (Appendix

C.7), and we prove Corollary 8 (Appendix C.8). We also apply Lemma 26 to prove Theorem 9 (Appendix C.9), and we prove Corollary 10 (Appendix C.10). We defer auxiliary calculations to Appendix C.11.

C.1 Useful lemmas about intermediate deterministic equivalents

The results in this section consider $Z_1 := \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I$, which we introduce when conditioning on the randomness of $\hat{\Sigma}_1$ when analyzing (T2) and (T3). We derive several properties of Z_1 and the effective regularizer $\kappa_1 = \kappa(1, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$ below.

The first set of lemmas relate the trace of various matrices involving κ_1 and Z_1 to deterministic quantities. A subtlety is that κ_1 and Z_1 are correlated, so we cannot directly apply Marčenko-Pastur, and instead we must indirectly analyze this quantity.

Lemma 15. *Consider the setup of Lemma 6, and assume the notation above. Assume $\alpha < 1$. Let $Z_1 = \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I$, and let $\kappa_1 = \kappa(1, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$. Suppose that B has bounded operator norm.*

$$\kappa_1 \operatorname{Tr}\left((\Sigma + \kappa_1 Z_1)^{-1} B\right) \sim \frac{(1-\alpha)\kappa}{\lambda} \operatorname{Tr}\left((\Sigma + \kappa I)^{-1} B\right)$$

Proof. By Claim 19, we know that:

$$\begin{aligned}(1-\alpha)\operatorname{Tr}\left(\left(\hat{\Sigma} + \lambda I\right)^{-1} B\right) &= \operatorname{Tr}\left(\left(\hat{\Sigma}_2 + Z_1\right)^{-1} B\right) \\ &\sim_{(A)} \kappa_1 \operatorname{Tr}\left((\Sigma + \kappa_1 I)^{-1} B\right).\end{aligned}$$

where (A) applies Lemma 36 and Claim 20.

Furthermore, by Lemma 36, it holds that:

$$\lambda \operatorname{Tr}\left(\left(\hat{\Sigma} + \lambda I\right)^{-1} B\right) \sim \kappa \operatorname{Tr}\left((\Sigma + \kappa I)^{-1} B\right).$$

Putting this all together yields the desired result. □

Lemma 16. *Consider the setup of Lemma 6, and assume the notation above. Assume $\alpha < 1$. Let $Z_1 = \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I$, and let $\kappa_1 = \kappa(1, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$. Suppose that A and B have bounded operator norm. Then it holds that:*

$$(\kappa_1)^2\left(\operatorname{Tr}\left((\Sigma + \kappa_1 Z_1)^{-1} A (\Sigma + \kappa_1 Z_1)^{-1} B\right) + E_1\right) \sim \frac{(1-\alpha)^2\kappa^2}{\lambda^2}\left(\operatorname{Tr}\left((\Sigma + \kappa I)^{-1} A (\Sigma + \lambda I)^{-1} B\right) + E_2\right)$$

where

$$\begin{aligned}\kappa &= \kappa(\lambda, N, \Sigma) \\ E_1 &= \frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}(A(\Sigma + \kappa_1 Z_1)^{-1}\Sigma(\Sigma + \kappa_1 Z_1)^{-1})}{1 - \frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma + \kappa_1 Z_1)^{-1})\Sigma(\Sigma + \kappa_1 Z_1)^{-1}\Sigma} \cdot \operatorname{Tr}\left((\Sigma + \kappa_1 Z_1)^{-1}\Sigma(\Sigma + \kappa_1 Z_1)^{-1} B\right) \\ E_2 &= \frac{\frac{1}{N}\operatorname{Tr}(A\Sigma(\Sigma + \kappa I)^{-2})}{1 - \frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma + \kappa I)^{-2})} \cdot \operatorname{Tr}\left((\Sigma + \kappa I)^{-1}\Sigma(\Sigma + \kappa I)^{-1} B\right)\end{aligned}$$

Proof. By Claim 19, we know that:

$$(1-\alpha)^2 \operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda I\right)^{-1} A \left(\hat{\Sigma}+\lambda I\right)^{-1} B\right) = \operatorname{Tr}\left(\left(\hat{\Sigma}_2+Z_1\right)^{-1} A \left(\hat{\Sigma}_2+Z_1\right)^{-1} B\right)$$
$$\sim_{(A)} \kappa_1^2 \left(\operatorname{Tr}\left(\left(\Sigma+\kappa_1 Z_1\right)^{-1} A \left(\Sigma+\kappa_1 Z_1\right)^{-1} B\right) + E_1\right).$$

where (A) applies Lemma 36 and Claim 20.

Furthermore, by Lemma 36, it holds that:

$$\lambda^2 \operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda I\right)^{-1} A \left(\hat{\Sigma}+\lambda I\right)^{-1} B\right) \sim \kappa^2 \left(\operatorname{Tr}\left(\left(\Sigma+\kappa I\right)^{-1} A \left(\Sigma+\kappa I\right)^{-1} B\right) + E_2\right).$$

Putting this all together yields the desired result.

□

Lemma 17. *Consider the setup of Lemma 6, and assume the notation above. Assume $\alpha < 1$. Let $Z_1 = \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha} I$, and let $\kappa_1 = \kappa(1, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$. Then it holds that:*

$$\kappa_1^2 \frac{\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)} \sim \frac{(1-\alpha)^2\kappa^2}{\lambda^2}\frac{\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}$$

Proof. By Claim 19, we know that:

$$(1-\alpha)^2 \operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda I\right)^{-1} \Sigma \left(\hat{\Sigma}+\lambda I\right)^{-1} \Sigma\right)$$
$$= \operatorname{Tr}\left(\left(\hat{\Sigma}_2+Z_1\right)^{-1} \Sigma \left(\hat{\Sigma}_2+Z_1\right)^{-1} \Sigma\right)$$
$$\sim_{(A)} \kappa_1^2 \left(1+\frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\right)\operatorname{Tr}\left(\left(\Sigma+\kappa_1 Z_1\right)^{-1}\Sigma\left(\Sigma+\kappa_1 Z_1\right)^{-1}\Sigma\right)$$
$$= \kappa_1^2 \frac{\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}$$

where (A) applies Lemma 36 and Claim 20.

Furthermore, by Lemma 36, it holds that:

$$\lambda^2 \operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda I\right)^{-1} \Sigma \left(\hat{\Sigma}+\lambda I\right)^{-1} \Sigma\right)$$
$$\sim_{(A)} \kappa^2 \left(1+\frac{\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2}}\right)\operatorname{Tr}\left(\left(\Sigma+\kappa I\right)^{-1}\Sigma\left(\Sigma+\kappa I\right)^{-1}\Sigma\right)$$
$$= \kappa^2 \left(\frac{\operatorname{Tr}\left(\Sigma^2\left(\Sigma+\kappa I\right)^{-2}\right)}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2}}\right).$$

where (A) applies Lemma 36.

Putting this all together yields the desired result. □

Next, we relate the random effective regularizer κ_1 to the deterministic effective regularizer $\kappa(\lambda, N, \Sigma)$.

Lemma 18. *Consider the setup of Lemma 6, and assume the notation above. Assume $\alpha < 1$. Let $Z_1 = \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I$, and let $\kappa_1 = \kappa(1, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$. Let $\kappa = \kappa(\lambda, N, \Sigma)$. Then, it holds that $\lambda\kappa_1 \sim \kappa$.*

Proof. Recall that $\kappa_1 = \kappa(1, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$ is the unique value such that:

$$\frac{1}{\kappa_1} + \frac{1}{N(1-\alpha)}\operatorname{Tr}((Z_1^{-1/2}\Sigma Z_1^{-1/2} + \kappa_1 I)^{-1} Z_1^{-1/2}\Sigma Z_1^{-1/2}) = 1.$$

We can write this as:

$$1 + \frac{\kappa_1}{N(1-\alpha)}\operatorname{Tr}((\Sigma + \kappa_1 Z_1)^{-1}\Sigma) = \kappa_1.$$

Now we apply Lemma 15 to see that:

$$\kappa_1 = 1 + \frac{\kappa_1}{N(1-\alpha)}\operatorname{Tr}((\Sigma + \kappa_1 Z_1)^{-1}\Sigma) \sim 1 + \frac{1}{N(1-\alpha)}\frac{(1-\alpha)\kappa}{\lambda}\operatorname{Tr}((\Sigma + \kappa I)^{-1}\Sigma).$$

We can write this to see that:

$$\kappa_1 \sim \frac{\kappa}{\lambda}\left(\frac{\lambda}{\kappa} + \frac{1}{N}\operatorname{Tr}((\Sigma + \kappa I)^{-1}\Sigma)\right) = \frac{\kappa}{\lambda}.$$

This implies that $\lambda\kappa_1 \sim \kappa$ as desired.

□

The proofs of these results relied on the following facts.

Claim 19. *Consider the setup of Lemma 6, and assume the notation above. Assume $\alpha < 1$. Let $Z_1 = \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I$. Then it holds that:*

$$(\hat{\Sigma} + \lambda I)^{-1} = (1-\alpha)^{-1}(\hat{\Sigma}_2 + Z_1)^{-1}.$$

Proof. We observe that:

$$\begin{aligned}(1-\alpha)(\hat{\Sigma} + \lambda I)^{-1} &= (1-\alpha)(\alpha\hat{\Sigma}_1 + (1-\alpha)\hat{\Sigma}_2 + \lambda I)^{-1} \\ &= (1-\alpha)(1-\alpha)^{-1}\left(\hat{\Sigma}_2 + \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I\right)^{-1} \\ &= \left(\hat{\Sigma}_2 + Z_1\right)^{-1},\end{aligned}$$

where $Z_1 = \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I$. □

Claim 20. *Consider the setup of Lemma 6, and assume the notation above. Assume $\alpha < 1$. Let $Z_1 = \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I$. Then it holds that Z_1 and Z_1^{-1} both have bounded operator norm.*

Proof. Since $\hat{\Sigma}_1$ is PSD, we observe that:

$$\|Z_1\|_{op} = \frac{\alpha}{1-\alpha}\|\hat{\Sigma}_1\|_{op} + \frac{\lambda}{1-\alpha}.$$

The fact that $\|\hat{\Sigma}_1\|_{op}$ is bounded follows from the boundedness requirements from Assumption 1. This proves that $\|Z_1\|_{op}$ is bounded.

To see that $\|Z_1^{-1}\|$ is also bounded, note that:

$$\|Z_1^{-1}\|_{op} \geq \frac{1-\alpha}{\lambda}$$

□

C.2 Analysis of Term 1 (T1)

We show the following deterministic equivalent for term 1. This analysis is identical to the analysis of the deterministic equivalent for single-objective linear regression [Bach, 2023, Wei et al., 2022], and we include it for completeness.

Lemma 21. *Consider the setup of Lemma 6, and assume the notation above. Then it holds that:*

$$\lambda^2 \operatorname{Tr}((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}B^{sn}) \sim \frac{\kappa^2}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\cdot \operatorname{Tr}(\Sigma(\Sigma+\kappa I)^{-2}B^{sn})$$

Proof. We apply Lemma 36 to see that:

$$\begin{aligned}
&\lambda^2 \operatorname{Tr}((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}B^{\mathsf{sn}})\\
&\sim \kappa^2 \operatorname{Tr}((\Sigma+\kappa I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}B^{\mathsf{sn}}) + \frac{\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\\
&= \frac{\kappa^2}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\cdot \operatorname{Tr}(\Sigma(\Sigma+\kappa I)^{-2}B^{\mathsf{sn}}),
\end{aligned}$$

as desired. □

C.3 Analysis of Term 2 (T2)

We show the following deterministic equivalent for term 2.

Lemma 22. *Consider the setup of Lemma 6, and assume the notation above. Then it holds that:*

$$\begin{aligned}
&(1-\alpha)^2 \operatorname{Tr}\left(\hat{\Sigma}_2(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{df}\right)\\
&\sim \frac{(1-\alpha)^2}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\left(\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma\,(\Sigma+\kappa I)^{-1}\Sigma B^{df}\Sigma\right)\right)\\
&+\frac{(1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\cdot\left(\operatorname{Tr}\left(\Sigma B^{df}\right)-2(1-\alpha)\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma B^{df}\Sigma\right)\right)
\end{aligned}$$

The key idea of the proof is to unwrap the randomness in layers. First, we condition on $\hat{\Sigma}_1$ and replace the randomness $\hat{\Sigma}_2$ with a deterministic equivalent where the effective regularizer κ_1 depends on $\hat{\Sigma}_1$ (Lemma 23). At this stage, we unfortunately cannot directly deal with the randomness $\hat{\Sigma}_1$ with deterministic equivalence due to the presence of terms κ_1 which depend on $\hat{\Sigma}_1$, and we instead apply the sublemmas from the previous section.

The following lemma replaces the randomness $\hat{\Sigma}_2$ with a deterministic equivalent.

Lemma 23. *Consider the setup of Lemma 6, and assume the notation above. Assume that $\alpha < 1$. Let $Z_1 = \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I$, and let $\kappa_1 = \kappa(1, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$. Then it holds that:*

$$\begin{aligned}
&(1-\alpha)^2 \operatorname{Tr}\left(\hat{\Sigma}_2(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{df}\right)\\
&\sim \frac{\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma\,(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{df}\Sigma\right)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\\
&+\frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\cdot\left(\operatorname{Tr}\left(\Sigma B^{df}\right)-2\operatorname{Tr}\left(\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{df}\right)\right).
\end{aligned}$$

Proof. By Claim 19 we have that:

$$
\begin{aligned}
(1-\alpha)^2 \operatorname{Tr}\left(\hat{\Sigma}_2(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{\mathtt{df}}\right) &= \operatorname{Tr}\left(\hat{\Sigma}_2\left(\hat{\Sigma}_2+Z_1\right)^{-1}\Sigma\left(\hat{\Sigma}_2+Z_1\right)^{-1}\hat{\Sigma}_2 B^{\mathtt{df}}\right)\\
&\sim_{(A)} \operatorname{Tr}\left(\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{df}}\right)+E\\
&= \operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)+E
\end{aligned}
$$

where (A) follows from Lemma 39 and Claim 20, and E is defined such that

$$
E := \frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}(\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\cdot(\kappa_1)^2\operatorname{Tr}\left(Z_1\left(\Sigma+\kappa_1 Z_1\right)^{-1}\Sigma\left(\Sigma+\kappa_1 Z_1\right)^{-1}Z_1 B^{\mathtt{df}}\right).
$$

and $\kappa_1 = \kappa(\lambda, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$.

Note that:

$$
\begin{aligned}
&(\kappa_1)^2\operatorname{Tr}\left(Z_1\left(\Sigma+\kappa_1 Z_1\right)^{-1}\Sigma\left(\Sigma+\kappa_1 Z_1\right)^{-1}Z_1 B^{\mathtt{df}}\right)\\
&=\operatorname{Tr}\left((\kappa_1 Z_1)\left(\Sigma+\kappa_1 Z_1\right)^{-1}\Sigma\left(\Sigma+\kappa_1 Z_1\right)^{-1}(\kappa_1 Z_1)B^{\mathtt{df}}\right)\\
&=\operatorname{Tr}\left(\left(I-\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\right)\Sigma\left(I-\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\right)^T B^{\mathtt{df}}\right)\\
&=\operatorname{Tr}\left(\Sigma B^{\mathtt{df}}\right)-2\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)+\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right).
\end{aligned}
$$

Note that:

$$
\begin{aligned}
&\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)\\
&+\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)\cdot\frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}(\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\\
&=\frac{\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}(\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}
\end{aligned}
$$

□

Now we are ready to prove Lemma 22.

Proof of Lemma 22. The statement follows trivially if $\alpha = 1$. By Lemma 23, it holds that:

$$\begin{aligned}
&(1-\alpha)^2 \operatorname{Tr}\left(\hat{\Sigma}_2(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{\mathtt{df}}\right) \\
&\sim \frac{\operatorname{Tr}\left(\left(\Sigma+\kappa_1 Z_1\right)^{-1}\Sigma\left(\Sigma+\kappa_1 Z_1\right)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)} \\
&+ \frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)} \cdot \left(\operatorname{Tr}\left(\Sigma B^{\mathtt{df}}\right) - 2\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)\right) \\
&\sim_{(A)} (1-\alpha)^2\left(\operatorname{Tr}\left(\left(\Sigma+\kappa I\right)^{-1}\Sigma\left(\Sigma+\kappa I\right)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)\right) \\
&+ \frac{\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})} \cdot (1-\alpha)^2 \cdot \operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right) \\
&+ \frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)} \cdot \left(\operatorname{Tr}\left(\Sigma B^{\mathtt{df}}\right) - 2(1-\alpha)\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)\right) \\
&= \frac{(1-\alpha)^2}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\left(\operatorname{Tr}\left(\left(\Sigma+\kappa I\right)^{-1}\Sigma\left(\Sigma+\kappa I\right)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)\right) \\
&+ \frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa_1 Z_1)^{-2})}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa_1 Z_1)^{-2})} \cdot \left(\operatorname{Tr}\left(\Sigma B^{\mathtt{df}}\right) - 2(1-\alpha)\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)\right) \\
&\sim_{(B)} \frac{(1-\alpha)^2}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\left(\operatorname{Tr}\left(\left(\Sigma+\kappa I\right)^{-1}\Sigma\left(\Sigma+\kappa I\right)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)\right) \\
&+ (1-\alpha)\frac{\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})} \cdot \left(\operatorname{Tr}\left(\Sigma B^{\mathtt{df}}\right) - 2(1-\alpha)\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma B^{\mathtt{df}}\Sigma\right)\right)
\end{aligned}$$

where (A) applies Lemma 16, Lemma 15, and (B) uses Lemma 17 and Lemma 18.

□

C.4 Analysis of Term 3 (T3)

We show the following deterministic equivalent for term 3.

Lemma 24. *Consider the setup of Lemma 6 and assume the notation above. Let $B^{\mathtt{mx}} = (\beta_1 - \beta_2)\beta_1^T$, and let $\kappa = \kappa(\lambda, N, \Sigma)$. Then it holds that:*

$$\begin{aligned}
&2\lambda(1-\alpha)\operatorname{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{\mathtt{mx}}\right) \\
&\sim \frac{2(1-\alpha)\kappa}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}\Sigma B^{\mathtt{mx}}\right) \\
&- 2\frac{(1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})} \cdot \kappa\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma B^{\mathtt{mx}}\right)
\end{aligned}$$

The analysis follows a similar structure to the analysis of (T2); we similarly unwrap the randomness in layers.

Lemma 25. *Consider the setup of Lemma 6 and assume the notation above. Assume $\alpha < 1$. Let $Z_1 = \frac{\alpha}{1-\alpha}\hat{\Sigma}_1 + \frac{\lambda}{1-\alpha}I$, and let $\kappa_1 = \kappa(1, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$. Then it holds that:*

$$
\begin{aligned}
&2\lambda(1-\alpha)^2 \operatorname{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{\mathtt{mx}}\right)\\
&\sim 2\frac{\lambda\kappa_1}{(1-\alpha)}\frac{\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{mx}}\right)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\\
&-2\frac{\lambda\kappa_1}{(1-\alpha)}\cdot\frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa_1 Z_1)^{-2})}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\cdot\operatorname{Tr}\left(\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{mx}}\right).
\end{aligned}
$$

Proof. By Claim 19 we have that:

$$
\begin{aligned}
2\lambda(1-\alpha)\operatorname{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{\mathtt{mx}}\right) &= 2\frac{\lambda}{(1-\alpha)}\operatorname{Tr}\left(\left(\hat{\Sigma}_2+Z_1\right)^{-1}\Sigma\left(\hat{\Sigma}_2+Z_1\right)^{-1}\hat{\Sigma}_2 B^{\mathtt{mx}}\right)\\
&\sim_{(A)} 2\frac{\lambda}{(1-\alpha)}\left(\kappa_1\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{mx}}\right)-E\right)
\end{aligned}
$$

where (A) follows from Lemma 40 and Claim 20, and E is defined such that

$$
E := \frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\cdot(\kappa_1)^2\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma\,(\Sigma+\kappa_1 Z_1)^{-1}Z_1 B^{\mathtt{mx}}\right).
$$

and $\kappa_1 = \kappa(\lambda, N(1-\alpha), Z_1^{-1/2}\Sigma Z_1^{-1/2})$.

Note that:

$$
\begin{aligned}
&(\kappa_1)^2\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma\,(\Sigma+\kappa_1 Z_1)^{-1}Z_1 B^{\mathtt{mx}}\right)\\
&=\kappa_1\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma\,(\Sigma+\kappa_1 Z_1)^{-1}(\kappa_1 Z_1)B^{\mathtt{mx}}\right)\\
&=\kappa_1\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma\left(I-(\Sigma+\kappa_1 Z_1)^{-1}\Sigma\right)B^{\mathtt{mx}}\right)\\
&=\kappa_1\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{mx}}\right)-\kappa_1\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{mx}}\right)
\end{aligned}
$$

Moreover, note that:

$$
\begin{aligned}
&2\frac{\lambda\kappa_1}{(1-\alpha)}\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{mx}}\right)\\
&+2\frac{\lambda}{(1-\alpha)}\cdot\frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\cdot\kappa_1\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{mx}}\right)\\
&=2\frac{\lambda}{(1-\alpha)}\frac{\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\mathtt{mx}}\right)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\cdot\kappa_1.
\end{aligned}
$$

□

Now we are ready to prove Lemma 22.

Proof of Lemma 22. The statement follows trivially if $\alpha = 1$. By Lemma 23, it holds that:

$$\begin{aligned}
&2\lambda(1-\alpha)^2 \operatorname{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}_2 B^{\texttt{mx}}\right) \\
&\sim 2\frac{\lambda\kappa_1}{(1-\alpha)}\frac{\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\texttt{mx}}\right)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)} \\
&-\frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\cdot 2\frac{\lambda\kappa_1}{(1-\alpha)}\operatorname{Tr}\left((\Sigma+\kappa_1 Z_1)^{-1}\Sigma B^{\texttt{mx}}\right) \\
&\sim_{(A)} 2(1-\alpha)\kappa\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}\Sigma B^{\texttt{mx}}\right) \\
&+2(1-\alpha)\kappa\frac{\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}\Sigma B^{\texttt{mx}}\right) \\
&-2\frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\cdot\kappa\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma B^{\texttt{mx}}\right) \\
&=2\frac{(1-\alpha)\kappa}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}\Sigma B^{\texttt{mx}}\right) \\
&-2\frac{\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}{1-\frac{1}{N(1-\alpha)}\operatorname{Tr}((\Sigma+\kappa_1 Z_1)^{-1}\Sigma(\Sigma+\kappa_1 Z_1)^{-1}\Sigma)}\cdot\kappa\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma B^{\texttt{mx}}\right) \\
&\sim_{(B)} 2\frac{(1-\alpha)\kappa}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}\Sigma B^{\texttt{mx}}\right) \\
&-2(1-\alpha)\frac{\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}{1-\frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma+\kappa I)^{-2})}\cdot\kappa\operatorname{Tr}\left((\Sigma+\kappa I)^{-1}\Sigma B^{\texttt{mx}}\right)
\end{aligned}$$

where (A) applies Lemma 16, Lemma 15, and Lemma 18, and (B) uses Lemma 17 and Lemma 18. □

C.5 Proof of Lemma 6

Lemma 6 follows from the sublemmas in this section.

Proof. We apply Claim 14 to decompose the error in terms (T1), (T2), and (T3). We replace these terms with deterministic equivalents using Lemma 21, Lemma 22, and Lemma 24. The statement follows from adding these terms. □

C.6 Reformulation of Lemma 6 using assumptions from Section 2.3

Under the assumptions from Section 2.3, we show the following:

Lemma 26. *Suppose that power scaling holds for the eigenvalues and alignment coefficients with scaling $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Suppose that $\lambda \in (0, 1)$, and $N \geq 1$. Let $L_1^{\texttt{det}} := L_1^{\texttt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$ be the deterministic equivalent from Lemma 6. Let $\kappa = \kappa(\lambda, N, \Sigma)$ from Definition 2. Let $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma(\beta_1 - \beta_2)]$. Then it* *holds that:*

$$Q \cdot \mathbb{E}_{\mathcal{D}_W}[L_1^{det}] = \kappa^2(1-2(1-\alpha)^2(1-\rho)) \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + (1-\alpha)^2 L^*(\rho)$$

$$+ 2\kappa(1-\rho)(1-\alpha)(1-2(1-\alpha)) \sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2}$$

$$+ 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot (1-2(1-\alpha)) \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa},$$

where $Q = 1 - \frac{1}{N}\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}$.

Before proving Lemma 26, we prove a number of sublemmas where we analyze each of the terms in Lemma 6 using the assumptions from Section 2.3. In the proofs in this section, we use the notation $F \approx F'$ to denote that $F = \Theta(F')$ where the Θ is allowed to hide dependence on the scaling exponents γ and δ. Moreover let $\Sigma = V\Lambda V^T$ be the eigendecomposition of Σ, where Λ is a diagonal matrix consisting of the eigenvalues.

Lemma 27. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0,1)$, suppose that $P = \infty$. Assume the notation from Lemma 6. Let $\nu = \min(2(1+\gamma), \gamma+\delta)$. Then it holds that:*

$$\mathbb{E}_{\mathcal{D}_W}[T_1] := \kappa^2 \cdot \mathrm{Tr}(\Sigma\Sigma_\kappa^{-2}\mathbb{E}_{\mathcal{D}_W}[B^{sn}]) = \kappa^2 \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2}.$$

Proof. Observe that:

$$\begin{aligned}\mathrm{Tr}(\Sigma\Sigma_\kappa^{-2}\mathbb{E}_{\mathcal{D}_W}[B^{\mathtt{sn}}]) &= \mathrm{Tr}(\Lambda(\Lambda+\kappa I)^{-2}\mathbb{E}_{\mathcal{D}_W}[V^T\beta_1\beta_1^T V]) \\ &= \sum_{i=1}^{P} \frac{i^{-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \cdot \mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i\rangle^2] \\ &= \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2}\end{aligned}$$

□

Lemma 28. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0,1)$, suppose that $P = \infty$. Assume the notation from Lemma 6. Then it holds that:*

$$\mathbb{E}_{\mathcal{D}_W}[T_2] := (1-\alpha)^2\left(\mathrm{Tr}\left(\Sigma_\kappa^{-2}\Sigma^3\mathbb{E}_{\mathcal{D}_W}[B^{df}]\right)\right) = 2(1-\alpha)^2(1-\rho)\sum_{i=1}^{P} \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2}.$$

Proof. First, we observe that

$$\mathbb{E}_{\mathcal{D}_W}[\langle \beta_1-\beta_2, v_i\rangle^2] = \mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i\rangle^2]+\mathbb{E}_{\mathcal{D}_W}[\langle \beta_2, v_i\rangle^2]-2\mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i\rangle\langle \beta_2, v_i\rangle] = i^{-\delta}+i^{-\delta}-2\rho i^{-\delta} = 2(1-\rho)i^{-\delta}.$$

It is easy to see that:

$$\begin{aligned}
\operatorname{Tr}\left(\Sigma_\kappa^{-2}\Sigma^3\mathbb{E}_{\mathcal{D}_W}[B^{\mathtt{df}}]\right) &= \operatorname{Tr}(\Lambda^3(\Lambda+\kappa I)^{-2}\mathbb{E}_{\mathcal{D}_W}[V^T(\beta_1-\beta_2)(\beta_1-\beta_2)^T V]) \\
&= \sum_{i=1}^P \frac{i^{-3(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2}\cdot\mathbb{E}_{\mathcal{D}_W}[\langle\beta_1-\beta_2, v_i\rangle^2] \\
&= 2(1-\rho)\sum_{i=1}^P \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2}.
\end{aligned}$$

□

Lemma 29. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0,1)$, suppose that $P = \infty$. Assume the notation from Lemma 6. Then it holds that:*

$$\mathbb{E}_{\mathcal{D}_W}[T_3] := 2(1-\alpha)\kappa\cdot\operatorname{Tr}\left(\Sigma_\kappa^{-2}\Sigma^2 B^{\mathtt{mx}}\right) = 2(1-\alpha)\kappa(1-\rho)\sum_{i=1}^P\frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}.$$

Proof. First, we observe that

$$\mathbb{E}_{\mathcal{D}_W}[\langle\beta_1-\beta_2, v_i\rangle\langle\beta_1, v_i\rangle] = \mathbb{E}_{\mathcal{D}_W}[\langle\beta_1, v_i\rangle^2] - \mathbb{E}_{\mathcal{D}_W}[\langle\beta_1, v_i\rangle\langle\beta_2, v_i\rangle] = i^{-\delta} - \rho i^{-\delta} = (1-\rho)i^{-\delta}.$$

Observe that:

$$\begin{aligned}
\operatorname{Tr}\left(\Sigma_\kappa^{-2}\Sigma^2 B^{\mathtt{mx}}\right) &= \operatorname{Tr}(\Lambda^2(\Lambda+\kappa I)^{-2}\mathbb{E}_{\mathcal{D}_W}[V^T(\beta_1-\beta_2)\beta_1^T V]) \\
&= \sum_{i=1}^P\frac{i^{-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2}\cdot\mathbb{E}_{\mathcal{D}_W}[\langle\beta_1-\beta_2, v_i\rangle\langle\beta_1, v_i\rangle] \\
&= (1-\rho)\sum_{i=1}^P\frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}.
\end{aligned}$$

This means that:

$$\mathbb{E}_{\mathcal{D}_W}[T_3] = 2(1-\alpha)\kappa(1-\rho)\sum_{i=1}^P\frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}.$$

□

Lemma 30. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0,1)$, suppose that $P = \infty$. Assume the notation from Lemma 6. Then it holds that:*

$$\begin{aligned}
|\mathbb{E}_{\mathcal{D}_W}[T_4]| &:= 2\kappa(1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2})\cdot\operatorname{Tr}\left(\Sigma_\kappa^{-1}\Sigma\mathbb{E}_{\mathcal{D}_W}[B^{\mathtt{mx}}]\right) \\
&= 2\kappa(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^P\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\left(\sum_{i=1}^P\frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa}\right)
\end{aligned}$$

Proof. First, we observe that

$$\mathbb{E}_{\mathcal{D}_W}[\langle \beta_1 - \beta_2, v_i\rangle\langle \beta_1, v_i\rangle] = \mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i\rangle^2] - \mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i\rangle\langle \beta_2, v_i\rangle] = i^{-\delta} + -\rho i^{-\delta} = (1-\rho)i^{-\delta}.$$

Observe that:

$$\begin{aligned}\operatorname{Tr}\left(\Sigma_\kappa^{-1}\Sigma\mathbb{E}_{\mathcal{D}_W}[B^{\mathtt{mx}}]\right) &= \operatorname{Tr}(\Lambda(\Lambda+\kappa I)^{-1}\mathbb{E}_{\mathcal{D}_W}[V^T(\beta_1-\beta_2)\beta_1^T V]) \\ &= \sum_{i=1}^{P} \frac{i^{-1-\gamma}}{i^{-1-\gamma}+\kappa} \cdot \mathbb{E}_{\mathcal{D}_W}[\langle \beta_1 - \beta_2, v_i\rangle\langle \beta_1, v_i\rangle] \\ &= (1-\rho)\sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa}.\end{aligned}$$

Now, apply Lemma 32, we see that:

$$\begin{aligned}|\mathbb{E}_{\mathcal{D}_W}[T_4]| &:= 2\kappa(1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2}) \cdot \operatorname{Tr}\left(\Sigma_\kappa^{-1}\Sigma B^{\mathtt{mx}}\right) \\ &=_{(A)} 2\kappa(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\left(\sum_{i=1}^{P}\frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa}\right)\end{aligned}$$

where (A) follows from Lemma 32. □

Lemma 31. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0,1)$, suppose that $P = \infty$. Assume the notation from Lemma 6, and similarly let*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_W}[T_5] &:= (1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2}) \cdot \left(\operatorname{Tr}\left(\Sigma\mathbb{E}_{\mathcal{D}_W}[B^{\mathit{df}}]\right) - 2(1-\alpha)\operatorname{Tr}\left(\Sigma_\kappa^{-1}\Sigma^2\mathbb{E}_{\mathcal{D}_W}[B^{\mathit{df}}]\right)\right) \\ &= 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot \left(\sum_{i=1}^{P} i^{-\delta-1-\gamma} - 2(1-\alpha)\cdot\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)}\right).\end{aligned}$$

Proof. First, we observe that

$$\mathbb{E}_{\mathcal{D}_W}[\langle \beta_1 - \beta_2, v_i\rangle^2] = \mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i\rangle^2] + \mathbb{E}_{\mathcal{D}_W}[\langle \beta_2, v_i\rangle^2] - 2\mathbb{E}_{\mathcal{D}_W}[\langle \beta_1, v_i\rangle\langle \beta_2, v_i\rangle] = i^{-\delta} + i^{-\delta} - 2\rho i^{-\delta} = 2(1-\rho)i^{-\delta}.$$

Now, observe that:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_W}[T_5] &:= (1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2}) \cdot \left(\operatorname{Tr}\left(\Sigma\mathbb{E}_{\mathcal{D}_W}[B^{\mathtt{df}}]\right) - 2(1-\alpha)\operatorname{Tr}\left(\Sigma_\kappa^{-1}\Sigma^2\mathbb{E}_{\mathcal{D}_W}[B^{\mathtt{df}}]\right)\right) \\ &= (1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2}) \cdot \left(\operatorname{Tr}\left(\Lambda\mathbb{E}_{\mathcal{D}_W}[V^T(\beta_1-\beta_2)(\beta_1-\beta_2)^T V]\right)\right) \\ &- (1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2}) \cdot \left(2(1-\alpha)\operatorname{Tr}\left((\Lambda+\kappa I)^{-1}\Lambda^2\mathbb{E}_{\mathcal{D}_W}[V^T(\beta_1-\beta_2)(\beta_1-\beta_2)^T V]\right)\right) \\ &= (1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2}) \cdot \left(\sum_{i=1}^{P} i^{-1-\gamma}\langle \beta_1-\beta_2, v_i\rangle^2 - 2(1-\alpha)\cdot\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)}\langle \beta_1-\beta_2, v_i\rangle^2\right) \\ &= 2(1-\alpha)(1-\rho)\frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2}) \cdot \left(\sum_{i=1}^{P} i^{-\delta-1-\gamma} - 2(1-\alpha)\cdot\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)}\right) \\ &=_{(A)} 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot \left(\sum_{i=1}^{P} i^{-\delta-1-\gamma} - 2(1-\alpha)\cdot\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)}\right).\end{aligned}$$

where (A) uses Lemma 32. □

The proofs of these sublemmas use the following fact.

Lemma 32. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, suppose that $P = \infty$. Assume the notation from Lemma 6. Then it holds that:*

$$\operatorname{Tr}\left(\Sigma^2(\Sigma + \kappa I)^{-2}\right) = \sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma} + \kappa)^2}.$$

Proof. We see that:

$$\begin{aligned}(1-\alpha)\frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2}) &= (1-\alpha)\frac{1}{N}\operatorname{Tr}(V\Lambda^2(\Lambda + \kappa I)^{-2}V^T) \\ &= (1-\alpha)\frac{1}{N}\operatorname{Tr}(\Lambda^2(\Lambda + \kappa I)^{-2}) \\ &= \sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma} + \kappa)^2}.\end{aligned}$$

□

Now, we are ready to prove Lemma 26.

Proof of Lemma 26. By Lemma 32, we know:

$$Q = 1 - \frac{1}{N}\operatorname{Tr}(\Sigma^2(\Sigma + \kappa I)^{-2}) = 1 - \frac{1}{N}\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma} + \kappa)^2}.$$

Moreover, we have that:

$$\begin{aligned}Q \cdot \mathbb{E}_{\mathcal{D}_W}[L_1^{\mathtt{det}}] &=_{(A)} \mathbb{E}_{\mathcal{D}_W}[T_1 + T_2 + T_3 + T_4 + T_5] \\ &=_{(B)} \kappa^2 \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma} + \kappa)^2} + 2(1-\alpha)^2(1-\rho)\sum_{i=1}^{P} \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma} + \kappa)^2} \\ &+ 2\kappa(1-\rho)(1-\alpha)\sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma} + \kappa)^2} \\ &- 2\kappa(1-\rho)(1-\alpha)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma} + \kappa)^2}\right)\left(\sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma} + \kappa}\right) \\ &+ 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma} + \kappa)^2}\right) \cdot \left(\sum_{i=1}^{P} i^{-\delta-1-\gamma} - 2(1-\alpha) \cdot \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma} + \kappa)}\right).\end{aligned}$$

where (A) follows from Lemma 6, and (B) follows from Lemmas 27-31.

By Claim 13, we know that:

$$L^*(\rho) = 2(1-\rho)\sum_{i=1}^{P} i^{-\delta-1-\gamma} = 2(1-\rho)\sum_{i=1}^{P} \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma})^2}.$$

This means that:

$$
\begin{aligned}
&L^*(\rho) - 2(1-\rho)\sum_{i=1}^{P} \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&= 2(1-\rho)\sum_{i=1}^{P}\left(\frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma})^2} - \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2}\right) \\
&= 2(1-\rho)\sum_{i=1}^{P}\left(\frac{i^{-\delta-3(1+\gamma)}\cdot((i^{-1-\gamma}+\kappa)^2-(i^{-1-\gamma})^2)}{(i^{-1-\gamma})^2\cdot(i^{-1-\gamma}+\kappa)^2}\right) \\
&= 2\kappa^2(1-\rho)\sum_{i=1}^{P}\left(\frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma})^2\cdot(i^{-1-\gamma}+\kappa)^2}\right) + 4\kappa(1-\rho)\sum_{i=1}^{P}\left(\frac{i^{-\delta-3(1+\gamma)}\cdot i^{-1-\gamma}}{(i^{-1-\gamma})^2\cdot(i^{-1-\gamma}+\kappa)^2}\right) \\
&= 2\kappa^2(1-\rho)\sum_{i=1}^{P}\left(\frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) + 4\kappa(1-\rho)\sum_{i=1}^{P}\left(\frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2}\right)
\end{aligned}
$$

Applying this and some other algebraic manipulations, we obtain that:

$$
\begin{aligned}
Q\cdot L_1^{\texttt{det}} &= \kappa^2(1-2(1-\alpha)^2(1-\rho))\sum_{i=1}^{P}\frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + (1-\alpha)^2 L^*(\rho) \\
&+ 2\kappa(1-\rho)(1-\alpha)(1-2(1-\alpha))\sum_{i=1}^{P}\frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&- 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\left(\sum_{i=1}^{P} i^{-\delta-1-\gamma} - \sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa}\right) \\
&+ 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\cdot\left(\sum_{i=1}^{P} i^{-\delta-1-\gamma} - 2(1-\alpha)\cdot\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)}\right) \\
&= \kappa^2(1-2(1-\alpha)^2(1-\rho))\sum_{i=1}^{P}\frac{i^{-\delta-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + (1-\alpha)^2 L^*(\rho) \\
&+ 2\kappa(1-\rho)(1-\alpha)(1-2(1-\alpha))\sum_{i=1}^{P}\frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\cdot(1-2(1-\alpha))\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa}.
\end{aligned}
$$

□

C.7 Proof of Theorem 7

We now prove Theorem 7. In the proof, we again use the notation $F \approx F'$ to denote $F = \Theta(F')$. The main ingredient is Lemma 26, coupled with the auxiliary calculations in Appendix C.11.

Proof. The proof boils down to three steps: (1) obtaining an exact expression, (2) obtaining an up-to-constants asymptotic expression in terms of κ and Q, and (3) substituting in κ and Q.

Step 1: Exact expression. We apply Lemma 26 to see that:

$$Q \cdot L_1^{\mathtt{det}} = \kappa^2(1-2(1-\alpha)^2(1-\rho))\sum_{i=1}^{P}\frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + (1-\alpha)^2 L^*(\rho)$$
$$+ 2\kappa(1-\rho)(1-\alpha)(1-2(1-\alpha))\sum_{i=1}^{P}\frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2}$$
$$+ 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\cdot(1-2(1-\alpha))\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa},$$

where $Q = 1 - \frac{1}{N}\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}$, where $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1-\beta_2)^T\Sigma(\beta_1-\beta_2)]$, and where $\kappa = \kappa(\Sigma, N, \lambda)$ as defined in Definition 2.

Step 2: Asymptotic expression in terms of κ and Q. We show that

$$Q \cdot L_1^{\mathtt{det}} \approx \kappa^{\frac{\nu}{1+\gamma}} + (1-\alpha)^2(1-\rho) + (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}.$$

We analyze this expression term-by-term and repeatedly apply Lemma 33. We see that:

$$\kappa^2(1-2(1-\alpha)^2(1-\rho))\sum_{i=1}^{P}\frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \approx_{(A)} \kappa^{\frac{\nu}{1+\gamma}}(1-2(1-\alpha)^2(1-\rho)) \approx_{(B)} \kappa^{\frac{\nu}{1+\gamma}},$$

where (A) uses Lemma 33 and (B) uses that $\alpha \geq 0.5$. Moreover, we observe that:

$$(1-\alpha)^2 L^*(\rho) \approx_{(C)} (1-\alpha)^2(1-\rho),$$

where (C) uses Claim 13. Moreover, we see that:

$$\begin{aligned}
2\kappa(1-\rho)(1-\alpha)(1-2(1-\alpha))\sum_{i=1}^{P}\frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} &\approx_{(D)} (1-\alpha)(1-\rho)(1-2(1-\alpha))\max\left(\kappa, \kappa^{\frac{\delta+\gamma}{1+\gamma}}\right) \\
&=_{(E)} O\left((1-\alpha)\sqrt{1-\rho}\max\left(\kappa, \kappa^{\frac{\delta+\gamma}{2(1+\gamma)}}\right)\right) \\
&= O\left(\sqrt{(1-\alpha)^2(1-\rho)\cdot\kappa^{\frac{\min(2(1+\gamma),\gamma+\delta)}{1+\gamma}}}\right) \\
&=_{(F)} O\left(\kappa^{\frac{\min(2(1+\gamma),\gamma+\delta)}{1+\gamma}} + (1-\alpha)^2(1-\rho)\right) \\
&= O\left(\kappa^{\frac{\nu}{1+\gamma}} + (1-\alpha)^2(1-\rho)\right)
\end{aligned}$$

where (D) uses Lemma 33, (E) uses that $1-\rho \leq 1$ and that $\kappa = O(1)$ (which follows from Lemma 35 and the assumption that $\lambda \in (0,1)$) and (F) follows from AM-GM. Finally, observe that:

$$\begin{aligned}
&2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\cdot(1-2(1-\alpha))\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\
&\approx (1-2(1-\alpha))\cdot(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\
&\approx_{(G)} (1-2(1-\alpha))\cdot(1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}
\end{aligned}$$

where (G) uses Lemma 33 twice.

Putting this all together, we see that:

$$Q \cdot L_1^{\mathtt{det}} \approx \kappa^{\frac{\nu}{1+\gamma}} + (1-\alpha)^2(1-\rho) + (1-2(1-\alpha)) \cdot (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}.$$

We split into two cases based on α. When $\alpha \geq 0.75$, we observe that

$$(1-2(1-\alpha)) \cdot (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N} \approx (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N},$$

and when $\alpha \in [0.5, 0.75]$, we observe that

$$(1-2(1-\alpha)) \cdot (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N} = O\left((1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}\right)$$

and

$$(1-\alpha)^2(1-\rho) \approx_{(H)} (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}$$

where (H) follows from the fact that $\kappa = \Omega(N^{-1-\gamma})$ by Lemma 35. Altogether, this implies that:

$$Q \cdot L_1^{\mathtt{det}} \approx \kappa^{\frac{\nu}{1+\gamma}} + (1-\alpha)^2(1-\rho) + (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N},$$

as desired.

Step 2: Substitute in κ and Q. Finally, we apply Lemma 34 to see that:

$$Q^{-1} = \left(1 - \frac{1}{N}\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)^{-1} = \Theta(1).$$

We apply Lemma 35 to see that

$$\kappa = \kappa(\Sigma, N, \Sigma) = \max(N^{-1-\gamma}, \lambda).$$

Plugging this into the expression derived in Step 2, we obtain the desired expression. □

C.8 Proof of Corollary 8

We prove Corollary 8 using Theorem 7.

Proof. We apply Theorem 7 to see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^{\mathtt{det}}] = \Theta\left(\underbrace{\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu})}_{\text{finite data error}} + \underbrace{(1-\alpha)^2 \cdot (1-\rho)}_{\text{mixture error}} + \underbrace{(1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)(1-\rho)}_{\text{overfitting error}}\right).$$

We split into three cases: $N \leq (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$, $(1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$, and $N \geq (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$.

Case 1: $N \le (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$**.** We observe that the finite data error dominates regardless of λ. This is because the condition implies that

$$\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) \ge (1-\alpha)(1-\rho),$$

which dominates both the mixture error and the overfitting error.

Case 2: $(1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \le N \le (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$**.** We show that the finite error term and overfitting error dominate. Let $\tilde{N} = \min(\lambda^{-\frac{1}{1+\gamma}}, N)$. We can bound the sum of the finite data error and the overfitting error as:

$$\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) + (1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)(1-\rho) = \tilde{N}^{-\nu} + (1-\alpha)(1-\rho)\frac{\tilde{N}}{N}.$$

Taking a derivative (and verifying the second order condition), we see that this expression is minimized when:

$$\nu \cdot \tilde{N}^{-\nu-1} = \frac{(1-\alpha)(1-\rho)}{N}$$

which solves to:

$$\tilde{N} = \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{-\frac{1}{1+\nu}}\right).$$

The lower bound on N guarantees that:

$$\tilde{N} = \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{-\frac{1}{1+\nu}}\right) = O\left(\left((1-\alpha)^{1+\frac{1}{\nu}}(1-\rho)^{1+\frac{1}{\nu}}\right)^{-\frac{1}{1+\nu}}\right) = O\left((1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}}\right) = O(N)$$

which ensures that $\tilde{N}$ can be achieved by some choice of λ. In particular, we can take $\lambda = \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{1+\gamma}{\nu+1}}\right)$.

The resulting sum of the finite error and the overfitting error is:

$$\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) + (1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right) = \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{\nu}{\nu+1}}\right).$$

The upper bound on N guarantees that this dominates the mixture error:

$$\Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{\nu}{\nu+1}}\right) = \Omega\left(\left((1-\alpha)^{1+\frac{2+\nu}{\nu}}(1-\rho)^{1+\frac{1}{\nu}}\right)^{\frac{\nu}{\nu+1}}\right) = \Omega((1-\alpha)^2(1-\rho))$$

as desired.

Case 3: $N \ge (1-\alpha)^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$**.** We show that the mixture and the overfitting error terms dominate. First, we observe that the sum of the mixture error and the finite data error is:

$$(1-\alpha)^2(1-\rho)+(1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)(1-\rho) = \Theta\left((1-\alpha)(1-\rho)\left(1-\alpha+\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)\right).$$

This is minimized by taking $\lambda = \Theta((N(1-\alpha))^{-1-\gamma})$, which yields $\Theta((1-\alpha)^2(1-\rho))$.

The upper bound on N and the setting of λ guarantees that this term dominates the finite data error:

$$\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) = O((N(1-\alpha))^{-\nu}) \le O\left((1-\alpha)^{-\nu}(1-\alpha)^{2+\nu}(1-\rho)\right) = O((1-\alpha)^2(1-\rho),$$

as desired.

□

C.9 Proof of Theorem 9

We prove Theorem 9.

Proof of Theorem 9. Like the proof of Theorem 7, the proof boils down to three steps: (1) obtaining an exact expression, (2) obtaining an up-to-constants asymptotic expression in terms of κ, and (3) substituting in κ.

Step 1: Exact expression. We first apply Lemma 26 to obtain the precise loss:

$$
\begin{aligned}
Q \cdot \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_E, N, \alpha_E)] &= \kappa^2(1 - 2(1-\alpha)^2(1-\rho)) \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + (1-\alpha)^2 L^*(\rho) \\
&+ 2\kappa(1-\rho)(1-\alpha)(1-2(1-\alpha)) \sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot (1-2(1-\alpha)) \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa},
\end{aligned}
$$

where $Q = 1 - \frac{1}{N}\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}$ and where $\kappa = \kappa(\Sigma, N, \lambda)$ as defined in Definition 2. This can be written as:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_E, N, \alpha_E)] - (1-\alpha)^2 L^*(\rho) \\
&= Q^{-1} \cdot \kappa^2(1 - 2(1-\alpha)^2(1-\rho)) \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ Q^{-1} \cdot 2\kappa(1-\rho)(1-\alpha)(1-2(1-\alpha)) \sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ Q^{-1} \cdot 2(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot (1-2(1-\alpha)) \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\
&+ \frac{1-Q}{Q}(1-\alpha)^2 L^*(\rho).
\end{aligned}
$$

Step 2: Asymptotic expression in terms of κ. We use the notation $F \approx F'$ to denote that

$F = \Theta(F')$. We obtain:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_E, N, \alpha_E)] - (1-\alpha)^2 L^*(\rho) \\
&\approx_{(A)} \kappa^2 (1 - 2(1-\alpha)^2(1-\rho)) \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ \kappa(1-\rho)(1-\alpha)(1-2(1-\alpha)) \sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ (1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot (1-2(1-\alpha)) \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\
&+ (1-Q)(1-\alpha)^2 L^*(\rho) \\
&\approx_{(B)} \kappa^2 \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + \kappa(1-\rho)(1-\alpha) \sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ (1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\
&+ (1-\alpha)^2 L^*(\rho) \cdot \frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right).
\end{aligned}$$

where (A) uses that Q^{-1} is a constant by Lemma 34 and (B) uses that $\alpha \geq 0.75$ and the definition of Q. Now, using the bounds from Lemma 33, and the bound from Claim 13, we obtain:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_E, N, \alpha_E)] - (1-\alpha)^2 L^*(\rho) \\
&\approx \kappa^{\frac{\min(2(1+\gamma), \gamma+\delta)}{1+\gamma}} + (1-\rho)(1-\alpha)\max\left(\kappa, \kappa^{\frac{\gamma+\delta}{1+\gamma}}\right) + (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N} + \frac{\kappa^{-\frac{1}{1+\gamma}}}{N}(1-\alpha)^2(1-\rho) \\
&\approx \kappa^{\frac{\nu}{1+\gamma}} + (1-\rho)(1-\alpha)\kappa^{\frac{\nu'}{1+\gamma}} + (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}.
\end{aligned}$$

Step 3: Substituting in κ. Finally, we apply Lemma 35 to see that

$$\kappa = \kappa(\Sigma, N, \Sigma) = \max(N^{-1-\gamma}, \lambda).$$

Plugging this into the expression derived in Step 2, we obtain the desired expression. □

C.10 Proof of Corollary 10

We prove Corollary 10 using Theorem 9.

Proof. We apply Theorem 7 to see that:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_1^{\texttt{det}} - L_1(\beta(\alpha, 0))] \\
&= \Theta\left(\underbrace{\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu})}_{\text{finite data error}} + \underbrace{(1-\rho)(1-\alpha)\max(\lambda^{\frac{\nu'}{1+\gamma}}, N^{-\nu'})}_{\text{mixture finite data error}} + \underbrace{(1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)(1-\rho)}_{\text{overfitting error}}\right).
\end{aligned}$$

We split into three cases: $N \leq (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$, $(1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}$, and $N \geq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}$.

Case 1: $N \leq (1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}}$**.** We observe that the finite data error dominates regardless of λ. This is because the condition implies that

$$\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) \geq (1-\alpha)(1-\rho),$$

which dominates both the mixture finite data error and the overfitting error.

Case 2: $(1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \leq N \leq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}$**.** We show that the finite error term and overfitting error dominate. Let $\tilde{N} = \min(\lambda^{-\frac{1}{1+\gamma}}, N)$. We can bound the sum of the finite data error and the overfitting error as:

$$\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) + (1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)(1-\rho) = \tilde{N}^{-\nu} + (1-\alpha)(1-\rho)\frac{\tilde{N}}{N}.$$

Taking a derivative (and verifying the second order condition), we see that this expression is minimized when:

$$\nu \cdot \tilde{N}^{-\nu-1} = \frac{(1-\alpha)(1-\rho)}{N}$$

which solves to:

$$\tilde{N} = \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{-\frac{1}{1+\nu}}\right).$$

The lower bound on N guarantees that:

$$\tilde{N} = \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{-\frac{1}{1+\nu}}\right) = O\left(\left((1-\alpha)^{1+\frac{1}{\nu}}(1-\rho)^{1+\frac{1}{\nu}}\right)^{-\frac{1}{1+\nu}}\right) = O\left((1-\alpha)^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}}\right) = O(N)$$

which ensures that $\tilde{N}$ can be achieved by some choice of λ. In particular, we can take $\lambda = \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{1+\gamma}{\nu+1}}\right)$.

The resulting sum of the finite error and the overfitting error is:

$$\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) + (1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right) = \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{\nu}{\nu+1}}\right).$$

The upper bound on N and the choice of λ guarantees that this dominates the mixture finite data error, as shown below:

$$
\begin{aligned}
&(1-\rho)(1-\alpha)\max(\lambda^{\frac{\nu'}{1+\gamma}}, N^{-\nu'}) \\
&= \Theta\left((1-\rho)(1-\alpha)\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{\nu'}{\nu+1}}\right) \\
&= \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{\nu}{\nu+1}}(1-\alpha)(1-\rho)\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{\nu'-\nu}{\nu+1}}\right) \\
&= \Theta\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{\nu}{\nu+1}}(1-\alpha)^{\frac{\nu'+1}{\nu+1}}(1-\rho)^{\frac{\nu'+1}{\nu+1}}N^{\frac{\nu-\nu'}{\nu+1}}\right) \\
&= O\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{\nu}{\nu+1}}(1-\alpha)^{\frac{\nu'+1}{\nu+1}}(1-\rho)^{\frac{\nu'+1}{\nu+1}}(1-\alpha)^{-\frac{\nu'+1}{\nu+1}}(1-\rho)^{-\frac{\nu'+1}{\nu+1}}\right) \\
&= O\left(\left(\frac{(1-\alpha)(1-\rho)}{N}\right)^{\frac{\nu}{\nu+1}}\right)
\end{aligned}
$$

as desired.

Case 3: $N \geq (1-\alpha)^{-\frac{\nu'+1}{\nu-\nu'}}(1-\rho)^{-\frac{\nu'+1}{\nu-\nu'}}$**.** We show that the mixture finite data error and the overfitting error terms dominate. First, we observe that the sum of the mixture error and the finite data error is:

$$
\begin{aligned}
&(1-\rho)(1-\alpha)\max(\lambda^{\frac{\nu'}{1+\gamma}}, N^{-\nu'}) + (1-\alpha)\left(\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)(1-\rho) \\
&= \Theta\left((1-\alpha)(1-\rho)\left(\lambda^{\frac{\nu'}{1+\gamma}} + \frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)\right)
\end{aligned}
$$

This is minimized by taking $\lambda = \Theta(N^{-\frac{1+\gamma}{\nu'+1}})$, which yields $\Theta((1-\alpha)(1-\rho)N^{-\frac{\nu'}{\nu'+1}})$.

The upper bound on N and the setting of λ guarantees that this term dominates the finite data error:

$$
\begin{aligned}
\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) &= \Theta(N^{-\frac{\nu}{\nu'+1}}) \\
&\leq \Theta\left((1-\alpha)(1-\rho)N^{-\frac{\nu'}{\nu'+1}}(1-\alpha)^{-1}(1-\rho)^{-1}N^{-\frac{\nu-\nu'}{\nu'+1}}\right) \\
&= O\left((1-\alpha)(1-\rho)N^{-\frac{\nu'}{\nu'+1}}(1-\alpha)^{-1}(1-\rho)^{-1}(1-\alpha)(1-\rho)\right) \\
&= O\left((1-\alpha)(1-\rho)N^{-\frac{\nu'}{\nu'+1}}\right)
\end{aligned}
$$

as desired.

□

C.11 Auxiliary calculations under power scaling assumptions

We show the following auxiliary calculations which we use when analyzing the terms in Lemma 6 under the power scaling assumptions. Throughout this section, we again use the notation $F \approx F'$ to denote that $F = \Theta(F')$.

Lemma 33. *Suppose that power-law scaling holds for eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Let $\kappa = \kappa(\lambda, N, \Sigma)$ be defined according to Definition 2. Then the following holds:*

$$\sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \approx \kappa^{-2}\kappa^{\frac{\min(2(1+\gamma),\gamma+\delta)}{1+\gamma}}$$

$$\sum_{i=1}^{P} \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \approx 1$$

$$\sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \approx 1$$

$$\sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \approx \max(1, \kappa^{\frac{\delta-1}{1+\gamma}})$$

$$\sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa} \approx \max(1, \kappa^{\frac{\delta-1}{1+\gamma}})$$

$$\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2} \approx \kappa^{-\frac{1}{1+\gamma}}$$

$$\sum_{i=1}^{P} \frac{i^{-1-\gamma}}{i^{-1-\gamma}+\kappa} \approx \kappa^{-\frac{1}{1+\gamma}}$$

$$\sum_{i=1}^{P} \frac{i^{-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \approx \kappa^{-2}\kappa^{\frac{\gamma}{1+\gamma}}$$

Proof. To prove the first statement, observe that:

$$\begin{aligned}
\sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} &= \sum_{i \le \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + \sum_{i \ge \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\
&\approx \sum_{i \le \kappa^{-\frac{1}{1+\gamma}}} i^{1+\gamma-\delta} + \kappa^{-2} \sum_{i \ge \kappa^{-\frac{1}{1+\gamma}}} i^{-\delta-1-\gamma} \\
&\approx \max(1, \kappa^{-\frac{2+\gamma-\delta}{1+\gamma}}) + \kappa^{-2}\kappa^{\frac{\delta+\gamma}{1+\gamma}} \\
&= \kappa^{-2}\max(\kappa^2, \kappa^{\frac{\gamma+\delta}{1+\gamma}}) + \kappa^{-2}\kappa^{\frac{\delta+\gamma}{1+\gamma}} \\
&\approx \kappa^{-2}\max(\kappa^2, \kappa^{\frac{\gamma+\delta}{1+\gamma}}) \\
&\approx \kappa^{-2}\kappa^{\frac{\min(2(1+\gamma),\gamma+\delta)}{1+\gamma}}.
\end{aligned}$$

To prove the second statement, we use Lemma 35 and the assumption that $\lambda \in (0, 1)$ to see $\kappa = \Theta(\max(\lambda, N^{-1-\gamma})) = O(1)$. This means that

$$\sum_{i=1}^{P} \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} = \Omega\left(\sum_{i=1}^{P} i^{-\delta-3(1+\gamma)}\right) = \Omega(1).$$

Moreover, we see that:

$$\sum_{i=1}^{P} \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} = O\left(\sum_{i=1}^{P} \frac{i^{-\delta-3(1+\gamma)}}{(i^{-1-\gamma})^2}\right) = O\left(\sum_{i=1}^{P} i^{-\delta-1-\gamma)}\right) = \Omega(1).$$

To prove the third statement, we use Lemma 35 and the assumption that $\lambda \in (0,1)$ to see $\kappa = \Theta(\max(\lambda, N^{-1-\gamma})) = O(1)$. This means that

$$\sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} = \Omega\left(\sum_{i=1}^{P} i^{-\delta-2(1+\gamma)}\right) = \Omega(1).$$

Moreover, we see that:

$$\sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{i^{-1-\gamma}+\kappa} = O\left(\sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{i^{-1-\gamma}}\right) = O\left(\sum_{i=1}^{P} i^{-\delta-1-\gamma}\right) = O(1).$$

To prove the fourth statement, observe that:

$$\begin{aligned}
\sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2} &\approx \sum_{i \le \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2} + \sum_{i \ge \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-\delta-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\
&\approx \sum_{i \le \kappa^{-\frac{1}{1+\gamma}}} i^{-\delta} + \kappa^{-2} \sum_{i \ge \kappa^{-\frac{1}{1+\gamma}}} i^{-\delta-2-2\gamma} \\
&\approx \max(1, \kappa^{-\frac{1-\delta}{1+\gamma}}) + \kappa^{-2}\kappa^{\frac{\delta+1+2\gamma}{1+\gamma}} \\
&\approx \max(1, \kappa^{\frac{\delta-1}{1+\gamma}}).
\end{aligned}$$

To prove the fifth statement, observe that:

$$\begin{aligned}
\sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa} &= \sum_{i \le \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa} + \sum_{i \ge \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa} \\
&\approx \sum_{i \le \kappa^{-\frac{1}{1+\gamma}}} i^{-\delta} + \kappa^{-1} \sum_{i \ge \kappa^{-\frac{1}{1+\gamma}}} i^{-\delta-1-\gamma} \\
&\approx \max(1, \kappa^{-\frac{1-\delta}{1+\gamma}}) + \kappa^{-1}\kappa^{\frac{\delta+\gamma}{1+\gamma}} \\
&\approx \max(1, \kappa^{\frac{\delta-1}{1+\gamma}}).
\end{aligned}$$

To prove the sixth statement, observe that:

$$\begin{aligned}
\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2} &= \sum_{i \le \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2} + \sum_{i \ge \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\
&\approx \sum_{i \le \kappa^{-\frac{1}{1+\gamma}}} 1 + \kappa^{-2} \sum_{i \ge \kappa^{-\frac{1}{1+\gamma}}} i^{-2-2\gamma} \\
&\approx \kappa^{-\frac{1}{1+\gamma}} + \kappa^{-2}\kappa^{\frac{1+2\gamma}{1+\gamma}} \\
&\approx \kappa^{-\frac{1}{1+\gamma}}.
\end{aligned}$$

To prove the seventh statement, observe that:

$$\begin{aligned}\sum_{i=1}^{P} \frac{i^{-1-\gamma}}{i^{-1-\gamma}+\kappa} &= \sum_{i \leq \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-1-\gamma}}{i^{-1-\gamma}+\kappa} + \sum_{i \geq \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-1-\gamma}}{i^{-1-\gamma}+\kappa} \\ &\approx \sum_{i \leq \kappa^{-\frac{1}{1+\gamma}}} 1 + \kappa^{-1} \sum_{i \geq \kappa^{-\frac{1}{1+\gamma}}} i^{-1-\gamma} \\ &\approx \kappa^{-\frac{1}{1+\gamma}} + \kappa^{-1}\kappa^{\frac{\gamma}{1+\gamma}} \\ &\approx \kappa^{-\frac{1}{1+\gamma}}.\end{aligned}$$

To prove the eighth statement, observe that:

$$\begin{aligned}\sum_{i=1}^{P} \frac{i^{-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} &= \sum_{i \leq \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + \sum_{i \geq \kappa^{-\frac{1}{1+\gamma}}} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\ &\approx \sum_{i \leq \kappa^{-\frac{1}{1+\gamma}}} i^{1+\gamma} + \kappa^{-2} \sum_{i \geq \kappa^{-\frac{1}{1+\gamma}}} i^{-1-\gamma} \\ &\approx \max(1, \kappa^{-\frac{2+\gamma}{1+\gamma}}) + \kappa^{-2}\kappa^{\frac{\gamma}{1+\gamma}} \\ &= \kappa^{-2}\max(\kappa^2, \kappa^{\frac{\gamma}{1+\gamma}}) + \kappa^{-2}\kappa^{\frac{\gamma}{1+\gamma}} \\ &\approx \kappa^{-2}\max(\kappa^2, \kappa^{\frac{\gamma}{1+\gamma}}) \\ &\approx \kappa^{-2}\kappa^{\frac{\gamma}{1+\gamma}}\end{aligned}$$

□

Lemma 34. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, suppose that $P = \infty$. Assume the notation from Lemma 6, and similarly let*

$$Q := 1 - \frac{1}{N}\operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2}).$$

Then it holds that $Q^{-1} = \Theta(1)$.

Proof. Let $\Sigma = V\Lambda V^T$ be the eigendecomposition of Σ, where Λ is a diagonal matrix consisting of the eigenvalues. By Definition 2, we see that:

$$\frac{\lambda}{\kappa} + \frac{1}{N}\operatorname{Tr}(\Sigma\Sigma_\kappa^{-1}) = 1.$$

This implies that:

$$\begin{aligned}Q &= 1 - \frac{1}{N}\operatorname{Tr}(\Sigma\Sigma_\kappa^{-1}) + \frac{1}{N}\left(\operatorname{Tr}(\Sigma\Sigma_\kappa^{-1}) - \operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2})\right) \\ &= \frac{\lambda}{\kappa} + \frac{1}{N}\left(\operatorname{Tr}(\Sigma\Sigma_\kappa^{-1}) - \operatorname{Tr}(\Sigma^2\Sigma_\kappa^{-2})\right).\end{aligned}$$

Observe that:

$$\begin{aligned}\mathrm{Tr}(\Sigma\Sigma_\kappa^{-1}) - \mathrm{Tr}(\Sigma^2\Sigma_\kappa^{-2}) &= \mathrm{Tr}(\Lambda(\Lambda+\kappa I)^{-1}) - \mathrm{Tr}(\Lambda^2(\Lambda+\kappa I)^{-2}) \\ &= \sum_{i=1}^{P}\left(\frac{i^{-1-\gamma}}{i^{-1-\gamma}+\kappa} - \frac{i^{-2-\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \\ &= \kappa\sum_{i=1}^{P}\frac{i^{-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2}.\end{aligned}$$

This means that:

$$\begin{aligned}Q &= \frac{\lambda}{\kappa} + \frac{\kappa}{N}\sum_{i=1}^{P}\frac{i^{-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\ &\approx_{(A)} \frac{\lambda}{\kappa} + \Theta(\Big(\frac{\kappa}{N}\kappa^{-2}\kappa^{\frac{\gamma}{1+\gamma}}\Big)) \\ &= \frac{\lambda}{\kappa} + \Theta\left(\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}\right).\end{aligned}$$

where (A) uses Lemma 33.

Case 1: $\kappa = \Theta(\lambda)$. In this case, we see that

$$Q = \frac{\lambda}{\kappa} + \Theta\left(\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}\right) = \Theta(1).$$

This means that $Q^{-1} = \Theta(1)$.

Case 2: $\kappa = \Theta(N^{-1-\gamma})$. In this case, we see that

$$Q = \frac{\lambda}{\kappa} + \Theta\left(\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}\right) = \Omega\left(\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}\right) = \Omega(1).$$

This means that $Q^{-1} = \Theta(1)$.

□

Lemma 35. *Suppose that power-law scaling holds for the eigenvalues with scaling exponent γ, and suppose that $P = \infty$. Then it holds that $\kappa(\lambda, M, \Sigma) = \Theta(\max(\lambda, M^{-1-\gamma}))$.*

Proof. Let $\Sigma = V\Lambda V^T$ be the eigendecomposition of Σ, where Λ is a diagonal matrix consisting of the eigenvalues. Observe that:

$$\begin{aligned}\mathrm{Tr}((\Sigma+\kappa I)^{-1}\Sigma) &= \mathrm{Tr}(\Lambda(\Lambda+\kappa I)^{-1}) \\ &= \sum_{i=1}^{P}\frac{i^{-1-\gamma}}{i^{-1-\gamma}+\kappa} \\ &\approx_{(A)} \kappa^{-\frac{1}{1+\gamma}}.\end{aligned}$$

where (A) follows from Lemma 33. Using Definition 2, we see that for $\kappa = \kappa(\lambda, M, \Sigma)$, it holds that:

$$\frac{\lambda}{\kappa} + \frac{1}{M}\Theta(\kappa^{-1-\gamma}) = 1.$$

This implies that $\kappa = \Theta(\max(\lambda, M^{-1-\gamma}))$ as desired. □

D Machinery from random matrix theory

In this section, we introduce machinery from random matrix theory that serves as the backbone for our analysis of multi-objective scaling laws in Appendix C. In Appendix D.1, we give a recap of known Marčenko-Pastur properties. In Appendix D.2, we use these known properties to derive random matrix theory results which are tailored to our analysis.

D.1 Recap of Marčenko-Pastur properties

We introduce Marčenko-Pastur properties, following the treatment in Bach [2023]. Informally speaking, Marčenko-Pastur laws show that a random matrix $(\hat{\Sigma} + \lambda I)^{-1}$ (where $\hat{\Sigma}$ is a sample covariance) behaves similarly to a deterministic matrix of the form $(\hat{\Sigma} + \kappa I)^{-1}$, where $\kappa = \kappa(\lambda, M, \Sigma)$ is an *effective regularizer*.

Deriving this formally requires placing several structural assumptions on number of data points $N \geq 1$, the number of parameters $P \geq 1$, the distribution $\mathcal{D}_F$, and the vectors β_1 and β_2. We adopt assumptions from Bach [2023] which guarantee that a Marčenko-Pastur law holds for Σ, and we further introduce a boundedness assumption for technical reasons.

Assumption 1. *We assume that: (1) $X \sim \mathcal{D}_F$ takes the form $X = Z\Sigma^{1/2}$ where Z has bounded subgaussian i.i.d components with mean zero and unit variance, (2) N and P approach ∞ with $\frac{P}{N}$ tending to $\gamma > 0$, (3) the spectral measure $\frac{1}{P}\sum_{i=1}^{P} \delta_{\lambda_i}$ of Σ converges to a probability measure with compact support, and Σ is invertible and bounded in operator norm, and (4) for $j \in \{1, 2\}$, the measure $\sum_{i=1}^{P} \langle v_i, \beta_j \rangle^2$ converges to a measure with bounded mass, and β_j has bounded ℓ_2 norm.*

The effective regularizer $\kappa(\lambda, M, \Sigma)$ is defined as follows.

Definition 2 (Effective regularizer)**.** *For $\lambda \geq 0$, $M \geq 1$, and a P-dimensional positive semidefinite matrix Σ with eigenvalues λ_i for $1 \leq i \leq P$, the value $\kappa(\lambda, M, \Sigma)$ is the unique value $\kappa \geq 0$ such that:*

$$\frac{\lambda}{\kappa} + \frac{1}{N}\sum_{i=1}^{P} \frac{\lambda_i}{\lambda_i + \kappa} = 1.$$

We are now ready to state the key random matrix theory results proven in Bach [2023]. Following Bach [2023], the asymptotic equivalence notation $u \sim v$ means that u/v tends to 1 as N and P go to ∞.

Lemma 36 (Restatement of Proposition 1 in Bach [2023])**.** *Let $\hat{\Sigma} = \frac{1}{M}\sum_{i=1}^{M} X_i X_i^T$ be the sample covariance matrix from M i.i.d. samples from $X_1, \ldots, X_M \sim \mathcal{D}_F$. Let $\kappa = \kappa(\lambda, N, \Sigma)$. Suppose that A and B have bounded operator norm. Then it holds that:*

$$\lambda \operatorname{Tr}\left((\hat{\Sigma} + \lambda I)^{-1} A\right) \sim \kappa \operatorname{Tr}\left((\Sigma + \kappa I)^{-1} A\right)$$

$$\lambda^2 \operatorname{Tr}\left((\hat{\Sigma} + \lambda I)^{-1} A (\hat{\Sigma} + \lambda I)^{-1} B\right) \sim \kappa^2 \operatorname{Tr}\left((\Sigma + \kappa I)^{-1} A (\Sigma + \kappa I)^{-1} B\right)$$
$$+ \kappa^2 \frac{\frac{1}{N}\operatorname{Tr}\left(A\Sigma(\Sigma + \kappa I)^{-2}\right)}{1 - \frac{1}{N}\operatorname{Tr}\left(\Sigma^2(\Sigma + \kappa I)^{-2}\right)} \operatorname{Tr}\left((\Sigma + \kappa I)^{-1}\Sigma(\Sigma + \kappa I)^{-1} B\right).$$

We note that the requirement that B has bounded operator norm in Lemma 36 is what forces us to require that $\|\beta_1\|$ and $\|\beta_2\|$ are bounded. However, Wei et al. [2022] showed that the norm can be unbounded in several real-world settings, and thus instead opt to assume a local Marčenko-Pastur law and derive scaling laws based on this assumption. We suspect it may be possible to derive our scaling law with an appropriate analogue of the local Marčenko-Pastur law, which would also have the added benefit of allowing one to relax other requirements in Assumption 1 such as gaussianity. We view such an extension as an interesting direction for future work.

D.2 Useful random matrix theory facts

We derive several corollaries of Lemma 36 tailored to random matrices that arise in our analysis of multi-objective scaling laws.

Lemma 37. *Assume that $\mathcal{D}_F$ satisfies the Marčenko-Pastur property (Assumption 1). Let Z be a positive definite matrix such that Z^{-1} has bounded operator norm, and let A be a matrix with bounded operator norm. Let $\hat{\Sigma} = \frac{1}{M}\sum_{i=1}^{M} X_i X_i^T$ be the sample covariance matrix from M i.i.d. samples from $X_1, \ldots, X_M \sim \mathcal{D}_F$. Then it holds that:*

$$\lambda \cdot \mathrm{Tr}((\hat{\Sigma} + \lambda Z)^{-1} A) \sim \kappa \cdot \mathrm{Tr}((\Sigma + \kappa Z)^{-1} A). \tag{3}$$

If A also has bounded trace and Z has bounded operator norm, then it holds that:

$$\mathrm{Tr}(\hat{\Sigma}(\hat{\Sigma} + \lambda Z)^{-1} A) \sim \mathrm{Tr}(\Sigma \cdot (\Sigma + \kappa Z)^{-1} A) \tag{4}$$

where $\kappa = \kappa(\lambda, M, Z^{-1/2}\Sigma Z^{-1/2})$.

Proof. For (3), observe that:

$$\begin{aligned}
\lambda \cdot \mathrm{Tr}((\hat{\Sigma} + \lambda Z)^{-1} A) &= \lambda \cdot \mathrm{Tr}(Z^{-1/2}(Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I)^{-1} Z^{-1/2} A) \\
&= \lambda \cdot \mathrm{Tr}((Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I)^{-1} Z^{-1/2} A Z^{-1/2}) \\
&\sim_{(A)} \kappa \cdot \mathrm{Tr}((Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2} A Z^{-1/2}) \\
&= \kappa \cdot \mathrm{Tr}(Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2} A) \\
&= \kappa \cdot \mathrm{Tr}((\Sigma + \kappa Z)^{-1} A).
\end{aligned}$$

where (A) applies Lemma 36 (using the fact that since A and Z^{-1} have bounded operator norm, it holds that $Z^{-1/2} A Z^{-1/2}$ has bounded operator norm).

For (4), observe that:

$$\begin{aligned}
\mathrm{Tr}(\hat{\Sigma}(\hat{\Sigma} + \lambda Z)^{-1} A) &=_{(A)} \mathrm{Tr}\left(\left(I - \lambda Z^{1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I\right)^{-1} Z^{-1/2}\right) A\right) \\
&=_{(B)} \mathrm{Tr}(A) - \lambda \cdot \mathrm{Tr}\left(\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I\right)^{-1} Z^{-1/2} A Z^{1/2}\right) \\
&\sim_{(C)} \mathrm{Tr}(A) - \kappa \cdot \mathrm{Tr}\left(\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2} A Z^{1/2}\right) \\
&=_{(D)} \mathrm{Tr}\left(\left(I - \kappa Z^{1/2}\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2}\right) A\right) \\
&=_{(E)} \mathrm{Tr}(\Sigma(\Sigma + \kappa Z)^{-1} A)
\end{aligned}$$

where (A) and (E) follows from Claim 41, (B) and (D) use the fact that $\mathrm{Tr}(A)$ is bounded, and (C) follows from Lemma 36 (using the fact that since A, Z, and Z^{-1} have bounded operator norm, it holds that $Z^{-1/2} A Z^{1/2}$ has bounded operator norm).

□

Lemma 38. *Assume that $\mathcal{D}_F$ satisfies the Marčenko-Pastur property (Assumption 1). Let Z be any positive definite matrix such that Z and Z^{-1} have bounded operator norm, and let A and B have bounded operator norm. Let $\hat{\Sigma} = \frac{1}{M}\sum_{i=1}^{M} X_i X_i^T$ be the sample covariance matrix from M i.i.d. samples from $X_1, \ldots, X_M \sim \mathcal{D}_F$. Then it holds that:*

$$
\begin{aligned}
&\lambda^2 \operatorname{Tr}((\hat{\Sigma} + \lambda Z)^{-1} A (\hat{\Sigma} + \lambda Z)^{-1} B) \\
&= \lambda^2 \operatorname{Tr}(Z^{-1/2}(Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I)^{-1} Z^{-1/2} A Z^{-1/2} (Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I)^{-1} B) \\
&\sim \kappa^2 \operatorname{Tr}((\Sigma + \kappa Z)^{-1} A (\Sigma + \kappa Z)^{-1} B) \\
&+ \kappa^2 \frac{\frac{1}{M}\operatorname{Tr}((\Sigma + \kappa Z)^{-1}\Sigma(\Sigma + \kappa Z)^{-1} A)}{1 - \frac{1}{M}\operatorname{Tr}((\Sigma + \kappa Z)^{-1}\Sigma(\Sigma + \kappa Z)^{-1}\Sigma)} \operatorname{Tr}((\Sigma + \kappa Z)^{-1}\Sigma(\Sigma + \kappa Z)^{-1} B)
\end{aligned}
$$

where $\kappa = \kappa(\lambda, M, Z^{-1/2}\Sigma Z^{-1/2})$.

Proof. Let $q = \frac{\frac{1}{M}\operatorname{Tr}(Z^{-1/2}\Sigma Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-2} Z^{-1/2} A Z^{-1/2})}{1 - \frac{1}{M}\operatorname{Tr}(Z^{-1/2}\Sigma Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-2} Z^{-1/2}\Sigma Z^{-1/2})}$.

Observe that:

$$
\begin{aligned}
&\lambda^2 \operatorname{Tr}((\hat{\Sigma} + \lambda Z)^{-1} A (\hat{\Sigma} + \lambda Z)^{-1} B) \\
&\lambda^2 \operatorname{Tr}\left(Z^{-1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I\right)^{-1} Z^{-1/2} A Z^{-1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I\right)^{-1} Z^{-1/2} B\right) \\
&= \lambda^2 \operatorname{Tr}\left(\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I\right)^{-1} Z^{-1/2} A Z^{-1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2} + \lambda I\right)^{-1} Z^{-1/2} B Z^{-1/2}\right) \\
&\sim_{(A)} \kappa^2 \operatorname{Tr}\left(\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2} A Z^{-1/2}\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2} B Z^{-1/2}\right) \\
&+ \kappa^2 q \operatorname{Tr}\left(\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2}\Sigma Z^{-1/2}\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2} B Z^{-1/2}\right) \\
&= \kappa^2 \operatorname{Tr}\left(Z^{-1/2}\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2} A Z^{-1/2}\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2} B\right) \\
&+ \kappa^2 q \operatorname{Tr}\left(Z^{-1/2}\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2}\Sigma Z^{-1/2}\left(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2} B\right) \\
&= \kappa^2 \operatorname{Tr}\left((\Sigma + \kappa Z)^{-1} A (\Sigma + \kappa Z)^{-1} B\right) + q\kappa^2 \operatorname{Tr}\left((\Sigma + \kappa Z)^{-1}\Sigma(\Sigma + \kappa Z)^{-1} B\right),
\end{aligned}
$$

where (A) follows from Lemma 36 (using the fact that since A, B, Z, and Z^{-1} have bounded operator norm, it holds that $Z^{-1/2} A Z^{1/2}$, Σ, and $Z^{-1/2} B Z^{1/2}$ have bounded operator norm).

We can simplify q as follows:

$$
\begin{aligned}
q &= \frac{\frac{1}{M}\operatorname{Tr}(Z^{-1/2}\Sigma Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-2} Z^{-1/2} A Z^{-1/2})}{1 - \frac{1}{M}\operatorname{Tr}(Z^{-1/2}\Sigma Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-2} Z^{-1/2}\Sigma Z^{-1/2})} \\
&= \frac{\frac{1}{M}\operatorname{Tr}((Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2}\Sigma Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2} A Z^{-1/2})}{1 - \frac{1}{M}\operatorname{Tr}((Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2}\Sigma Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2}\Sigma Z^{-1/2})} \\
&= \frac{\frac{1}{M}\operatorname{Tr}(Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2}\Sigma Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2} A)}{1 - \frac{1}{M}\operatorname{Tr}(Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2}\Sigma Z^{-1/2}(Z^{-1/2}\Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2}\Sigma)} \\
&= \frac{\frac{1}{M}\operatorname{Tr}((\Sigma + \kappa Z)^{-1}\Sigma(\Sigma + \kappa Z)^{-1} A)}{1 - \frac{1}{M}\operatorname{Tr}((\Sigma + \kappa Z)^{-1}\Sigma(\Sigma + \kappa Z)^{-1}\Sigma)}.
\end{aligned}
$$

□

Lemma 39. *Assume that $\mathcal{D}_F$ satisfies the Marčenko-Pastur property (Assumption 1). Let Z be any positive definite matrix such that Z and Z^{-1} have bounded operator norm. Let A and B have bounded operator norm, and suppose also that $\operatorname{Tr}(AB)$ is bounded. Let $\hat{\Sigma} = \frac{1}{M}\sum_{i=1}^{M} X_i X_i^T$ be the sample covariance matrix from M i.i.d. samples from $X_1, \ldots, X_M \sim \mathcal{D}_F$. Then it holds that:*

$$\operatorname{Tr}(\hat{\Sigma}(\hat{\Sigma}+\lambda Z)^{-1}A(\hat{\Sigma}+\lambda Z)^{-1}\hat{\Sigma}B) \sim \operatorname{Tr}(\Sigma(\Sigma+\kappa Z)^{-1}A(\Sigma+\kappa Z)^{-1}\Sigma B) + E, \tag{5}$$

where:

$$E := \frac{\frac{1}{M}\operatorname{Tr}((\Sigma+\kappa Z)^{-1}\Sigma(\Sigma+\kappa Z)^{-1}A)}{1-\frac{1}{M}\operatorname{Tr}((\Sigma+\kappa Z)^{-1}\Sigma(\Sigma+\kappa Z)^{-1}\Sigma)} \cdot \kappa^2 \operatorname{Tr}\left((\Sigma+\kappa Z)^{-1}\Sigma\,(\Sigma+\kappa Z)^{-1} ZBZ\right),$$

and $\kappa = \kappa(\lambda, M, Z^{-1/2}\Sigma Z^{-1/2})$.

Proof. Observe that:

$$\begin{aligned}
&\operatorname{Tr}(\hat{\Sigma}(\hat{\Sigma}+\lambda Z)^{-1}A(\hat{\Sigma}+\lambda Z)^{-1}\hat{\Sigma}B)\\
&= \operatorname{Tr}(\hat{\Sigma}(\hat{\Sigma}+\lambda Z)^{-1}A\left(\hat{\Sigma}(\hat{\Sigma}+\lambda Z)^{-1}\right)^T B)\\
&=_{(A)} \operatorname{Tr}\left(\left(I-\lambda Z^{1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{-1/2}\right)A\left(I-\lambda Z^{1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{-1/2}\right)^T B\right)\\
&=_{(B)} \operatorname{Tr}(AB) - \lambda\operatorname{Tr}\left(A\left(Z^{1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{-1/2}\right)^T B\right)\\
&-\lambda\operatorname{Tr}\left(Z^{1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{-1/2}AB\right)\\
&+\lambda^2\operatorname{Tr}\left(Z^{1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{-1/2}A\left(Z^{1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{-1/2}\right)^T B\right)\\
&= \operatorname{Tr}(AB) - \lambda\operatorname{Tr}\left(AZ^{-1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{1/2}B\right)\\
&-\lambda\operatorname{Tr}\left(Z^{1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{-1/2}AB\right)\\
&+\lambda^2\operatorname{Tr}\left(Z^{1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{-1/2}AZ^{-1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{1/2}B\right)\\
&= \operatorname{Tr}(AB) - \underbrace{\lambda\operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda Z\right)^{-1}ZBA\right)}_{(1)} - \underbrace{\lambda\operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda Z\right)^{-1}ABZ\right)}_{(2)}\\
&+\underbrace{\lambda^2\operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda Z\right)^{-1}A\left(\hat{\Sigma}+\lambda Z\right)^{-1}ZBZ\right)}_{(3)}
\end{aligned}$$

where (A) follows from Claim 41, (B) uses that $\operatorname{Tr}(AB)$ is bounded,

For term (1) and term (2), we apply Lemma 37 to see that:

$$\begin{aligned}
\lambda\operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda Z\right)^{-1}ZBA\right) &\sim \kappa\lambda\operatorname{Tr}\left((\Sigma+\kappa Z)^{-1}ZBA\right)\\
\lambda\operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda Z\right)^{-1}ABZ\right) &\sim \kappa\operatorname{Tr}\left((\Sigma+\kappa Z)^{-1}ABZ\right).
\end{aligned}$$

For term (3), we apply Lemma 38 to see that

$$\begin{aligned}
&\lambda^2 \operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda Z\right)^{-1} A\left(\hat{\Sigma}+\lambda Z\right)^{-1} ZBZ\right)\\
&\sim \kappa^2 \operatorname{Tr}\left((\Sigma+\kappa Z)^{-1} A\,(\Sigma+\kappa Z)^{-1} ZBZ\right)\\
&+\kappa^2 \frac{\frac{1}{M}\operatorname{Tr}((\Sigma+\kappa Z)^{-1}\Sigma(\Sigma+\kappa Z)^{-1}A)}{1-\frac{1}{M}\operatorname{Tr}((\Sigma+\kappa Z)^{-1}\Sigma(\Sigma+\kappa Z)^{-1}\Sigma)} \operatorname{Tr}\left((\Sigma+\kappa Z)^{-1}\Sigma(\Sigma+\kappa Z)^{-1}ZBZ\right)\\
&\sim \kappa^2 \operatorname{Tr}\left((\Sigma+\kappa Z)^{-1} A\,(\Sigma+\kappa Z)^{-1} ZBZ\right)+E
\end{aligned}$$

This means that:

$$\begin{aligned}
\operatorname{Tr}\left(\hat{\Sigma}(\hat{\Sigma}+\lambda Z)^{-1}A(\hat{\Sigma}+\lambda Z)^{-1}\hat{\Sigma}\right) &\sim \operatorname{Tr}(AB)-\kappa\operatorname{Tr}\left((\Sigma+\kappa Z)^{-1} ZBA\right)-\kappa\operatorname{Tr}\left((\Sigma+\kappa Z)^{-1} ABZ\right)\\
&+\kappa^2 \operatorname{Tr}\left((\Sigma+\kappa Z)^{-1} A\,(\Sigma+\kappa Z)^{-1} ZBZ\right)+E\\
&=_{(C)} \operatorname{Tr}(\Sigma(\Sigma+\kappa Z)^{-1}A(\Sigma+\kappa Z)^{-1}\Sigma B)+E,
\end{aligned}$$

where (C) uses an analogous analysis to the beginning of the proof. □

Lemma 40. *Assume that $\mathcal{D}_F$ satisfies the Marčenko-Pastur property (Assumption 1). Let Z be any positive definite matrix such that Z and Z^{-1} have bounded operator norm, and let A and B have bounded operator norm. Let $\hat{\Sigma}=\frac{1}{M}\sum_{i=1}^{M} X_i X_i^T$ be the sample covariance matrix from M i.i.d. samples from $X_1,\ldots,X_M\sim\mathcal{D}_F$. Then it holds that:*

$$\lambda\operatorname{Tr}\left((\hat{\Sigma}+\lambda Z)^{-1}A(\hat{\Sigma}+\lambda Z)^{-1}\hat{\Sigma}B\right)\sim\kappa\operatorname{Tr}\left((\Sigma+\kappa Z)^{-1}A(\Sigma+\kappa Z)^{-1}\Sigma B\right)-E, \tag{6}$$

where:

$$E:=\frac{\frac{1}{M}\operatorname{Tr}((\Sigma+\kappa Z)^{-1}\Sigma(\Sigma+\kappa Z)^{-1}A)}{1-\frac{1}{M}\operatorname{Tr}((\Sigma+\kappa Z)^{-1}\Sigma(\Sigma+\kappa Z)^{-1}\Sigma)}\cdot\kappa^2\operatorname{Tr}\left((\Sigma+\kappa Z)^{-1}\Sigma\,(\Sigma+\kappa Z)^{-1}ZB\right)$$

and $\kappa=\kappa(\lambda,N,Z^{-1/2}\Sigma Z^{-1/2})$.

Proof. Observe that:

$$\begin{aligned}
&\lambda\operatorname{Tr}\left((\hat{\Sigma}+\lambda Z)^{-1}A(\hat{\Sigma}+\lambda Z)^{-1}\hat{\Sigma}B\right)\\
&=_{(A)}\lambda\operatorname{Tr}\left(Z^{-1/2}(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I)^{-1}Z^{-1/2}A\left(I-\lambda Z^{-1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{1/2}\right)B\right)\\
&=\lambda\operatorname{Tr}\left(Z^{-1/2}(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I)^{-1}Z^{-1/2}AB\right)\\
&-\lambda^2\operatorname{Tr}\left(Z^{-1/2}\left(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I\right)^{-1}Z^{-1/2}AZ^{-1/2}(Z^{-1/2}\hat{\Sigma}Z^{-1/2}+\lambda I)^{-1}Z^{1/2}B\right)\\
&=\underbrace{\lambda\operatorname{Tr}\left((\hat{\Sigma}+\lambda Z)^{-1}AB\right)}_{(1)}-\underbrace{\lambda^2\operatorname{Tr}\left(\left(\hat{\Sigma}+\lambda Z\right)^{-1}A(\hat{\Sigma}+\lambda Z)^{-1}ZB\right)}_{(2)}
\end{aligned}$$

where (A) follows from Claim 41.

For term (1), we apply Lemma 37 see that:

$$\lambda \operatorname{Tr}\left((\hat{\Sigma} + \lambda Z)^{-1} AB\right) \sim \kappa \operatorname{Tr}\left((\Sigma + \kappa Z)^{-1} AB\right).$$

For term (2), we apply Lemma 38 to see that

$$\begin{aligned}
&\lambda^2 \operatorname{Tr}\left(\left(\hat{\Sigma} + \lambda Z\right)^{-1} A(\hat{\Sigma} + \lambda Z)^{-1} ZB\right) \\
&\sim \kappa^2 \operatorname{Tr}\left((\Sigma + \kappa Z)^{-1} A \left(\Sigma + \kappa Z\right)^{-1} ZB\right) \\
&+ \kappa^2 \frac{\frac{1}{M} \operatorname{Tr}((\Sigma + \kappa Z)^{-1} \Sigma (\Sigma + \kappa Z)^{-1} A)}{1 - \frac{1}{M} \operatorname{Tr}((\Sigma + \kappa Z)^{-1} \Sigma (\Sigma + \kappa Z)^{-1} \Sigma)} \operatorname{Tr}\left((\Sigma + \kappa Z)^{-1} \Sigma (\Sigma + \kappa Z)^{-1} ZB\right) \\
&\sim \kappa^2 \operatorname{Tr}\left((\Sigma + \kappa Z)^{-1} A \left(\Sigma + \kappa Z\right)^{-1} ZB\right) + E.
\end{aligned}$$

This means that:

$$\begin{aligned}
&\lambda \operatorname{Tr}\left((\hat{\Sigma} + \lambda Z)^{-1} A(\hat{\Sigma} + \lambda Z)^{-1} \hat{\Sigma} B\right) \\
&\sim \kappa \operatorname{Tr}\left((\Sigma + \kappa Z)^{-1} AB\right) + \kappa^2 \operatorname{Tr}\left((\Sigma + \kappa Z)^{-1} A \left(\Sigma + \kappa Z\right)^{-1} ZB\right) - E \\
&= \kappa \operatorname{Tr}\left(Z^{-1/2} (Z^{-1/2} \Sigma Z^{-1/2} + \kappa I)^{-1} Z^{-1/2} A \left(I - \kappa Z^{1/2} \left(Z^{-1/2} \Sigma Z^{-1/2} + \kappa I\right)^{-1} Z^{-1/2}\right) B\right) - E \\
&=_{(A)} \kappa \operatorname{Tr}\left((\Sigma + \kappa Z)^{-1} A (\Sigma + \kappa Z)^{-1} \Sigma B\right) - E,
\end{aligned}$$

where (A) uses an analogous analysis to the beginning of the proof. □

The proofs of these results relied on the following basic matrix fact.

Claim 41. *Let A be any matrix and let B be any symmetric positive definite matrix. Then it holds that:*

$$A(A + \lambda B)^{-1} = I - \lambda B^{1/2} \left(B^{-1/2} A B^{-1/2} + \lambda I\right)^{-1} B^{-1/2}.$$

Proof. Observe that:

$$\begin{aligned}
&A(A + \lambda B)^{-1} \\
&= A B^{-1/2} \left(B^{-1/2} A B^{-1/2} + \lambda I\right)^{-1} B^{-1/2} \\
&= B^{1/2} \left(B^{-1/2} A B^{-1/2}\right) \left(B^{-1/2} A B^{-1/2} + \lambda I\right)^{-1} B^{-1/2} \\
&= B^{1/2} \left(B^{-1/2} A B^{-1/2} + \lambda I\right) \left(B^{-1/2} A B^{-1/2} + \lambda I\right)^{-1} B^{-1/2} - B^{1/2} \lambda \left(B^{-1/2} A B^{-1/2} + \lambda I\right)^{-1} B^{-1/2} \\
&= I - \lambda B^{1/2} \left(B^{-1/2} A B^{-1/2} + \lambda I\right)^{-1} B^{-1/2}.
\end{aligned}$$

□

E Extension: Market-entry threshold with richer form for L_2^*

In this section, we modify the safety requirement to take into account the impact of dataset size N and regularization parameter λ, and we extend our model and analysis of the market-entry threshold accordingly. We show that the characterization in Theorem 1 directly applies to this setting, and we also show relaxed versions of Theorem 4 and Theorem 5. Altogether, these extended results illustrate that our qualitative insights from Sections 3-4 hold more generally.

We define a modified approximation of the safety violation $\tilde{L}_2(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$. This modified approximation is defined analogously to $L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$. To formalize this, we define a deterministic equivalent $L_2^{\texttt{det}}$ for the safety violation to be

$$L_2^{\texttt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha) := L_1^{\texttt{det}}(\beta_2, \beta_1, \mathcal{D}_F, \lambda, N, 1-\alpha). \tag{7}$$

It follows from Lemma 6 that $L_2(\hat{\beta}(\alpha, \lambda, X)) \sim L_2^{\texttt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$: here, we use the fact that $L_2(\hat{\beta}(\alpha, \lambda, X))$ is distributed identically to $L_1(\hat{\beta}(1-\alpha, \lambda, X))$. Now, using this deterministic equivalent, we define $\tilde{L}_2(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha) = L_2^{\texttt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$.

Using this formulation of $\tilde{L}_2$, we define a modified market entry threshold where we replace all instances of original approximation L_2^* with the modified approximation $\tilde{L}_2$. In particular, a company C faces reputational damage if:

$$\mathbb{E}_{(\beta_1,\beta_2)\sim\mathcal{D}_W} \tilde{L}_2(\beta_1, \beta_2, \mathcal{D}_F, \alpha_C) \geq \tau_C.$$

The company selects $\alpha \in [0.5, 1]$ and $\lambda \in (0, 1)$ to maximize their performance subject to their safety constraint, as formalized by the following optimization program:[12]

$$(\tilde{\alpha}_C, \tilde{\lambda}_C) = \operatorname{argmin}_{\alpha\in[0.5,1],\lambda\in(0,1)} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N_C, \alpha)] \text{ s.t. } \mathbb{E}_{\mathcal{D}_W}[\tilde{L}_2(\beta_1, \beta_2, \mathcal{D}_F, \alpha)] \leq \tau_C.$$

We define the modified market-entry threshold as follows.

Definition 3. *The modified market-entry threshold $\tilde{N}_E^*(N_I, \tau_I, \tau_E, \mathcal{D}_W, \mathcal{D}_F)$ is the minimum value of $N_E \in \mathbb{Z}_{\geq 1}$ such that $\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_E, N_E, \tilde{\alpha}_E)] \leq \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)]$.*

In this section, we analyze the modified market entry threshold $\tilde{N}_E^*(N_I, \tau_I, \tau_E, \mathcal{D}_W, \mathcal{D}_F)$. We show an extension of Theorem 1 (Appendix E.1). We then derive a simplified version of the deterministic equivalent $L_2^{\texttt{det}}$ (Appendix E.2). Finally, we show a weakened extension of Theorem 4 (Appendix E.3) and a weakened extension of Theorem 5 (Appendix E.4). These weakened extensions derive upper bounds (rather than tight bounds) on the modified market entry threshold, and also assume that $\delta \leq 1$.

E.1 Extension of Theorem 1

We study the market entry $\tilde{N}_E^*$ threshold in the environment of Theorem 1 where the incumbent has infinite data and the new company faces no safety constraint. We show that the modified market entry threshold takes the same form as the market entry threshold in Theorem 1.

Theorem 42 (Extension of Theorem 1)**.** *Suppose that power-law scaling holds for the eigenvalues and alignment coefficients, with scaling exponents $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Suppose that the incumbent company has infinite data (i.e., $N_I = \infty$), and that the entrant faces no constraint on their safety (i.e., $\tau_E = \infty$). Suppose that the safety constraint τ_I satisfies (1). Then, it holds that:*

$$\tilde{N}_E^*(\infty, \tau_I, \infty, \mathcal{D}_W, \mathcal{D}_F) = \Theta\left(\left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^{-2/\nu}\right),$$

where $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)] = \Theta(1 - \rho)$, and where $\nu := \min(2(1+\gamma), \delta + \gamma)$.

[12] Unlike in Section 2, there might not exist $\alpha \in [0.5, 1]$ and $\lambda \in (0, 1)$ which satisfy the safety constraint, if N_C is too small.

Theorem 42 shows that the qualitative insights from Theorem 1—including that the new company can enter with finite data—readily extend to this setting.

To prove Theorem 42, we build on the notation and analysis from Appendix A. It suffices to show that each company C will select $\alpha_C = \tilde{\alpha}_C$ and $\lambda_C = \tilde{\lambda}_C$. This follows trivially for the entrant $C = E$ since they face no safety constraint, and there is no different between the two settings. The key ingredient of the proof is to compute $\tilde{\alpha}_I$ and $\tilde{\lambda}_I$ for the incumbent (i.e., an analogue of Lemma 12 in Appendix A).

To do this, we first upper bound the following function of the safety loss and performance loss for general parameters λ and α.

Lemma 43. *For any α and λ, it holds that:*

$$\sqrt{\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha,\lambda))]} + \sqrt{\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha,\lambda))]} \geq \sqrt{\mathbb{E}_{\mathcal{D}_W}[(\beta_1-\beta_2)^T\Sigma(\beta_1-\beta_2)^T]}.$$

Proof. Note that:

$$\begin{aligned}
T &:= \sqrt{\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha,\lambda))]} + \sqrt{\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha,\lambda))]} \\
&= \sqrt{(\beta_1-\beta(\alpha,\lambda))^T\Sigma(\beta_1-\beta(\alpha,\lambda))} + \sqrt{(\beta_2-\beta(\alpha,\lambda))^T\Sigma(\beta_2-\beta(\alpha,\lambda))} \\
&= \sqrt{(\lambda\beta_1+(1-\alpha)\Sigma(\beta_1-\beta_2))^T\Sigma(\Sigma+\lambda I)^{-2}(\lambda\beta_1+(1-\alpha)\Sigma(\beta_1-\beta_2))} \\
&+ \sqrt{(\lambda\beta_2+\alpha\Sigma(\beta_2-\beta_1))^T\Sigma(\Sigma+\lambda I)^{-2}(\lambda\beta_2+\alpha\Sigma(\beta_2-\beta_1))} \\
&= \sqrt{(\lambda\beta_1+(1-\alpha)\Sigma(\beta_1-\beta_2))^T\Sigma(\Sigma+\lambda I)^{-2}(\lambda\beta_1+(1-\alpha)\Sigma(\beta_1-\beta_2))} \\
&+ \sqrt{(-\lambda\beta_2+\alpha\Sigma(\beta_1-\beta_2))^T\Sigma(\Sigma+\lambda I)^{-2}(-\lambda\beta_2+\alpha\Sigma(\beta_1-\beta_2))}.
\end{aligned}$$

Now note that for any PSD matrix Σ' and any distribution, note that the following triangle inequality holds:

$$\sqrt{\mathbb{E}[(X_1+X_2)^T\Sigma'(X_1+X_2)]} \leq \sqrt{\mathbb{E}[X_1^T\Sigma' X_1]} + \sqrt{\mathbb{E}[X_2^T\Sigma' X_2]}.$$

We apply this for $X_1 = \lambda\beta_1 + (1-\alpha)\Sigma(\beta_1-\beta_2)$, $X_2 = -\lambda\beta_2 + \alpha\Sigma(\beta_1-\beta_2)$, and distribution $\mathcal{D}_W$. This means that we can lower bound:

$$\begin{aligned}
T &\geq \sqrt{\mathbb{E}_{\mathcal{D}_W}[((\Sigma+\lambda I)(\beta_1-\beta_2))^T\Sigma(\Sigma+\lambda I)^{-2}((\Sigma+\lambda I)(\beta_1-\beta_2))]} \\
&= \sqrt{\mathbb{E}_{\mathcal{D}_W}[(\beta_1-\beta_2))^T\Sigma(\beta_1-\beta_2))]}
\end{aligned}$$

as desired. □

Now, we are ready to compute $\tilde{\alpha}_I$ and $\tilde{\lambda}_I$ for the incumbent.

Lemma 44. *Let $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1-\beta_2)^T\Sigma(\beta_1-\beta_2)^T]$. Suppose that $N_I = \infty$, and suppose that the safety constraint τ_I satisfies (1). Then it holds that $\alpha_I = \sqrt{\frac{\min(\tau_I, L^*(\rho))}{L^*(\rho)}}$, and $\lambda_I = 0$ is optimal for the incumbent. Moreover, it holds that:*

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1,\beta_2,\mathcal{D}_F,\tilde{\lambda}_I,\infty,\tilde{\alpha}_I)] = \left(\sqrt{L^*(\rho)} - \sqrt{\min(L^*(\rho),\tau_I)}\right)^2.$$

Proof. First, we apply Lemma 46 with $N = \infty$ to see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, \infty, \alpha)] = \mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, \lambda))]$$

and

$$\mathbb{E}_{\mathcal{D}_W}[L_2^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, \infty, \alpha)] = \mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha, \lambda))].$$

Let $\alpha^* = \sqrt{\frac{\min(\tau_I, L^*(\rho))}{L^*(\rho)}}$. By the assumption in the lemma statement, we know that:

$$\alpha^* \geq \sqrt{\frac{\mathbb{E}_{\mathcal{D}_W}[L_2^*(\beta_1, \beta_2, \mathcal{D}_F, 0.5)]}{L^*(\rho)}} = 0.5.$$

Observe that:

$$\begin{aligned}
&\sqrt{\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha^*, 0))]} + \sqrt{\min(\tau_I, L^*(\rho))} \\
&= \sqrt{\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha^*, 0))]} + \sqrt{\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha^*, 0))]} \\
&= \sqrt{(1-\alpha^*)^2 \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)^T]} + \sqrt{(\alpha^*)^2 \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)^T]} \\
&= \sqrt{\mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)^T]}
\end{aligned}$$

We show that $(\tilde{\alpha}_I, \tilde{\lambda}_I) = (\alpha^*, 0)$. Assume for sake of contradiction that $(\alpha, \lambda) \neq (\alpha^*, 0)$ satisfies the safety constraint $\mathbb{E}_{\mathcal{D}_W}[\tilde{L}_2(\beta_1, \beta_2, \mathcal{D}_F, \alpha)] \leq \min(\tau_I, L^*(\rho))$ and achieves strictly better performance loss:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, \infty, \alpha)] < \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, 0, \infty, \alpha^*)].$$

Then it would hold that:

$$\begin{aligned}
\sqrt{\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha, \lambda))]} + \sqrt{\mathbb{E}_{\mathcal{D}_W}[L_2(\beta(\alpha, \lambda))]} &< \sqrt{\mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\alpha^*, 0))]} + \sqrt{\min(\tau_I, L^*(\rho))} \\
&= \sqrt{\mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)^T]},
\end{aligned}$$

which contradicts Lemma 43.

To analyze the loss, note that:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, \infty, \tilde{\alpha}_I)] \\
&\qquad = \mathbb{E}_{\mathcal{D}_W}[L_1(\beta(\tilde{\alpha}_I, \tilde{\lambda}_I))] \\
&\qquad = (1 - \tilde{\alpha}_I)^2 L^*(\rho) \\
&\qquad = (\sqrt{L^*(\rho)} - \sqrt{\min(L^*(\rho), \tau_I)})^2
\end{aligned}$$

□

We now prove Theorem 42.

Proof of Theorem 42. We analyze $(\tilde{\alpha}_C, \tilde{\lambda}_C)$ first for the incumbent $C = I$ and then for the entrant $C = E$.

Analysis of the incumbent $C = I$. By Lemma 44, we see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, \infty, \tilde{\alpha}_I)] = \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^2.$$

Analysis of the entrant $C = E$. This analysis follows identically to the analogous case in the proof of Theorem 1, and we repeat the proof for completeness. Since the entrant faces no safety constraint, the entrant can choose any $\alpha \in [0.5, 1]$. We apply Corollary 8 to see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda_E, N, \alpha_E)] = \inf_{\alpha \in [0.5,1]} \inf_{\lambda > 0} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)] = \Theta\left(N^{-\nu}\right),$$

which means that:

$$N_E^*(\infty, \tau_I, \infty, \mathcal{D}_W, \mathcal{D}_F) = \Theta\left(\left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho)}\right)^{-2/\nu}\right)$$

as desired. We can further apply Claim 13 to see that $L^*(\rho) = \Theta(1-\rho)$. □

E.2 Bounds on the excess loss for safety

We bound the excess loss $\alpha^2 L^*(\rho) - \mathbb{E}_{\mathcal{D}_W}[L_2^{\mathsf{det}}]$. We assume that $\alpha \geq 0.5$ and we further assume that $\delta \leq 1$.

Lemma 45. *Suppose that power scaling holds for the eigenvalues and alignment coefficients with scaling $\gamma > 0$ and $\delta \in (0, 1]$, and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Suppose that $\alpha \geq 0.5$, $\lambda \in (0, 1)$, and $N \geq 1$. Let $L_2^{\mathsf{det}} := L_2^{\mathsf{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$ be defined according to (7). Let $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)]$. Then it holds that:*

$$\alpha^2 L^*(\rho) - \mathbb{E}_{\mathcal{D}_W}[L_2^{\mathsf{det}}] = O\left(\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu})\right)$$

and

$$\mathbb{E}_{\mathcal{D}_W}[L_2^{\mathsf{det}}] - \alpha^2 L^*(\rho) = O\left(\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) + (1-\alpha)(1-\rho)\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right),$$

where $\nu = \min(2(1+\gamma), \delta + \gamma) = \delta + \gamma$.

To prove Lemma 45, we first simplify the deterministic equivalent $L_2^{\mathsf{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$ using the assumptions from Section 2.3.

Lemma 46. *Suppose that power scaling holds for the eigenvalues and alignment coefficients with scaling $\gamma, \delta > 0$ and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Suppose that $\lambda \in (0, 1)$, and $N \geq 1$. Let $L_2^{\mathsf{det}} := L_2^{\mathsf{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)$ be defined according to (7). Let $\kappa = \kappa(\lambda, N, \Sigma)$ from Definition 2. Let $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)]$. Then it holds that:*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_W}[L_2^{\mathsf{det}}] - L^*(\rho) = Q^{-1} \cdot \kappa^2 \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + Q^{-1} 2\kappa\alpha(1-\alpha)(1-\rho) \sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\ + Q^{-1} 2\alpha(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\ - 2\alpha^2\kappa(1-\rho)\sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa},\end{aligned}$$

where $Q = 1 - \frac{1}{N}\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}$.

Proof. First, we apply Lemma 26, coupled with the fact that $L_2^{\mathtt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha) := L_1^{\mathtt{det}}(\beta_2, \beta_1, \mathcal{D}_F, \lambda, N, 1-\alpha)$, to see that:

$$\begin{aligned}
Q \cdot \mathbb{E}_{\mathcal{D}_W}[L_2^{\mathtt{det}}] &= \kappa^2(1-2\alpha^2(1-\rho))\sum_{i=1}^{P}\frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + \alpha^2 L^*(\rho) \\
&\quad + 2\kappa(1-\rho)\alpha(1-2\alpha)\sum_{i=1}^{P}\frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&\quad + 2\alpha(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\cdot(1-2\alpha)\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa},
\end{aligned}$$

where $Q = 1 - \frac{1}{N}\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}$. Using that $(Q^{-1}-1)\alpha^2 L^*(\rho) = Q^{-1}\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)2\alpha^2(1-\rho)\left(\sum_{i=1}^{P} i^{-\delta-1-\gamma}\right)$, this means that:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_W}[L_2^{\mathtt{det}}] - \alpha^2 L^*(\rho) &= Q^{-1}\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)2\alpha^2(1-\rho)\left(\sum_{i=1}^{P} i^{-\delta-1-\gamma}\right) \\
&\quad + Q^{-1}\cdot\kappa^2(1-2\alpha^2(1-\rho))\sum_{i=1}^{P}\frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\
&\quad + Q^{-1}2\kappa(1-\rho)\alpha(1-2\alpha)\sum_{i=1}^{P}\frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&\quad + Q^{-1}2\alpha(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\cdot(1-2\alpha)\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa}
\end{aligned}$$

By expanding some of these terms, we see that:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_W}[L_2^{\mathtt{det}}] - \alpha^2 L^*(\rho) &= Q^{-1}2\alpha^2(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\cdot\sum_{i=1}^{P} i^{-\delta-1-\gamma} \\
&\quad + Q^{-1}\cdot\kappa^2\sum_{i=1}^{P}\frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} - Q^{-1}2\alpha^2(1-\rho)\cdot\kappa^2\sum_{i=1}^{P}\frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\
&\quad + Q^{-1}2\kappa(1-\rho)\alpha(1-\alpha)\sum_{i=1}^{P}\frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} - Q^{-1}2\kappa(1-\rho)\alpha^2\sum_{i=1}^{P}\frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&\quad + Q^{-1}2\alpha(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\cdot\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\
&\quad - Q^{-1}2\alpha^2(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P}\frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right)\cdot\sum_{i=1}^{P}\frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa}.
\end{aligned}$$

When we collect terms, we obtain:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_W}[L_2^{\texttt{det}}] - \alpha^2 L^*(\rho) &= Q^{-1} \cdot \kappa^2 \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + Q^{-1} 2\kappa(1-\rho)\alpha(1-\alpha) \sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ Q^{-1} 2\alpha(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\
&- Q^{-1} 2\kappa(1-\rho)\alpha^2 \left(\sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} + \sum_{i=1}^{P} \frac{\kappa \cdot i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \\
&+ Q^{-1} 2\alpha^2(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot \left(\sum_{i=1}^{P} i^{-\delta-1-\gamma} - \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa}\right) \\
&= Q^{-1} \cdot \kappa^2 \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} + Q^{-1} 2\kappa(1-\rho)\alpha(1-\alpha) \sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ Q^{-1} 2\alpha(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\
&- Q^{-1} 2\kappa(1-\rho)\alpha^2 \left(\sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)}\right) \\
&+ Q^{-1} 2\kappa\alpha^2(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot \frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa}.
\end{aligned}$$

Combining the last two terms gives us the desired statement. □

Now, we are ready to prove Lemma 45.

Proof. For the first bound, we observe that:

$$\begin{aligned}
&\alpha^2 L^*(\rho) - \mathbb{E}_{\mathcal{D}_W}[L_2^{\texttt{det}}] \\
&\leq_{(A)} 2\alpha^2\kappa(1-\rho) \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{i^{-1-\gamma}+\kappa} \\
&=_{(B)} O\left(\alpha^2(1-\rho)\kappa^{\frac{\min(1+\gamma,\delta+\gamma)}{1+\gamma}}\right) \\
&=_{(C)} O\left(\kappa^{\frac{\nu}{1+\gamma}}\right) \\
&=_{(D)} O\left(\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu})\right)
\end{aligned}$$

where (A) uses Lemma 46, (B) uses Lemma 33, (C) uses that $\delta \leq 1$ and $\rho \in [0,1)$, and (D) uses Lemma 35.

For the second bound, we observe that:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_2^{\mathtt{det}}] - \alpha^2 L^*(\rho) \\
&\leq_{(A)} Q^{-1} \cdot \kappa^2 \sum_{i=1}^{P} \frac{i^{-\delta-1-\gamma}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ Q^{-1} 2\kappa\alpha(1-\alpha)(1-\rho) \sum_{i=1}^{P} \frac{i^{-\delta-2(1+\gamma)}}{(i^{-1-\gamma}+\kappa)^2} \\
&+ Q^{-1} 2\alpha(1-\alpha)(1-\rho)\frac{1}{N}\left(\sum_{i=1}^{P} \frac{i^{-2-2\gamma}}{(i^{-1-\gamma}+\kappa)^2}\right) \cdot \sum_{i=1}^{P} \frac{i^{-\delta-2-2\gamma}}{i^{-1-\gamma}+\kappa} \\
&=_{(B)} O\left(\kappa^{\frac{\min(2(1+\gamma),\gamma+\delta)}{1+\gamma}} + \alpha(1-\alpha)(1-\rho)\kappa^{\frac{\min(1+\gamma,\gamma+\delta)}{1+\gamma}} + \alpha(1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}\right) \\
&=_{(C)} O\left(\kappa^{\frac{\gamma+\delta}{1+\gamma}} + (1-\alpha)(1-\rho)\kappa^{\frac{\gamma+\delta}{1+\gamma}} + (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}\right) \\
&= O\left(\kappa^{\frac{\gamma+\delta}{1+\gamma}} + (1-\alpha)(1-\rho)\frac{\kappa^{-\frac{1}{1+\gamma}}}{N}\right) \\
&=_{(D)} O\left(\max(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}) + (1-\alpha)(1-\rho)\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N)}{N}\right)
\end{aligned}
$$

where (A) uses Lemma 46, (B) uses Lemma 33 and Lemma 34, (C) uses that $\delta \leq 1$ and $\alpha \geq 0.5$, and (D) uses Lemma 35. □

E.3 Extension of Theorem 4

We next study the market entry $\tilde{N}_E^*$ threshold in the environment of Theorem 4 where the incumbent has *finite data* and the new company faces no safety constraint. We place the further assumption that $\delta \leq 1$. We compute the following upper bound on the modified market entry threshold.

Theorem 47 (Extension of Theorem 4)**.** *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma > 0, \delta \in (0, 1]$ and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Assume that $\tau_E = \infty$. Suppose that the safety constraint τ_I satisfies (1). Then we have that $\tilde{N}_E^* = \tilde{N}_E^*(N_I, \tau_I, \infty, \mathcal{D}_W, \mathcal{D}_F)$ satisfies:*

$$
\tilde{N}_E^* := \begin{cases}
O\left(N_I\right) & \textit{if } N_I \leq \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \\
O\left(N_I^{\frac{1}{\nu+1}} \cdot \tilde{G}_I^{-\frac{1}{2(\nu+1)}}(1-\rho)^{-\frac{1}{2(\nu+1)}}\right) & \textit{if } \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \leq N_I \leq \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}} \\
O\left(\tilde{G}_I^{-\frac{1}{\nu}}\right) & \textit{if } N_I \geq \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}},
\end{cases}
$$

where $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma(\beta_1 - \beta_2)] = \Theta(1-\rho)$, where $\alpha^* = \sqrt{\frac{\min(\tau_I, L^*(\rho))}{L^*(\rho)}}$, where $\tilde{\alpha} := \sqrt{(1-\alpha^*) + (\alpha^*)^2}$, where $\tilde{G}_I = (1-\tilde{\alpha})^2(1-\rho)$, and where $\nu = \min(2(1+\gamma), \gamma+\delta) = \gamma+\delta$.

Theorem 47 shows that the key qualitative finding from Theorem 4—that the new company can enter with $N_E = o(N_I)$ data as long as the incumbent's dataset size is sufficiently large—readily extends to this setting. We note that the bound in Theorem 47 and the bound in Theorem 4 take slightly different forms: the term $G_I = (\sqrt{L^*(\rho)} - \sqrt{\min(L^*(\rho), \tau_I)})^2 = \Theta((1-\alpha^*)^2(1-\rho))$ is replaced by $\tilde{G}_I = (1-\tilde{\alpha})^2(1-\rho)$. We expect some of these differences arise because the bound in Theorem 47 is not tight, rather than fundamental distinctions between the two settings. Proving a tight bound on the modified market entry threshold is an interesting direction for future work.

To prove this, we compute a lower bound on the incumbent's loss $\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)]$.

Lemma 48. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma > 0, \delta \in (0, 1]$ and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Assume that $\tau_E = \infty$. Suppose that the safety constraint τ_I satisfies (1). Then we have that:*

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)]$$
$$= \begin{cases} \Omega\left(N_I^{-\nu}\right) & \text{if } N_I \leq \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \\ \Omega\left(N_I^{-\frac{\nu}{\nu+1}} \cdot \tilde{G}_I^{\frac{\nu}{2(\nu+1)}}(1-\rho)^{\frac{\nu}{2(\nu+1)}}\right) & \text{if } \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \leq N_I \leq \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}} \\ \Omega\left(\tilde{G}_I\right) & \text{if } N_I \geq \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}}. \end{cases}$$

where $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta_2)] = \Theta(1-\rho)$, where $\alpha^* = \sqrt{\frac{\min(\tau_I, L^*(\rho))}{L^*(\rho)}}$, where $\tilde{\alpha} := \sqrt{(1-\alpha^*) + (\alpha^*)^2}$, where $\tilde{G}_I = (1-\tilde{\alpha})^2(1-\rho)$ and where $\nu = \min(2(1+\gamma), \gamma+\delta) = \gamma+\delta$.

Proof. By Corollary 8 and Lemma 35, we know that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)] = \Omega(\kappa^{\frac{\nu}{1+\gamma}}) = \Omega(\max(\lambda^{\frac{\nu}{1+\gamma}}, N_I^{-\nu})).$$

Let $C_{\delta,\gamma}$ be an implicit constant[13] such that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)] \geq C_{\delta,\gamma} \max(\lambda^{\frac{\nu}{1+\gamma}}, N_I^{-\nu}) \tag{8}$$

By Lemma 45, there also exists an implicit constant $C'_{\delta,\gamma}$ such that:

$$\alpha^2 L^*(\rho) - \mathbb{E}_{\mathcal{D}_W}[L_2^{\texttt{det}}(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N_I, \alpha)] \leq C'_{\delta,\gamma} \max(\lambda^{\frac{\nu}{1+\gamma}}, N_I^{-\nu}). \tag{9}$$

We now split into two cases: (1) $\frac{C'_{\delta,\gamma}}{C_{\delta,\gamma}} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)] \geq (1-\alpha^*)L^*(\rho)$, and (2) $\frac{C'_{\delta,\gamma}}{C_{\delta,\gamma}} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)] \leq (1-\alpha^*)L^*(\rho)$.

Case 1: $\frac{C'_{\delta,\gamma}}{C_{\delta,\gamma}} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)] \geq (1-\alpha^*)L^*(\rho)$. It follows from (8) that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)] \geq C_{\delta,\gamma} \max(\lambda^{\frac{\nu}{1+\gamma}}, N_I^{-\nu}) \geq C_{\delta,\gamma} N_I^{-\nu}.$$

Using the condition for this case, this implies that:

$$\begin{aligned} N_I &\leq \left(\frac{1}{C_{\delta,\gamma}} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)]\right)^{-\frac{1}{\nu}} \\ &\leq \left(\frac{1}{C'_{\delta,\gamma}}(1-\alpha^*)L^*(\rho)\right)^{-\frac{1}{\nu}} \\ &= O\left(((1-\tilde{\alpha})(1-\rho))^{-\frac{1}{\nu}}\right) \\ &= O\left(\tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}}\right). \end{aligned}$$

[13] We need to introduce an implicit constant because of $O()$ is permitted to hide constants that depend on δ and γ.

This proves that N_I is up to constants within the first branch of the expression in the lemma statement. Since the bound in the lemma statement only changes by constants (that depend on δ and γ) between the first branch and second branch, this proves the desired expression for this case.

Case 2: $\frac{C'_{\delta,\gamma}}{C_{\delta,\gamma}}\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1,\beta_2,\mathcal{D}_F,\tilde{\lambda}_I,N_I,\tilde{\alpha}_I)] \le (1-\alpha^*)L^*(\rho)$**.** Note that $\alpha^* = \sqrt{\frac{\min(\tau_I,L^*(\rho))}{L^*(\rho)}}$ is the mixture parameter that achieves the safety constraint in the infinite-data ridgeless setting. The incumbent's safety constraint means that:

$$\mathbb{E}_{\mathcal{D}_W}[L_2^{\texttt{det}}(\beta_1,\beta_2,\mathcal{D}_F,\tilde{\lambda}_I,N_I,\tilde{\alpha}_I)] \le (\alpha^*)^2 L^*(\rho).$$

By (9), this implies that

$$(\tilde{\alpha}_I)^2 L^*(\rho) \le C'_{\delta,\gamma}\cdot\max(\lambda^{\frac{\delta+\gamma}{1+\gamma}}, N_I^{-\delta-\gamma}) + (\alpha^*)^2 L^*(\rho).$$

Now, applying (8) and the assumption for this case, we see that:

$$\begin{aligned}(\tilde{\alpha}_I)^2 L^*(\rho) &\le \frac{C'_{\delta,\gamma}}{C_{\delta,\gamma}}\cdot\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1,\beta_2,\mathcal{D}_F,\tilde{\lambda}_I,N_I,\tilde{\alpha}_I)] + (\alpha^*)^2 L^*(\rho)\\ &\le (1-\alpha^*)L^*(\rho) + (\alpha^*)^2 L^*(\rho).\end{aligned}$$

This implies that:

$$\tilde{\alpha}_I \le \sqrt{(1-\alpha^*)+(\alpha^*)^2}.$$

Let $\tilde{\alpha} := \sqrt{(1-\alpha^*)+(\alpha^*)^2}$. Plugging this into Corollary 8, we see that:

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1,\beta_2,\mathcal{D}_F,\tilde{\lambda}_I,N_I,\tilde{\alpha}_I)]\\ &\ge \inf_{\alpha\in[0.5,\tilde{\alpha}]}\inf_{\lambda>0}\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1,\beta_2,\mathcal{D}_F,\lambda,N_I,\alpha)]\\ &= \Theta\left(\inf_{\lambda>0}\mathbb{E}_{\mathcal{D}_W}\left[L_1^*\left(\beta_1,\beta_2,\Sigma,\lambda,N_I,\tilde{\alpha}\right)\right]\right)\\ &= \begin{cases}\Theta\left(N_I^{-\nu}\right) & \text{if } N_I \le (1-\tilde{\alpha})^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}}\\ \Theta\left(\left(\frac{N_I}{(1-\tilde{\alpha})(1-\rho)}\right)^{-\frac{\nu}{\nu+1}}\right) & \text{if } (1-\tilde{\alpha})^{-\frac{1}{\nu}}(1-\rho)^{-\frac{1}{\nu}} \le N_I \le (1-\tilde{\alpha})^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}}\\ \Theta((1-\tilde{\alpha})^2(1-\rho)) & \text{if } N_I \ge (1-\tilde{\alpha})^{-\frac{2+\nu}{\nu}}(1-\rho)^{-\frac{1}{\nu}},\end{cases}\\ &= \begin{cases}\Theta\left(N_I^{-\nu}\right) & \text{if } N_I \le \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}}\\ \Theta\left(N_I^{-\frac{\nu}{\nu+1}}\cdot\tilde{G}_I^{\frac{\nu}{2(\nu+1)}}(1-\rho)^{\frac{\nu}{2(\nu+1)}}\right) & \text{if } \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \le N_I \le \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}}\\ \Theta\left(\tilde{G}_I\right) & \text{if } N_I \ge \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}}.\end{cases}\end{aligned}$$

The statement follows in this case. □

We are now ready to prove Theorem 47.

Proof of Theorem 47. We analyze $(\tilde{\alpha}_C,\tilde{\lambda}_C)$ first for the incumbent $C = I$ and then for the entrant $C = E$. Like in the theorem statement, let $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1-\beta_2)^T\Sigma(\beta_1-\beta_2)] = \Theta(1-\rho)$ (Claim 13) and $G_I := (\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))})^2$, and $\nu = \min(2(1+\gamma),\delta+\gamma)$.

Analysis of the incumbent $C = I$. We apply Lemma 48 to see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, N_I, \tilde{\alpha}_I)]$$
$$= \begin{cases} \Omega\left(N_I^{-\nu}\right) & \text{if } N_I \le \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \\ \Omega\left(N_I^{-\frac{\nu}{\nu+1}} \cdot \tilde{G}_I^{\frac{\nu}{2(\nu+1)}}(1-\rho)^{\frac{\nu}{2(\nu+1)}}\right) & \text{if } \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \le N_I \le \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}} \\ \Omega\left(\tilde{G}_I\right) & \text{if } N_I \ge \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}}. \end{cases}.$$

Analysis of the entrant $C = E$. Since the entrant faces no safety constraint, the entrant can choose any $\alpha \in [0.5, 1]$. We apply Corollary 7 to see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_E, N, \tilde{\alpha}_E)] = \inf_{\alpha \in [0.5,1]} \inf_{\lambda > 0} \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \lambda, N, \alpha)] = \Theta\left(N^{-\nu}\right),$$

which means that:

$$N_E^*(N_I, \tau_I, \infty, \mathcal{D}_W, \mathcal{D}_F) = \begin{cases} O\left(N_I\right) & \text{if } N_I \le \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \\ O\left(N_I^{\frac{1}{\nu+1}} \cdot \tilde{G}_I^{-\frac{1}{2(\nu+1)}}(1-\rho)^{-\frac{1}{2(\nu+1)}}\right) & \text{if } \tilde{G}_I^{-\frac{1}{2\nu}}(1-\rho)^{-\frac{1}{2\nu}} \le N_I \le \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}} \\ O\left(\tilde{G}_I^{-\frac{1}{\nu}}\right) & \text{if } N_I \ge \tilde{G}_I^{-\frac{1}{2}-\frac{1}{\nu}}(1-\rho)^{\frac{1}{2}} \end{cases}$$

as desired.

□

E.4 Extension of Theorem 5

We next study the market entry $\tilde{N}_E^*$ threshold in the environment of Theorem 5 where the incumbent has infinite data and the new company faces a *nontrivial safety constraint.* We place the further assumption that $\delta \le 1$. We compute the following upper bound on the modified market entry threshold.

Theorem 49 (Extension of Theorem 5)**.** *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma > 0$, $\delta \in (0, 1]$, and correlation coefficient $\rho \in [0, 1)$, and suppose that $P = \infty$. Suppose that the safety constraints τ_I and τ_E satisfy (2). Then it holds that $\tilde{N}_E^* = \tilde{N}_E^*(\infty, \tau_I, \tau_E, \mathcal{D}_W, \mathcal{D}_F)$ satisfies:*

$$\tilde{N}_E^* := O\left(\max\left(\tilde{D}^{-\frac{1}{\nu}}, \tilde{D}^{-\frac{\nu+1}{\nu}}\left(G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}} + \frac{1}{2}G_I - \frac{1}{2}G_E\right)\right)\right),$$

where $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma(\beta_1 - \beta)] = \Theta(1-\rho)$, where $\nu = \min(2(1+\gamma), \delta+\gamma) = \delta + \gamma$, where $G_I := \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^2$ and $G_E := \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_E, L^*(\rho))}\right)^2$, and where:

$$\tilde{D} := \alpha_E^* \cdot (G_I - G_E) - \frac{(G_I - G_E)^2}{4 \cdot L^*(\rho)}.$$

Theorem 49 shows that the key qualitative finding from Theorem 5—that the new company can enter with finite data, as long as they face a strictly weaker safety constraint than the incumbent company—readily extends to this setting. We note that the bound in Theorem 49 and the bound in

Theorem 5 take slightly different forms. Some of these differences are superficial: while the bound in Theorem 49 contains two—rather than three—regimes, the third regime in Theorem 5 does not exist in the case where $\delta \leq 1$. Other differences are more substantial: for example, the bound in Theorem 49 scales with $\tilde{D}$ while the bound in Theorem 5 scales with D. However, we expect some of this difference arises because the bound in Theorem 49 is not tight, rather than fundamental distinctions between the two settings. Proving a tight bound on the modified market entry threshold is an interesting direction for future work.

We compute an upper bound on the number of data points N_E that the new company needs to achieve at most loss $\left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^2$ on performance.

Lemma 50. *Suppose that the power-law scaling holds for the eigenvalues and alignment coefficients with scaling exponents $\gamma > 0, \delta \in (0,1]$ and correlation coefficient $\rho \in [0,1)$, and suppose that $P = \infty$. Suppose that the safety constraints τ_I and τ_E satisfy (1). For sufficiently large constant $C_{\delta,\gamma}$, if*

$$N_E \geq C_{\delta,\gamma} \cdot \max\left(\tilde{D}^{-\frac{1}{\nu}}, \tilde{D}^{-\frac{\nu+1}{\nu}}\left(G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}} + \frac{1}{2}G_I - \frac{1}{2}G_E\right)\right),$$

then it holds that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_E, N_E, \tilde{\alpha}_E)] \leq G_I,$$

where $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta)] = \Theta(1-\rho)$, where $\nu = \min(2(1+\gamma), \delta+\gamma) = \delta + \gamma$, where $G_I := \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^2$ and $G_E := \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_E, L^*(\rho))}\right)^2$, and where:

$$\tilde{D} := \alpha_E^* \cdot (G_I - G_E) - \frac{(G_I - G_E)^2}{4 \cdot L^*(\rho)}.$$

Proof. It suffices to construct $\tilde{\alpha}$ and $\tilde{\lambda}$ such that

$$\mathbb{E}_{\mathcal{D}_W}[\tilde{L}_2(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}, N_E, \tilde{\alpha})] \leq \tau_E$$

and

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}, N_E, \tilde{\alpha})] \leq G_I$$

for $N_E = \Omega\left(\max\left(\tilde{D}^{-\frac{1}{\nu}}, \tilde{D}^{-\frac{\nu+1}{\nu}}\left(G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}} + \frac{1}{2}G_I - \frac{1}{2}G_E\right)\right)\right)$.

To define $\tilde{\alpha}$ and $\tilde{\lambda}$, it is inconvenient to work with the following intermediate quantities. Let $\alpha_E^* = \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_E, L^*(\rho))}\right)^2$ and let $\alpha_I^* = \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^2$. We define an error function:

$$f(N_E, \alpha, \lambda) := \max(\lambda^{\frac{\nu}{1+\gamma}}, N_E^{-\nu}) + (1-\alpha)(1-\rho)\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N_E)}{N_E}$$

We define:

$$\tilde{\alpha} := \alpha_E^* + \frac{1}{2}(1-\alpha_E^*)^2 - \frac{1}{2}(1-\alpha_I^*)^2 = \alpha_I^* + \frac{\alpha_E^* - \alpha_I^*}{2}.$$

and

$$\tilde{\lambda} := \inf_{\lambda \in (0,1)} f(N_E, \tilde{\alpha}, \lambda).$$

At these values of $\tilde{\alpha}$ and $\tilde{\lambda}$ and under the condition on N_E, observe that:

$$\begin{aligned}
f(N_E, \tilde{\alpha}, \tilde{\lambda}) &= \Theta\left(\max\left(N_E^{-\nu}, \left(\frac{N_E}{(1-\tilde{\alpha})(1-\rho)}\right)^{-\frac{\nu}{\nu+1}}\right)\right) \\
&= \Theta\left(\max\left(N_E^{-\nu}, \left(\frac{N_E}{G_E^{\frac{1}{2}}(1-\rho)^{\frac{1}{2}} + \frac{1}{2}G_I + \frac{1}{2}G_E}\right)^{-\frac{\nu}{\nu+1}}\right)\right) \\
&= O\left(\tilde{D}\right),
\end{aligned}$$

where the implicit constant can be reduced by increasing the implicit constant on N_E.

The remainder of the analysis boils down to showing that $\mathbb{E}_{\mathcal{D}_W}[\tilde{L}_2(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}, N_E, \tilde{\alpha})] \leq \tau_E$ and $\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}, N_E, \tilde{\alpha})] \leq G_I$. To show this, we first derive an error function and bound these losses in terms of the error function.

Bounding $\mathbb{E}_{\mathcal{D}_W}[\tilde{L}_2(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}, N_E, \tilde{\alpha})] \leq \tau_E$. Observe that:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[\tilde{L}_2(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}, N_E, \tilde{\alpha})] \\
&=_{(A)} \tilde{\alpha}^2 L^*(\rho) + O\left(\max(\lambda^{\frac{\nu}{1+\gamma}}, N_E^{-\nu}) + (1-\alpha)(1-\rho)\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N_E)}{N_E}\right) \\
&= (\alpha_E^* + \frac{1}{2}(1-\alpha_E^*)^2 - \frac{1}{2}(1-\alpha_I^*)^2)L^*(\rho) + O\left(f(N_E, \tilde{\alpha})\right) \\
&\leq \left((\alpha_E^*)^2 L^*(\rho) + \frac{((1-\alpha_I^*)^2 - (1-\alpha_E^*)^2)^2}{4} - \alpha_E^*((1-\alpha_I^*)^2 - (1-\alpha_E^*)^2)\right) L^*(\rho) + \tilde{D} \\
&= \tau_E + \frac{(G_I - G_E)^2}{4 \cdot L^*(\rho)} - \alpha_E^*(G_I - G_E)\alpha_E^* \cdot (G_I - G_E) - \frac{(G_I - G_E)^2}{4 \cdot L^*(\rho)} \\
&= \tau_E
\end{aligned}$$

where (A) follows from Lemma 45. This gives us the desired bound.

Bounding $\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}, N_E, \tilde{\alpha})]$. Observe that:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}, N_E, \tilde{\alpha})] \\
&=_{(A)} (1-\tilde{\alpha})^2 L^*(\rho) + O\left(\max(\lambda^{\frac{\nu}{1+\gamma}}, N_E^{-\nu}) + (1-\alpha)(1-\rho)\frac{\min(\lambda^{-\frac{1}{1+\gamma}}, N_E)}{N_E}\right) \\
&\leq (1 - \alpha_E^* - \frac{1}{2}(1-\alpha_E^*)^2 + \frac{1}{2}(1-\alpha_I^*)^2)^2 L^*(\rho) + O\left(f(N_E, \tilde{\alpha})\right) \\
&\leq \left((1-\alpha_E^*)^2 + \frac{((1-\alpha_I^*)^2 - (1-\alpha_E^*)^2)^2}{4} - (1-\alpha_E^*)((1-\alpha_I^*)^2 - (1-\alpha_E^*)^2)\right) L^*(\rho) + \tilde{D} \\
&\leq G_E + (G_I - G_E)(1-\alpha_E^*) + \frac{(G_I - G_E)^2}{4L^*(\rho)} + \alpha_E^* \cdot (G_I - G_E) - \frac{(G_I - G_E)^2}{4 \cdot L^*(\rho)} \\
&= G_I.
\end{aligned}$$

where (A) uses Theorem 9, coupled with the fact that $\delta \leq 1$ (which means that $\nu' = \nu$, so the mixture finite data error is subsumed by the finite data error) and coupled with Lemma 35. This gives us the desired bound.

□

We are now ready to prove Theorem 49.

Proof of Theorem 49. We analyze $(\tilde{\alpha}_C, \tilde{\lambda}_C)$ first for the incumbent $C = I$ and then for the entrant $C = E$. Like in the theorem statement, let $L^*(\rho) = \mathbb{E}_{\mathcal{D}_W}[(\beta_1 - \beta_2)^T \Sigma (\beta_1 - \beta)] = \Theta(1 - \rho)$, let $\nu = \min(2(1+\gamma), \delta + \gamma) = \delta + \gamma$, let $G_I := \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^2$ and $G_E := \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_E, L^*(\rho))}\right)^2$, and let:

$$\tilde{D} := \alpha_E^* \cdot (G_I - G_E) - \frac{(G_I - G_E)^2}{4 \cdot L^*(\rho)}.$$

Analysis of the incumbent $C = I$. To compute $\tilde{\alpha}_I$ and $\tilde{\lambda}_I$, we apply Lemma 44. The assumption $\tau_I \geq \mathbb{E}_{\mathcal{D}_W}[L_2(\beta_1, \beta_2, \Sigma, 0.5)]$ in the lemma statement can be rewritten as $\tau_I \geq 0.25 L^*(\rho)$, which guarantees the assumptions in Lemma 44 are satisfied. By Lemma 44, we see that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, \infty, \tilde{\alpha}_I)] = \left(\sqrt{L^*(\rho)} - \sqrt{\min(\tau_I, L^*(\rho))}\right)^2 = G_I.$$

Analysis of the entrant $C = E$. We apply Lemma 50 to see for sufficiently large constant $C_{\delta,\gamma}$, if

$$N_E \geq C_{\delta,\gamma} \cdot \max\left(\tilde{D}^{-\frac{1}{\nu}}, \tilde{D}^{-\frac{\nu+1}{\nu}} \left(G_E^{\frac{1}{2}} (1-\rho)^{\frac{1}{2}} + \frac{1}{2} G_I - \frac{1}{2} G_E\right)\right),$$

then it holds that:

$$\mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_E, N_E, \tilde{\alpha}_E)] \leq G_I = \mathbb{E}_{\mathcal{D}_W}[L_1^*(\beta_1, \beta_2, \mathcal{D}_F, \tilde{\lambda}_I, \infty, \tilde{\alpha}_I)].$$

This means that:

$$\tilde{N}_E^* = O\left(\max\left(\tilde{D}^{-\frac{1}{\nu}}, \tilde{D}^{-\frac{\nu+1}{\nu}} \left(G_E^{\frac{1}{2}} (1-\rho)^{\frac{1}{2}} + \frac{1}{2} G_I - \frac{1}{2} G_E\right)\right)\right)$$

as desired. □