Optimal Regularization for Performative Learning

Edwige Cyffers [1] **Alireza Mirrokni** [2] **Marco Mondelli** [2]

Abstract

In performative learning, the data distribution reacts to the deployed model—for example, because strategic users adapt their features to game it—which creates a more complex dynamic than in classical supervised learning. One should thus not only optimize the model for the current data but also take into account that the model might steer the distribution in a new direction, without knowing the exact nature of the potential shift. We explore how regularization can help cope with performative effects by studying its impact in high-dimensional ridge regression. We show that, while performative effects worsen the test risk in the population setting, when moving to the over-parameterized regime where the number of features exceeds the number of samples, the optimal regularization in the presence of performativity helps reduce the variance in the estimated parameters, thereby improving performance. Furthermore, we prove that the optimal regularization scales with the overall strength of the performative effect, making it possible to set the regularization in anticipation of this effect. We illustrate this finding through empirical evaluations of the optimal regularization parameter on both synthetic and real-world datasets.

1. Introduction

When machine learning predictions affect user outcomes, deployed models can induce shifts in the data distribution. These shifts may result from strategic user behavior—where individuals try to secure favorable outcomes such as loan approval or college admission (Bechavod et al., 2021; Narang et al., 2023; Wang et al., 2023a)—or from self-fulfilling prophecies, for example in economic forecasts, recommendation systems, or predictive policing (Morgenstern, 1928; Ensign et al., 2018; Ursu, 2015). Such distribution shifts can undermine predictive performance over time, amplifying bias and reducing model quality (Taori & Hashimoto, 2023; Pan et al., 2024). Performative learning (Perdomo et al., 2020) addresses this feedback loop by parameterizing the data distribution with the same parameter as the model. This allows optimization to account not only for the training loss but also for the steering of the data distribution.

Unfortunately, while optimizing model parameters is a classical problem in machine learning, estimating the performative effect on the distribution is generally infeasible, as the distribution is unknown to the learner. Several algorithms have been proposed to approximate this effect (Miller et al., 2021; Izzo et al., 2022; Cyffers et al., 2024), typically by assuming that it depends on a small number of parameters in a sufficiently simple way that can be learned across the first few deployments. However, these methods are limited to relatively toy examples in small dimensions and may be impractical in high-dimensional settings. In particular, many approaches require numerous repeated deployments, alternating between loss minimization and distribution steering. Yet in practice, deployment often happens only once after full training. This makes repeated risk minimization (RRM) (Perdomo et al., 2020)—where one trains until convergence before deployment, and the number of deployments is small—the default in many applications, even though it remains largely unaddressed by existing mitigation methods. This motivates a shift away from exact estimation of performative effects toward the study of principled choices of loss functions and models, and in particular regularization is a natural and tractable candidate.

In this work, we study how regularization mitigates performative effects in repeated retraining. Unlike estimation-based methods, regularization does not depend on a precise characterization of the distribution shift, avoids their limitations, and introduces little computational overhead. Prior work suggests its potential benefits: Perdomo et al. (2020) proved that retraining converges to an optimal solution under assumptions tied to the strong convexity of the loss, which ridge regularization can enforce; more recently, Cyffers et al. (2024) showed that, in classification, the performative optimum can be interpreted as a regularized version of the non-performative problem, with numerical evidence

[1]CNRS, LAMSADE, Dauphine PSL [2]Institute of Science and Technology Austria, Klosterneuburg, Austria. Correspondence to: Edwige Cyffers <edwige.cyffers@cnrs.fr>, Marco Mondelli <marco.mondelli@ist.ac.at>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

that ridge penalties perform well in small-dimensional classification tasks. However, these results, as most of the performative literature, do not cover the high-dimensional regime. In particular, regularization may encourage reliance on spurious features (Bombari & Mondelli, 2025) in higher dimension, especially when such features are reinforced by performativity.

To better understand these tradeoffs, we study the role of ridge regularization in linear regression in the presence of performativity and spurious features. We consider both *(i)* the population regime, with enough data to recover exactly the unknown vector of regression coefficients at each deployment, and *(ii)* the over-parameterized regime, where the number of data samples is a fixed fraction of the number of parameters. This last setting, though simple, captures behaviors relevant to deep learning, such as double descent (Belkin et al., 2019; Hastie et al., 2022), benign overfitting (Bartlett et al., 2020) and adversarial robustness (Fawzi et al., 2018; Ribeiro et al., 2023). In performative learning, it also brings the additional advantage that parameters and data live in the same space, simplifying the encoding of performative effects. The theoretical framework we develop enables us to provide strong evidence for the effectiveness of regularization under performativity, and to show how regularization should be scaled. More precisely, our contributions are summarized below.

1. In the population regime, we characterize how the risk depends on magnitude and direction of the performative effect, as well as on spurious features (Theorem 4.1). We find that the optimal regularization is proportional to the strength of the performative effect and it mitigates the performance loss due to performativity: zero excess risk is achieved with identity covariance and constant entries of the performative vector, while the risk remains significant in the presence of a complex covariance structure and highly variable entries of the performative vector (Corollary 4.2).

2. In the proportional regime with random data, we establish a deterministic equivalent of the performative fixed point, depending only on population covariance and regularization (Theorem 5.1). The analysis of this deterministic equivalent then unveils a remarkable phenomenology: for small noise variance, the optimal regularization moves in the same direction as the performative effect on predictive features, while it moves in the opposite direction as the performative effect on spurious features; remarkably, the optimally-regularized risk improves in the presence of a performative effect that reinforces existing trends.

3. We illustrate these behaviors on both synthetic data and real-world datasets (Housing, LSAC), showing empirically that our findings provide valuable insights beyond the assumptions in the theory and extend to other regularizers.

2. Related work

Performative learning. Performative learning was introduced by Perdomo et al. (2020), who showed that retraining converges under assumptions including strong convexity. Subsequent works demonstrated that retraining can enable adaptation over time (Li et al., 2022; Drusvyatskiy & Xiao, 2023; Brown et al., 2022; Wang et al., 2023b) but can also fail dramatically (Miller et al., 2021; Izzo et al., 2022; Cyffers et al., 2024). Several works evaluate performance on the initial distribution rather than the induced one (Demirel et al., 2024; Tsoy et al., 2025; Taori & Hashimoto, 2023). Label shift is standard in domain adaptation (Zhu et al., 2024; Cai & Wei, 2021) and can encode performative effects, such as placebo effects or traffic prediction in (Kirev et al., 2026; Hardt & Mendler-Dünner, 2023). The role of model choice has been studied in the related setting of collective action (Ben-Dov et al., 2024). Improved performance due to performativity was observed by Bechavod et al. (2021), and our work provides further evidence supporting this claim. Finally, performative learning effects tend to be harder to learn in high-dimensional settings, as noted by Jagadeesan et al. (2022); Bracale et al. (2025). Our work is the first to use tools from high-dimensional statistics to study performative learning, to our knowledge.

High-dimensional regression and role of ridge regularization. The high-dimensional setting where numbers of features and samples scale proportionally was considered by a rich line of work: the test error of ridgeless and ridge regression is characterized by Hastie et al. (2022); Wu & Xu (2020); Richards et al. (2021); Tsigler & Bartlett (2023); max-margin classification is studied by Montanari et al. (2025); Deng et al. (2022), model compression by Chang et al. (2021), distribution shift by Patil et al. (2024); Mallinar et al. (2024), transfer learning by Yang et al. (2025); Song et al. (2024) and learning from surrogate data by Kolossov et al. (2024); Jain et al. (2024); Rezaei et al. (2025).The role of ridge regularization has also been studied. Hastie et al. (2022) optimally tune the ridge penalty, while Richards et al. (2021) give conditions for the optimality of ridgeless interpolation. The sign of the optimal ridge penalty was studied for the standard in-distribution regression setup (Wu & Xu, 2020; Tsigler & Bartlett, 2023), as well as out-of-distribution (Patil et al., 2024): these works give conditions under which the optimal ridge is negative, associating the phenomenon of negative optimal regularization to over-parameterization. Our paper shows that such a phenomenon occurs also in the population setting, due to performative effects. The distribution of the empirical risk minimizer was established by Han & Xu (2023). Leveraging this characterization, spurious correlations were studied by Bombari &

Mondelli (2025) and weak-to-strong generalization by Ildiz et al. (2025). We will also build on these tools to analyze the risk of repeated risk minimization.

3. Preliminaries and problem setup

In this section, we introduce our performative regression setting. We consider a sequence of model deployments $(\theta_k)_{k\geq 0}$, and let $\mathcal{D}(\theta)$ be the dataset generated in reaction to the deployment of θ. At each deployment, n new samples are collected, and the model is fully retrained. This setting, known as repeated risk minimization (RRM) (Perdomo et al., 2020), reflects real-world scenarios where deployments are costly and thus limited in number. It also aligns with the fact that convergence to the fixed point is fast and requires only a few iterations to reach equilibrium in practice. We encode the performative effect as a shift in the label, where each feature's contribution varies depending on an additional linear term in the model parameter.

Assumption 3.1 (Regression performative model)**.** For $\theta \in \mathbb{R}^p$, samples from $\mathcal{D}(\theta)$ are taken i.i.d. with features x having zero mean and covariance Σ drawn independently of θ and with the label y given by

$$y = x^\top \theta^*_{\text{pop}} + x^\top D\theta + w, \quad w \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

We assume $p = 2d$, $(\theta^*_{\text{pop}})^\top = (a^\top, 0)$ with a having zero mean and covariance I_d/d, and $D = \text{diag}(b, c)$ where $b, c \in \mathbb{R}^d$ with $\|b\|_\infty, \|c\|_\infty < 1$.

This model generalizes the one-dimensional setting (Example 2.2) in Perdomo et al. (2020), where labels follow a binomial distribution with parameter $\frac{1}{2} + x\theta^* + x\bar{b}\theta$, for $\theta^* \in (0, \frac{1}{2})$ and $\bar{b} < \frac{1}{2} - \mu$. Focusing on label shifts is natural in regression: it keeps the feature distribution centered and unchanged across deployments (this can be enforced via pre-processing) despite performative effects, and it can encode scenario such as placebo effects (Kirev et al., 2026). We do not cover feature shifts. However, when a feature shift affects all data points in a regression task, replacing x by $x + \theta$ can be absorbed by centering the data, and prior work provides numerical evidence that regularization remains beneficial when feature shifts affect only one class in binary classification (Cyffers et al., 2024).

The performative term $x^\top D\theta$ enforces coordinate-wise effects. This is consistent with previous works (Cyffers et al., 2024; Izzo et al., 2022; Hardt & Mendler-Dünner, 2023) and also close to the model $y = x^\top \theta^*_{\text{pop}} + \mu^\top \theta + w$ studied by Miller et al. (2021), where the performative effect does not depend on x but only on a fixed vector μ. Assuming linearity in θ is reasonable, as performative effects are expected to be moderate to avoid iterations to diverge. We specify in the rest of the paper when the fact that D is diagonal is needed. Intuitively, diagonal coefficients can be interpreted directly as the modifications made by a strategic agent, depending on how the feature is used and the cost of modifying it. Most existing methods impose explicit constraints on the performative effect (Miller et al., 2021), and our setting is no more restrictive: we only require $\|b\|_\infty, \|c\|_\infty < 1$. We set the second half of θ^*_{pop} to zero to represent spurious features, and c captures the corresponding performative effect. This enables us to express correlations between predictive and spurious features via the block structure of the covariance

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_2 \end{bmatrix},$$

where Σ_1 denotes the covariance for the predictive part, Σ_2 the covariance for the spurious part, and Σ_{12} the covariance between the two blocks. Under this setting, RRM corresponds to solving

$$\theta_k = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \ell(x_i^{(k-1)}, y_i^{(k-1)}; \theta) + \frac{\lambda}{2} \|\theta\|_2^2 \right\}, \tag{2}$$

where $\{(x_i^{(k-1)}, y_i^{(k-1)})\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}(\theta_{k-1})$ and ℓ is the squared loss. This defines a recurring sequence in both population and over-parameterized regimes. In the population case, the sequence converges in parameter space to a fixed vector θ^*_{pop} (Section 4), while in the over-parameterized case the vector varies at each iteration but the excess risk still converges deterministically (Section 5).

We evaluate the test risk when the final model is deployed on the untouched distribution $\mathcal{D}(\theta = 0)$. Testing on $\mathcal{D}(\theta = 0)$ is particularly relevant for long-term fairness, as it prevents bias amplification over time (Ensign et al., 2018; Taori & Hashimoto, 2023) or steering the distribution toward undesirable regimes that decrease the risk by collapsing the data distribution to make it easier to predict (e.g., reducing entropy or producing a single possible label in a classification task) (Tsoy et al., 2025; Demirel et al., 2024). This choice also enables testing whether regularization increases reliance on spurious features (Bombari & Mondelli, 2025), which one wants to avoid. We thus aim to minimize the following excess risk:

$$\begin{aligned} \mathcal{R}(\Sigma, \theta, \theta^*_{\text{pop}}) :=& \mathbb{E}_{\mathcal{D}(\theta=0)} \left[(y - x^\top \theta)^2 \right] - \sigma^2 \\ =& \|\Sigma^{1/2}(\theta - \theta^*_{\text{pop}})\|_2^2, \end{aligned} \tag{3}$$

where we have subtracted the Bayes risk σ^2. In Section 4, we analyze this risk in the population setting, where enough data is available to exactly recover the parameter vector and the optimal solution is θ^*_{pop} (as suggested by the notation). In Section 5, we then focus on the over-parameterized regime where $p > n$.

We note that, in the population setting, testing on $\mathcal{D}(\theta)$ gives zero excess risk, thus trivializing the problem. We provide results for testing on $\mathcal{D}(\theta)$ in the over-parameterized setting at the end of Section 5.

4. Analysis in the population setting

In this section, we tackle the population regime where there are enough samples from $\mathcal{D}(\theta_k)$ at each deployment to compute exactly the next regressor, as would typically happen in a low-dimensional setting. The sequence $(\theta_k)_k$ is thus deterministically defined by

$$\begin{aligned}\theta_k &= (\Sigma + \lambda I_p)^{-1}\mathbb{E}_{(x,y)\sim\mathcal{D}(\theta_{k-1})}[xy] \\ &= (\Sigma + \lambda I_p)^{-1}(\Sigma\theta^*_{\text{pop}} + \Sigma D\theta_{k-1}),\end{aligned} \tag{4}$$

where in the second equality we plug back the definition of the current data distribution in (1).

Excess risk at the performative fixed point. By unrolling (4), for any arbitrary (possibly non-diagonal) matrix D, we have that the sequence $(\theta_k)_k$ converges at an exponential rate to the fixed point

$$\theta^\infty = (I_p + \lambda\Sigma^{-1} - D)^{-1}\theta^*_{\text{pop}}. \tag{5}$$

The formal statement, including an explicit convergence rate, is deferred to Lemma B.1 in Appendix B. By inserting (5) into (3) and taking the expectation with respect to θ^*_{pop}, we have

$$\begin{aligned}\mathbb{E}_{\theta^*_{\text{pop}}}\mathcal{R}(\Sigma, \theta^\infty, \theta^*_{\text{pop}}) &= \mathbb{E}_{\theta^*_{\text{pop}}}[(\theta^*_{\text{pop}})^\top A^\top \Sigma A \theta^*_{\text{pop}}] \\ &= \frac{1}{d}\operatorname{Tr}\left[(A^\top\Sigma A)_1\right],\end{aligned} \tag{6}$$

where we define $A := (\Sigma + \lambda I_p - \Sigma D)^{-1}\Sigma - I_p$, use $(\theta^*_{\text{pop}})^\top = (a^\top, 0)$ with a having zero mean and covariance I_d/d and, given a $p \times p$ matrix M, denote by $(M)_1$ its top-left $d \times d$ block. This leads to the following approximation for the excess risk, proved in Appendix B.

Theorem 4.1 (Excess risk – population)**.** *Let $F = D - \lambda\Sigma^{-1}$. Then, we have*

$$\begin{aligned}\mathbb{E}_{\theta^*_{\text{pop}}}\mathcal{R}(\Sigma, \theta^\infty, \theta^*_{\text{pop}}) &= \widetilde{\mathcal{R}}_{\text{pop}}(D, \lambda, \Sigma) + O(\|F\|^2_{\text{op}}), \\ \widetilde{\mathcal{R}}_{\text{pop}}(D, \lambda, \Sigma) &:= \frac{1}{d}\operatorname{Tr}[\operatorname{diag}(b^2)\Sigma_1] - 2\lambda\bar{b} + \frac{1}{d}\lambda^2\operatorname{Tr}(S_1),\end{aligned} \tag{7}$$

where $\|\cdot\|_{\text{op}}$ denotes the operator norm, $\bar{b} := \frac{1}{d}\operatorname{Tr}[\operatorname{diag}(b)] = \frac{1}{d}\sum_{i=1}^d b_i$, $b^2 := [b_1^2, \dots, b_d^2] \in \mathbb{R}^d$ and $S_1 = (\Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21})^{-1}$ is the Schur complement of Σ.

The matrix $F = D - \lambda\Sigma^{-1}$ naturally appears in the computation, and its norm can be bounded explicitly as a function of b, c, λ and Σ by applying Weyl's inequality (see Lemma B.2 in Appendix B) and ensuring that the approximation $O(\|F\|^2_{\text{op}})$ is tighter than $O(\max(\|b\|_\infty, \|c\|_\infty, \lambda)^2)$, since b and c can be partially canceled by λ. In fact, this cancellation occurs when λ is near-optimal, resulting in an accurate approximation, see Figure 1.

Optimal regularization and optimally regularized risk. Leveraging the risk expression for small F of Theorem 4.1, we next study the behavior of the optimal regularization and of the corresponding optimal risk. Formally, define

$$\begin{aligned}\lambda^*_{\text{pop}}(D, \Sigma) &:= \arg\min_{\lambda\in\mathbb{R}} \widetilde{\mathcal{R}}_{\text{pop}}(D, \lambda, \Sigma), \\ \mathcal{R}^*_{\text{pop}}(D, \Sigma) &:= \min_{\lambda\in\mathbb{R}} \widetilde{\mathcal{R}}_{\text{pop}}(D, \lambda, \Sigma).\end{aligned} \tag{8}$$

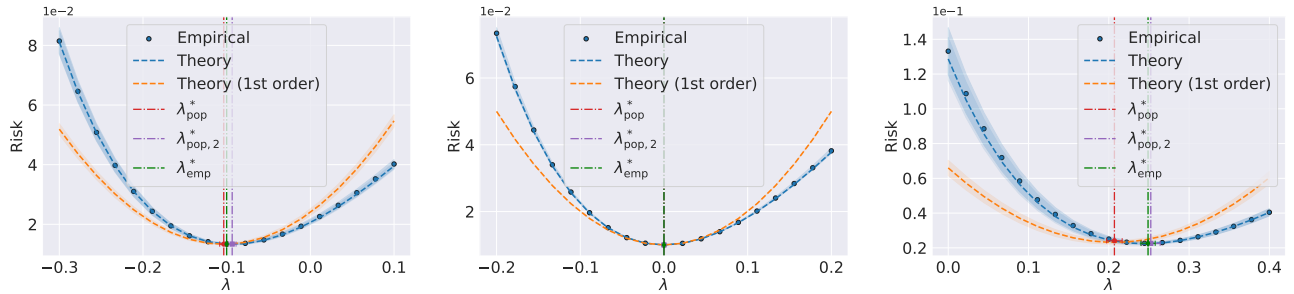
From (7), we note that $\widetilde{\mathcal{R}}_{\text{pop}}(D, \lambda, \Sigma)$ is quadratic in λ, so the minimization in (8) can be solved explicitly, leading to the expressions below.

Corollary 4.2 (Optimal regularization – population)**.** *In the setting described above, we have*

$$\begin{aligned}\lambda^*_{\text{pop}}(D, \Sigma) &= \frac{\bar{b}d}{\operatorname{Tr}(S_1)}, \\ \mathcal{R}^*_{\text{pop}}(D, \Sigma) &= \frac{1}{d}\operatorname{Tr}(\operatorname{diag}(b^2)\Sigma_1) - \frac{\bar{b}^2 d}{\operatorname{Tr}(S_1)}.\end{aligned} \tag{9}$$

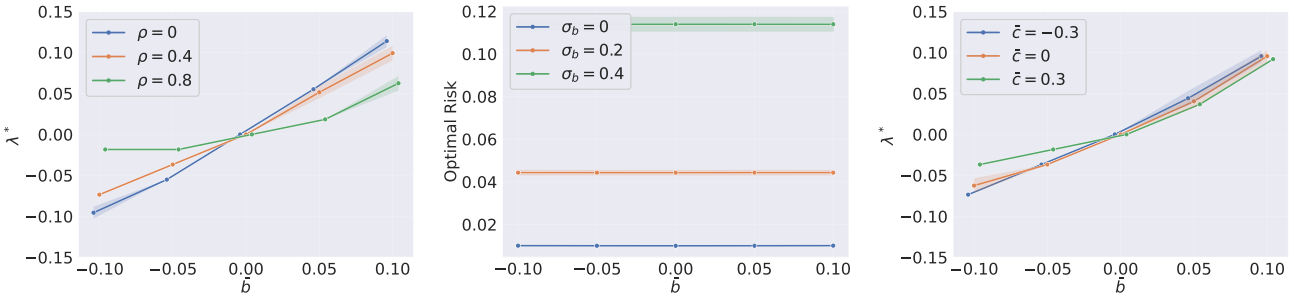
These formulas call for several comments. First, the optimal regularization $\lambda^*_{\text{pop}}(D, \Sigma)$ is proportional to the strength of the performative effect $\bar{b}$, see Figure 2a (and also the location of the minima in Figure 1). The fact that $\lambda^*_{\text{pop}}(D, \Sigma)$ grows with $\bar{b}$ captures an effect common in practice: when performativity reinforces existing trends—corresponding to "rich-get-richer" phenomena, such as a feature becoming more important over successive deployments—the optimal regularizer increases and helps to limit this effect. Conversely, when the performative effect already mitigates the influence of some feature, the optimal solution calls for less regularization. In the population case, this corresponds to negative regularization (see Figure 1a) which, although less common, has also been studied in the literature (Wu & Xu, 2020; Tsigler & Bartlett, 2023; Patil et al., 2024). Spurious correlations tend to reduce the optimal regularization, although their effect is mild, see Figure 2a.

Second, the optimal risk $\mathcal{R}^*_{\text{pop}}(D, \Sigma)$ is always positive and, thus, worse than in the non-performative scenario, where it is zero. More specifically, zero excess risk can only be reached with a non-zero b if $\Sigma = I_p$ and b is aligned with the all-1 vector, see the blue line in Figure 2b. Intuitively, as regularization impacts all features equally, it better compensates performativity in this uniform case. If the variance in the entries of b grows, then $\mathcal{R}^*_{\text{pop}}(D, \Sigma)$ increases, see Figure 2b. We finally note that $\widetilde{\mathcal{R}}_{\text{pop}}(D, \lambda, \Sigma)$ does not depend on c and, in fact, the effect of c is only visible at



(a) $\bar{b} = -0.1$ *(b)* $\bar{b} = 0$ *(c)* $\bar{b} = 0.2$

Figure 1. Excess risk at the performative fixed point θ^∞ in (5), as a function of ridge regularization λ, for $d = 100$, $\Sigma = I_p$, entries of b uniform in $\left[\min\{0, 2\bar{b}\}, \max\{0, 2\bar{b}\}\right]$, $c = 0$ and $\sigma = 0.1$. Empirical values (blue dots) are computed from 20 i.i.d. trials on a and 5 i.i.d. trials on b, with error band at 1 standard deviation. Theoretical predictions (blue dashed curves) are from (6) and match perfectly empirical values. First-order approximations (orange dashed curves) are given by $\widetilde{\mathcal{R}}_{\text{pop}}(D, \lambda, \Sigma)$ in (7) and still provide a good match when λ is near-optimal. The green vertical line is the optimal regularization obtained by numerically optimizing the excess risk of θ^∞ (λ^*_{emp}), the red one is the first-order approximation (λ^*_{pop} from (9)) and the violet one the second-order approximation ($\lambda^*_{\text{pop},2}$ minimizing (10)).



(a) Spurious correlations *(b)* Variance of entries of b *(c)* Spurious performative effect

Figure 2. Optimal regularization and risk for the performative fixed point θ^∞ in (5), with $d = 100$, $\Sigma_1 = \Sigma_2 = I_d$, $\Sigma_{12} = \rho I_d$. Values are computed from 20 i.i.d. trials on a and 5 i.i.d. trials on b, with error band at 1 standard deviation. (a) Optimal regularization as a function of $\bar{b}$ for $\rho \in \{0, 0.4, 0.8\}$. The entries of b are uniform in $\left[\min\{0, 2\bar{b}\}, \max\{0, 2\bar{b}\}\right]$ and $c = 0$. (b) Optimal risk as a function of $\bar{b}$. Different curves correspond to different variances σ_b^2 of the entries of b, which are uniform in $\left[\bar{b} - \sigma_b\sqrt{3}, \bar{b} + \sigma_b\sqrt{3}\right]$ for $\sigma_b \in \{0, 0.2, 0.4\}$. We pick $c = 0$ and $\rho = 0$. We note that, when $\rho = 0$, $\mathcal{R}^*_{\text{pop}}(D, \Sigma)$ equals the empirical variance of the entries of b and, as such, it does not depend on $\bar{b}$. (c) Optimal regularization as a function of $\bar{b}$ for $\bar{c} \in \{-0.3, 0, 0.3\}$. The entries of b are uniform in $\left[\min\{0, 2\bar{b}\}, \max\{0, 2\bar{b}\}\right]$, the entries of c are uniform in $[\min\{0, 2\bar{c}\}, \max\{0, 2\bar{c}\}]$, and $\rho = 0.5$.

higher order, as seen in this formula, proven in Appendix B:

$$\mathbb{E}_{\theta^*_{\text{pop}}} \mathcal{R}(\Sigma, \theta^\infty, \theta^*_{\text{pop}}) = \frac{1}{d}\Big(-2\lambda^3 \operatorname{Tr}\left[\left(\Sigma^{-2}\right)_1\right] + \lambda^2\big(\operatorname{Tr}[S_1] + 6\operatorname{Tr}\left[\operatorname{diag}(b)S_1\right]\big) - \lambda\Big(2\operatorname{Tr}\left[\operatorname{diag}(b)\Sigma_1 \operatorname{diag}(b)S_1\right] + 2\operatorname{Tr}\left[\operatorname{diag}(b)\Sigma_{12}\operatorname{diag}(c)S_{21}\right] + 2d\bar{b} + 4\operatorname{Tr}[\operatorname{diag}(b^2)]\Big) + \operatorname{Tr}\left[\operatorname{diag}(b^2)\Sigma_1\right] + 2\operatorname{Tr}\left[\operatorname{diag}(b^3)\Sigma_1\right]\Big) + O(\|F\|_{\text{op}}^4), \tag{10}$$

with $S_{21}^\top = -(\Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_2^{-1}$. Figure 1 illustrates that the minimizer of this second-order approximation ($\lambda^*_{\text{pop},2}$) is close to the minimizer obtained numerically (λ^*_{emp}). While c tends to steer the optimal regularizer in the opposite direction (less regularization in the case of a self-reinforcing performative effect), it does so only through the cross term $2\operatorname{Tr}[\operatorname{diag}(b)\Sigma_{12}\operatorname{diag}(c)S_{21}]$, which also depends on b. Figure 2c shows that c moves the optimal regularization in a direction opposite to its sign, but its effect remains rather limited.

5. Analysis in the over-parameterized setting

Next, we consider the case where n, p are both large and scale proportionally, with $p/n = \kappa > 1$. All *constants* (e.g., R, M) are intended to be positive values independent of n, p. For mathematical convenience, we opt for a different normalization w.r.t. (2), and the estimator θ_k is given by

$$\theta_k = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2p} \sum_{i=1}^n \ell\left(x_i^{(k-1)}, y_i^{(k-1)}; \theta\right) + \frac{\lambda}{2}\|\theta\|_2^2 \right\}. \tag{11}$$

Solving for θ_k yields

$$\theta_k = \frac{1}{p}\left(\frac{1}{p}X^{(k-1)\top}X^{(k-1)} + \lambda I_p\right)^{-1} X^{(k-1)\top} y^{(k-1)}, \tag{12}$$

where $X^{(k-1)} = [x_1^{(k-1)}, \dots, x_n^{(k-1)}] \in \mathbb{R}^{n\times p}$ and $y^{(k-1)} = [y_1^{(k-1)}, \dots, y_n^{(k-1)}] \in \mathbb{R}^n$. Note that, when θ_k is given by (12), the risk $\mathcal{R}(\Sigma, \theta_k, \theta^*_{\rm pop})$ as defined in (3) is a random quantity since the data $\{X^{(\ell)}\}_{\ell=1}^{k-1}$ and the noise contained in the labels $\{y^{(\ell)}\}_{\ell=1}^{k-1}$ are random. This makes it challenging to characterize optimal ridge penalty and optimally-tuned risk. To address the challenge, we first establish a *deterministic* equivalent of the risk at the performative fixed point. We next optimize such deterministic equivalent and study the effect of performativity on the optimal regularization.

Deterministic equivalent of the performative fixed point. First, note that, if we regard the performative effect as small and aim at characterizing its effect on the fixed point up to the leading (first) order, it suffices to do two iterations of the recursion in (12). In fact, the labels $y^{(k-1)}$ are linear in $D\theta_{k-1}$, so we expect that, after two iterations, the performative fixed point is reached up to fluctuations of order $O(\|D\|_{\rm op}^2)$. Now, the risk after two iterations is still a random quantity, so we apply techniques from (Han & Xu, 2023; Ildiz et al., 2025) to derive a deterministic equivalent. The formal statement is below and the proof is deferred to Appendix C.

Theorem 5.1 (Excess risk – over-parameterized)**.** *Let Assumption 3.1 hold with $x \sim \mathcal{N}(0, \Sigma)$. Let $R > 0$ be a constant s.t. $\|\theta^*_{\rm pop}\|_2, \|\theta_0\|_2 \le R$. Assume that $\kappa, \sigma, \lambda \in (1/M, M)$ and $\|\Sigma\|_{\rm op},\ \|\Sigma^{-1}\|_{\rm op} \le M$ for some constant $M > 1$. Then, there exists a constant $C = C(M, R)$ such that for any $\delta \in (0, 1/2]$, with probability at least $1 - Cpe^{-p\delta^4/C}$,*

$$\left|\mathcal{R}(\Sigma, \theta_2, \theta^*_{\rm pop}) - \mathcal{R}_{\rm eq}\left(\Sigma, \theta^*_{\rm pop}, D, \lambda\right)\right| \le \delta + O(\|D\|_{\rm op}^2), \tag{13}$$

where

$$\mathcal{R}_{\rm eq}\left(\Sigma, \theta^*_{\rm pop}, D, \lambda\right) = \tau\langle\theta^*_{\rm pop}, \Xi\left(\tau I_p - 2\Xi\Sigma^2 D\right)\Sigma\Xi\theta^*_{\rm pop}\rangle + \kappa\,{\rm Tr}\left[\Sigma^2\Xi^2\right]\frac{\sigma^2 + \tau^2\langle\theta^*_{\rm pop}, \Xi\left(I_p + 2\Xi\Sigma D\right)\Sigma\Xi\theta^*_{\rm pop}\rangle}{p - \kappa\,{\rm Tr}\left[\Sigma^2\Xi^2\right]}, \tag{14}$$

with $\Xi = (\Sigma + \tau I_p)^{-1}$ and τ is the unique solution of

$$\kappa^{-1} - \frac{\lambda}{\tau} = \frac{1}{p}\,{\rm Tr}\left[(\Sigma + \tau I_p)^{-1}\Sigma\right]. \tag{15}$$

In words, Theorem 5.1 shows that the risk $\mathcal{R}(\Sigma, \theta_2, \theta^*_{\rm pop})$ is well approximated by the quantity $\mathcal{R}_{\rm eq}\left(\Sigma, \theta^*_{\rm pop}, D, \lambda\right)$ defined in (14). We highlight that this quantity does not depend on the initialization θ_0: up a fluctuation of order $O(\|D\|_{\rm op}^2)$, *the risk has reached a fixed point* after two iterations. While the data (and, consequently, $\mathcal{R}(\Sigma, \theta_2, \theta^*_{\rm pop})$) are random, $\mathcal{R}_{\rm eq}\left(\Sigma, \theta^*_{\rm pop}, D, \lambda\right)$ provides a *deterministic equivalent* that depends only on the population covariance Σ, the ground-truth vector $\theta^*_{\rm pop}$, the matrix D capturing the performative effect and the regularization λ.

We note that the result of Theorem 5.1 holds for a general matrix D with bounded operator norm—not necessarily a diagonal D as in Assumption 3.1. The assumptions ($\|\theta^*_{\rm pop}\|_2, \|\theta_0\|_2 \le R$, $\kappa, \sigma, \lambda \in (1/M, M)$, $\|\Sigma\|_{\rm op},\ \|\Sigma^{-1}\|_{\rm op} \le M$) are all standard in the related literature (Han & Xu, 2023; Ildiz et al., 2025). We could handle the ridgeless case $\lambda = 0$ in a similar way to (Han & Xu, 2023; Ildiz et al., 2025). However, this requires changing some details and we have opted to avoid the notation clutter, since our focus is on the effect of regularization. The assumption on the features x being Gaussian can be also relaxed. In fact, the results of (Han & Xu, 2023) (see Theorem 2.4 therein) hold for $\Sigma^{-1/2}x$ having independent, zero mean, unit variance and uniformly subgaussian entries. We prove a formal extension of Theorem 5.1 to sub-Gaussian data in Appendix D.

Optimal regularization and optimally regularized risk. Leveraging the characterization of Theorem 5.1, we optimize the ridge regularization. We focus on the case $\Sigma = \begin{bmatrix} I_d & \rho I_d \\ \rho I_d & I_d \end{bmatrix}$ for small ρ and require D to be diagonal as in Assumption 3.1. While simplified, this setting captures the performative effect of both predictive and spurious features, which are mixed via the covariance matrix Σ, leading to an interesting phenomenology. Lemma E.1 in Appendix E computes $\mathbb{E}_{\theta^*_{\rm pop}}\mathcal{R}_{\rm eq}\left(\Sigma, \theta^*_{\rm pop}, D, \lambda\right)$, as well as the following expansion in ρ:

$$\mathbb{E}_{\theta^*_{\rm pop}}\mathcal{R}_{\rm eq}\left(\Sigma, \theta^*_{\rm pop}, D, \lambda\right) = \widetilde{\mathcal{R}}(D, \lambda, \rho) + O(\bar{b}\rho^2 + \rho^4),$$
$$\widetilde{\mathcal{R}}(D, \lambda, \rho) := \mathcal{R}_0(\lambda, \rho) + \bar{b}A_1(\lambda) + \bar{c}\rho^2 A_2(\lambda), \tag{16}$$

with $\bar{b} = {\rm Tr}[{\rm diag}(b)]/d, \bar{c} = {\rm Tr}[{\rm diag}(c)]/d$. Explicit expressions for $\mathcal{R}_0(\lambda, \rho), A_1(\lambda)$ and $A_2(\lambda)$ are given in (55) in Appendix E. We note that $\mathbb{E}_{\theta^*_{\rm pop}}\mathcal{R}_{\rm eq}\left(\Sigma, \theta^*_{\rm pop}, D, \lambda\right)$ is even in ρ, hence the odd powers of ρ are absent from (16). Now, we define optimal regularization and risk as

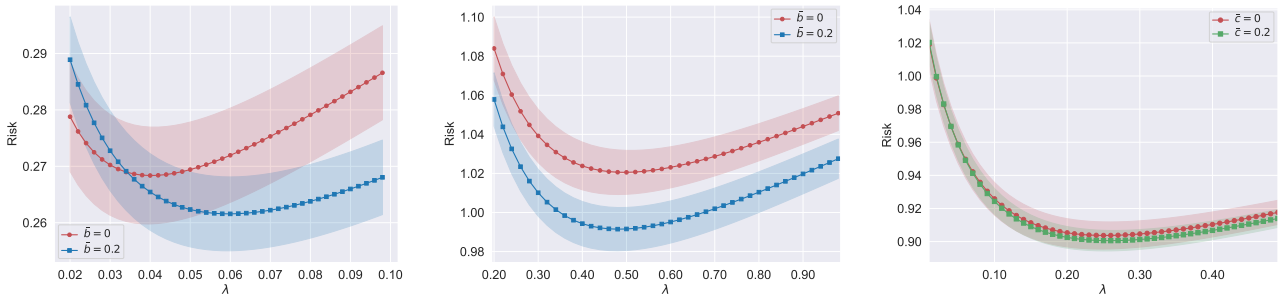
$$\lambda^*_{\rm eq}(D, \rho) := \arg\min_{\lambda\ge 0}\widetilde{\mathcal{R}}(D, \lambda, \rho),$$
$$\mathcal{R}^*_{\rm eq}(D, \rho) := \min_{\lambda\ge 0}\widetilde{\mathcal{R}}(D, \lambda, \rho). \tag{17}$$

Our goal is to characterize the performative effect on $\lambda^*_{\rm eq}(D, \lambda, \rho), \mathcal{R}^*_{\rm eq}(D, \lambda, \rho)$ and, to do so, we compare these quantities to their values when $D = 0$, defined as

$$\lambda^*_{{\rm eq}, D=0}(\rho) := \arg\min_{\lambda\ge 0}\mathcal{R}_0(\lambda, \rho), \mathcal{R}^*_{\rm eq}(\rho) = \min_{\lambda\ge 0}\mathcal{R}_0(\lambda, \rho). \tag{18}$$

This is formalized by the result below whose proof is deferred to Appendix E.

Theorem 5.2 (Optimal regularization – over-parameterized)**.** *In the setting described above, we have*



(a) $\kappa = 1.1, \sigma = 0.2, \rho = 0$ *(b)* $\kappa = 1.1, \sigma = 0.7, \rho = 0$ *(c)* $\kappa = 2, \sigma = 0.5, \rho = 0.5$

Figure 3. Excess risk as a function of ridge regularization λ with Gaussian data, for $n = 4000$, $\Sigma_1 = \Sigma_2 = I_d$, $\Sigma_{12} = \rho I_d$, entries of b equal to $\bar{b}$, and entries of c equal to $\bar{c}$. Values are computed from 20 i.i.d. trials, with error band at 1 standard deviation. We perform 5 steps of RRM to approximate the fixed point, as in the simulation setup of Section 6. (a) In the low-noise regime ($\sigma = 0.2$), taking $\bar{b} = 0.2$ instead of $\bar{b} = 0$ *increases the optimal regularization* and *reduces the optimal risk*. We set $\bar{c} = 0$ to emphasize the dependence on $\bar{b}$. (b) In the large-noise regime ($\sigma = 0.7$), taking $\bar{b} = 0.2$ instead of $\bar{b} = 0$ *reduces both optimal regularization and optimal risk*. As in (a), we set $\bar{c} = 0$. (c) Taking $\bar{c} = 0.2$ instead of $\bar{c} = 0$ *reduces the optimal risk*, although the impact of $\bar{c}$ is less pronounced. We set $\bar{b} = 0$ to emphasize the dependence on $\bar{c}$.

$$\lambda^*_{\text{eq}}(D,\rho) = \lambda^*_{\text{eq},D=0}(\rho) + \bar{b}(B_1(\sigma,\kappa) + O(\rho^2)) + \bar{c}\rho^2(C_1(\sigma,\kappa) + O(\rho^2)) + O(\bar{b}^2 + \bar{c}^2), \tag{19}$$

$$\mathcal{R}^*_{\text{eq}}(D,\rho) = \mathcal{R}^*_{\text{eq}}(\rho) + \bar{b}(B_2(\sigma,\kappa) + O(\rho^2)) + \bar{c}\rho^2(C_2(\sigma,\kappa) + O(\rho^2)) + O(\bar{b}^2 + \bar{c}^2), \tag{20}$$

where the functions $B_1(\sigma,\kappa), B_2(\sigma,\kappa), C_1(\sigma,\kappa), C_2(\sigma,\kappa)$ *depend only on* σ, κ *and they are explicitly given in* (60)-(62). *Furthermore, these functions satisfy*

$$B_1(\sigma,\kappa) \geq 0 \quad \textit{for } 0 \leq \sigma \leq \sigma_{B_1}(\kappa),\ \kappa > 1, \tag{21}$$
$$B_1(\sigma,\kappa) \leq 0 \quad \textit{for } \sigma > \sigma_{B_1}(\kappa),\ \kappa > 1, \tag{22}$$
$$C_1(\kappa,\sigma) \leq 0 \quad \textit{for } \sigma \geq 0, \kappa \geq 2, \tag{23}$$
$$B_2(\kappa,\sigma) \leq 0 \quad \textit{for } \sigma \geq 0, \kappa > 1, \tag{24}$$
$$C_2(\kappa,\sigma) \leq 0 \quad \textit{for } \sigma \geq 0, \kappa > 1, \tag{25}$$

with $\sigma^2_{B_1}(\kappa) = 1/2 - 7\kappa^{-1}/18 + O(\kappa^{-2})$.

In words, (19) gives a quantitative comparison between optimal regularization with performative effect ($\lambda^*_{\text{eq}}(D,\rho)$) and without it ($\lambda^*_{\text{eq},D=0}(\rho)$). Similarly, (20) compares optimally-regularized risks $\mathcal{R}^*_{\text{eq}}(D,\rho)$ and $\mathcal{R}^*_{\text{eq}}(\rho)$ respectively with and without performativity. The study of the signs of the auxiliary functions $B_1(\sigma,\kappa), B_2(\sigma,\kappa), C_1(\sigma,\kappa), C_2(\sigma,\kappa)$ leads to the considerations below:

- Equations (21) and (22) imply that *(i)* if the noise variance σ^2 is small, then the optimal regularization moves in the same direction as the performative effect on the predictive features; and *(ii)* if the noise variance is large, the effect is reversed and the optimal regularization moves in the opposite direction to the performative effect. From a Bayesian perspective, informally, this acts as the noise level controlling the model's confidence: if the noise variance increases, the model moves back towards its "prior", and thus the shift in the regularization due to performativity decreases. This is illustrated in Figures 3a and 3b.
- Equation (23) implies that, when $\kappa \geq 2$, the optimal regularization moves in the opposite direction to the performative effect on the spurious features. This effect is however significantly attenuated by the factor ρ^2 multiplying $\bar{c}$ in (19) and, as such, it is hardly noticeable both with Gaussian data (as considered in this section) and in real-world settings (as considered in Section 6).
- Equation (24) implies that, when performativity reinforces existing trends ($\bar{b} > 0$), the optimally-regularized risk improves in the presence of a performative effect on the predictive features. This occurs regardless of the size of the noise variance, and it is illustrated in Figures 3a and 3b. Instead, when performativity dampens existing trends ($\bar{b} < 0$), the effect is reversed and the optimal risk worsens.
- Finally, Equation (25) implies that the dependence of the optimally-regularized risk on the performative effect on the spurious features is analogous: the optimal risk decreases when $\bar{c} > 0$, and increases when $\bar{c} < 0$. However, as for $\lambda^*_{\text{eq}}(D,\rho)$, the impact of performativity on $\mathcal{R}^*_{\text{eq}}(D,\rho)$ is less pronounced for the spurious features, due to the factor ρ^2 multiplying $\bar{c}$ in (20). This is illustrated in Figure 3c.

Evaluating the model on the shifted distribution. By following similar steps, we can also provide an analysis in the over-parameterized setting of the model tested on the shifted distribution (i.e., on the same distribution used to train it). We defer the details to Appendix F and summarize the main results below. Theorem F.1 is the equivalent of

Theorem 5.1, and it provides a deterministic equivalent of the performative fixed point. Lemma F.2 specializes the risk expression from Theorem F.1 to a covariance of the form $\Sigma = \begin{bmatrix} I_d & \rho I_d \\ \rho I_d & I_d \end{bmatrix}$, and it is the equivalent of Lemma E.1 (which gives (16)). Next, we provide expressions for the optimal τ in Lemma F.3, for the optimal regularization parameter in (74)-(75), and for the optimal risk in Corollary F.4. Lemma F.5 shows that the optimal regularization moves in the same direction as the performative effect on the predictive features, as it is the case for testing over $\mathcal{D}(\theta = 0)$ provided that the noise variance is enough. Finally, Lemmas F.6 and F.7 show that the optimally regularized risk worsens in the presence of a performative effect (either on the predictive or on the spurious features) that reinforces existing trends ($\bar{b}, \bar{c} > 0$). This is in contrast with the behavior of the optimally regularized risk tested over $\mathcal{D}(\theta = 0)$, which instead decreases when either $\bar{b} > 0$ or $\bar{c} > 0$.

6. Numerical experiments

In this section, we test the effect of regularization and performative shifts on real data. Since no dataset currently provides a real performative shift, the shift must be encoded synthetically. In practice, we take a real-world dataset, randomly split the samples across time steps, train a model on one split, compute the parameter θ, and then shift the samples of the next split according to the theoretical model. This methodology follows previous work (Perdomo et al., 2020; Hardt & Mendler-Dünner, 2023; Zezulka & Genin, 2023). These experiments allow us to test whether the theory remains predictive when *(i)* the data is not i.i.d. random and θ^*_{pop} is fixed by the task, when *(ii)* the true relationship between the feature and the target is likely non-linear, and when *(iii)* the regularization is not the ridge penalty. The code to reproduce the experiments is available at `https://github.com/totilas/regularization-vs-perf`.

We consider two datasets. First, we use the Housing dataset,[1] where the goal is to predict house prices from housing features and local demographics. We follow the methodology of Cyffers et al. (2024) to choose performative features. The dataset has 8 features and 20,640 datapoints, which we split into five folds: four for training and one for test. Four training steps suffice experimentally to reach the fixed point, which is consistent with the theory, where the first-order effect stabilizes after only two iterations. Second, we use the Law School Admission Council (LSAC) dataset,[2] where the default task is to predict bar passage from demographic features and previous grades. We change the target to GPA to maintain a regression task, and randomly choose features affected by performativity. After dropping redundant columns or those too correlated with GPA, the dataset has 22 features and 20,427 samples, which again we split in five folds. We report detailed pre-processing, parameters and data covariance in Appendix G.

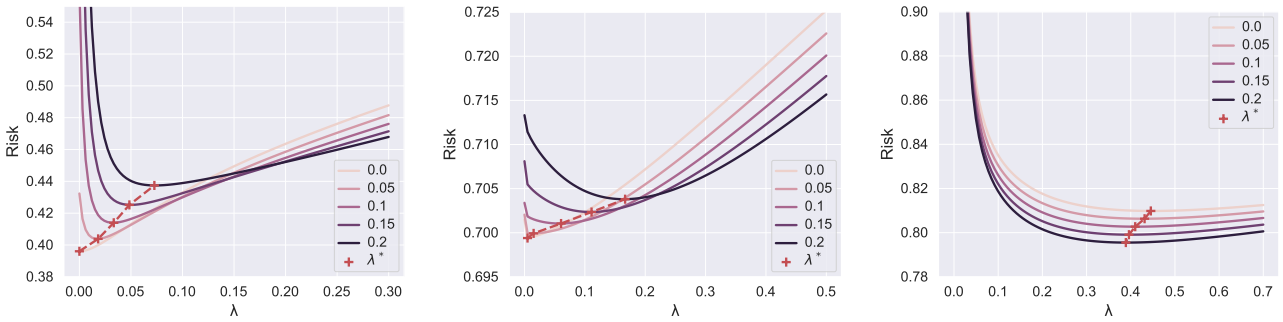
When $n = 4000$, both for the Housing (Figure 4a) and the LSAC (Figure 4b) dataset, we note that *(i)* the optimal regularizer increases proportionally to $\bar{b}$, and *(ii)* the optimally-regularized risk becomes worse as $\bar{b}$ grows. This can be attributed to the fact that $n \gg d$, and it is consistent with our theoretical results in the population setting (Corollary 4.2). In contrast, when training with very few samples on LSAC (Figure 4c), the behavior of the regularized risk follows the predictions of Theorem 5.1 for the proportional setting in the large-noise regime: as $\bar{b}$ grows, the optimal regularizer gets smaller and the risk improves. Note that, even if Figures 4b and 4c consider the same dataset, the ranges of excess risk and regularizer are not the same due to the different sample sizes. We did not find numerical evidence for the role of c, suggesting that its effect may be dominated by data noise, consistent with our theoretical findings on the limited impact of spurious features.

In Appendix A we showcase the results of the same experiments when the ridge regularization is replaced with *(i)* dropout (Figure 6), *(ii)* Lasso regularization (Figure 7), and *(iii)* elastic net (Figure 8). This demonstrates that the relationship between regularization and performative strength persists beyond ridge regression.

To further assess whether our findings extend beyond linear models, we also run an experiment with neural networks. We follow the strategic classification setting of Mofakhami et al. (2023), using the `GiveMeSomeCredit` dataset through the `whynot` package. In this setting, the strength of the performative effect is controlled by a parameter δ: when a data point would receive a negative classification under the previous model, it may strategically modify its manipulable features by copying features from another data point, with probability depending on δ. We build on the publicly available code of Mofakhami et al. (2023), use the largest neural network reported there, and keep the preprocessing, learning rate, and other hyperparameters unchanged. For each value of δ, we run the dynamics until convergence for several values of the ℓ_2 regularization parameter λ, sharing randomness across values of δ and λ. Each setting is run several times, and we report the final test accuracy with standard deviations. As shown in Figure 5, the same qualitative behavior appears: ℓ_2 regularization mitigates the sharp drop in accuracy induced by the performative shift, and the optimal amount of regularization increases with the strength of the performative effect.

[1] `https://www.openml.org/d/823`

[2] `https://storage.googleapis.com/lawschool_dataset/bar_pass_prediction.csv`



(a) Housing ($n = 4000$) *(b)* LSAC ($n = 4000$) *(c)* LSAC ($n = 100$)

Figure 4. Excess risk as a function of ridge regularization λ in real-world datasets (Housing, LSAC). Different curves (in different colors) correspond to different values of $\bar{b} \in \{0, 0.05, 0.1, 0.15, 0.2\}$, and we connect with a red dashed line the optima of the risk for various choices of $\bar{b}$. The plots in (a)-(b) use $n = 4000$ data points at each training step, which corresponds to the population setting ($n \gg d$); the plot in (c) uses $n = 100$, a value closer to the number of features $d = 22$.

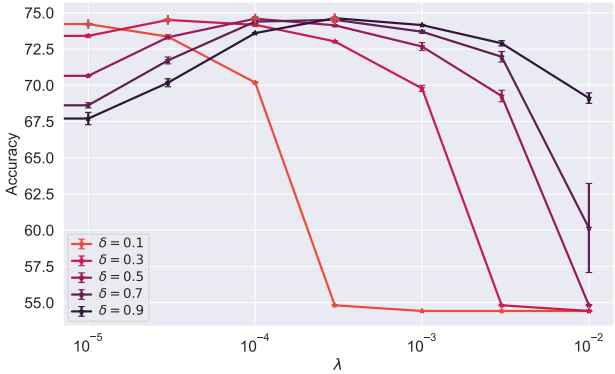


Figure 5. Final test accuracy of a neural network under performative shifts of increasing strength δ, as a function of the ℓ_2 regularization parameter λ. Regularization mitigates the loss in accuracy caused by the performative shift, and the best regularization level increases with δ.

7. Conclusions

In this work, we demonstrate that regularization and performative effect are strongly related, as one can partially cancel out the other. In the population regime, the excess risk is worsened by performative effects. However, optimal ridge regularization mitigates this issue, especially when the data is isotropic and the entries of the vector modeling performativity have little variability. In the proportional regime, we provide a deterministic equivalent of the performative fixed point for random data. This in turn unveils a remarkable phenomenology: in contrast with the population setting, the optimal risk improves when performativity reinforces existing trends; furthermore, the optimal regularization follows the direction of the performative effect on the predictive features when the noise is small, while it goes in the opposite direction when the noise is large. Although the theoretical results focus on random data and a linear target model, our experiments on real-world data follow the theoretical predictions, suggesting their generality. Overall, these findings indicate that regularization could help in a wider range of scenarios, which we leave for future work. Beyond studying more complex data or models, interesting directions include the impact of other forms of regularization, such as early stopping or pruning, to mitigate performative effects.

Impact Statement

This work studies performative learning, which contributes to the study of the long-term impact of machine learning when deployed. Taking into account the risk of modifying the data distribution is likely to improve the impact of machine learning. However, the two evaluations we propose, namely evaluation on the initial distribution and evaluation on the final distribution, can both lead to unfortunate situations. Evaluating on the initial distribution might amount to turning a blind eye to the evolution of society, while the second might encourage intentionally steering the data distribution toward less diverse distributions that are easier to tackle. We mitigate this risk by deriving results for both regimes, which allows readers to pick the setting that makes sense for a given use case, in the absence of alternative evaluation criteria currently proposed for performative learning.

The choice to use high-dimensional tools to study performative learning should open interesting research directions and contribute to widening the tools available to study performativity. High-dimensional theory has proven useful for better explaining deep learning in classical machine learning, and we hope for similar benefits in the context of performative learning.

The focus on regularization offers rather practical and actionable takeaways from the paper, advocating for regularization as a convenient and agnostic mitigation for performativity. We trust readers to understand that regularization is only one tool among many for building more resilient models.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/COE12. For the purpose of open access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. Alireza Mirrokni was an intern at ISTA while working on this project. Edwige Cyffers is supported by the National Research Agency under France 2030, reference "ANR-23-IACL-0008" and this work was partially done while visiting the Simons Institute for the Theory of Computing.

Conflict of Interest

The authors have no conflict of interest.

References

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48), 2020.

Bechavod, Y., Ligett, K., Wu, S., and Ziani, J. Gaming helps! learning from strategic interactions in natural dynamics. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, volume 130, 2021.

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

Ben-Dov, O., Fawkes, J., Samadi, S., and Sanyal, A. The role of learning algorithms in collective action. In *International Conference on Machine Learing (ICML)*, volume 235, 2024.

Bombari, S. and Mondelli, M. Spurious correlations in high dimensional regression: The roles of regularization, simplicity bias and over-parameterization. In *International Conference on Machine Learing (ICML)*, 2025.

Bracale, D., Maity, S., Banerjee, M., and Sun, Y. Learning the distribution map in reverse causal performative prediction, 2025.

Brown, G., Hod, S., and Kalemaj, I. Performative prediction in a stateful world. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*. PMLR, 2022.

Cai, T. T. and Wei, H. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1), February 2021. ISSN 0090-5364. doi: 10.1214/20-aos1949. URL http://dx.doi.org/10.1214/20-AOS1949.

Chang, X., Li, Y., Oymak, S., and Thrampoulidis, C. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. In *Conference on Artificial Intelligence (AAAI)*, volume 35, 2021.

Clara, G., Langer, S., and Schmidt-Hieber, J. Dropout regularization versus l2-penalization in the linear model. *Journal of Machine Learning Research*, 25(204), 2024.

Cyffers, E., Pydi, M. S., Atif, J., and Cappé, O. Optimal classification under performative distribution shift. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Demirel, B., Kong, L., Zhang, K., Karaletsos, T., Mendler-Dünner, C., and Locatello, F. Adjusting pretrained backbones for performativity, 2024.

Deng, Z., Kammoun, A., and Thrampoulidis, C. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2), 2022.

Drusvyatskiy, D. and Xiao, L. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2), May 2023.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency (FAccT)*, volume 81, 2018.

Fawzi, A., Fawzi, O., and Frossard, P. Analysis of classifiers' robustness to adversarial perturbations. *Machine learning*, 107(3):481–508, 2018.

Han, Q. and Xu, X. The distribution of ridgeless least squares interpolators. *arXiv preprint arXiv:2307.02044*, 2023.

Hardt, M. and Mendler-Dünner, C. Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608*, 2023.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2), 2022.

Ildiz, M. E., Gozeten, H. A., Taga, E. O., Mondelli, M., and Oymak, S. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025.

Izzo, Z., Zou, J., and Ying, L. How to learn when data gradually reacts to your model. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, volume 151, 2022.

Jagadeesan, M., Zrnic, T., and Mendler-Dünner, C. Regret minimization with performative feedback. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learing (ICML)*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Jain, A., Montanari, A., and Sasoglu, E. Scaling laws for learning with real and surrogate data. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Kirev, I., Baltadzhiev, L., and Konstantinov, N. Pac learnability in the presence of performativity, 2026. URL https://arxiv.org/abs/2510.08335.

Kolossov, G., Montanari, A., and Tandon, P. Towards a statistical theory of data selection under weak supervision. In *International Conference on Learning Representations (ICLR)*, 2024.

Li, Q., Yau, C.-Y., and Wai, H.-T. Multi-agent performative prediction with greedy deployment and consensus seeking agents. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Mallinar, N. R., Zane, A., Frei, S., and Yu, B. Minimum-norm interpolation under covariate shift. In *International Conference on Machine Learing (ICML)*, volume 235, 2024.

Miller, J., Perdomo, J. C., and Zrnic, T. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learing (ICML)*, volume 139, 2021.

Mofakhami, M., Mitliagkas, I., and Gidel, G. Performative prediction with neural networks. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11079–11093. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/mofakhami23a.html.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *Annals of statistics*, 53(2), 2025.

Morgenstern, O. *Wirtschaftsprognose: Eine Untersuchung ihrer Voraussetzungen und Möglichkeiten*. Springer, 1928. ISBN 978-3709121139.

Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M., and Ratliff, L. J. Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research*, 24(202), 2023.

Pan, A., Jones, E., Jagadeesan, M., and Steinhardt, J. Feedback loops with language models drive in-context reward hacking. In *International Conference on Machine Learing (ICML)*, volume 235, 2024.

Patil, P., Du, J.-H., and Tibshirani, R. J. Optimal ridge regularization for out-of-distribution prediction. In *International Conference on Machine Learing (ICML)*, volume 235, 2024.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learing (ICML)*, volume 119, 2020.

Rezaei, P., Kovacevic, F., Locatello, F., and Mondelli, M. High-dimensional analysis of synthetic data selection. *arXiv preprint arXiv:2510.08123*, 2025.

Ribeiro, A. H., Zachariah, D., Bach, F., and Schön, T. B. Regularization properties of adversarially-trained linear regression. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Richards, D., Mourtada, J., and Rosasco, L. Asymptotics of ridge (less) regression under general source condition. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, volume 130, 2021.

Song, Y., Bhattacharya, S., and Sur, P. Generalization error of min-norm interpolators in transfer learning. *arXiv preprint arXiv:2406.13944*, 2024.

Taori, R. and Hashimoto, T. Data feedback loops: Model-driven amplification of dataset biases. In *International Conference on Machine Learing (ICML)*, volume 202, 2023.

Tsigler, A. and Bartlett, P. L. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24 (123), 2023.

Tsoy, N., Kirev, I., Rahimiyazdi, N., and Konstantinov, N. On the impact of performative risk minimization for binary random variables. *International Conference on Machine Learing (ICML)*, abs/2502.02331, 2025.

Ursu, R. M. The power of rankings: Quantifying the effects of rankings on online consumer search and choice. *SSRN Electronic Journal*, 2015.

Wang, X., Yau, C.-Y., and Wai, H. T. Network effects in performative prediction games. In *International Conference on Machine Learing (ICML)*, volume 202, 2023a.

Wang, Z., Liu, C., Parisini, T., Zavlanos, M. M., and Johansson, K. H. Constrained optimization with decision-dependent distributions, 2023b.

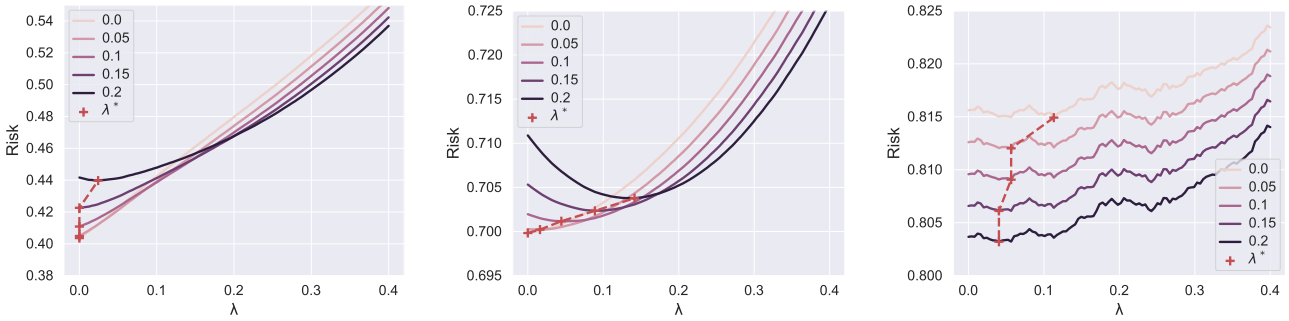
Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Yang, F., Zhang, H. R., Wu, S., Re, C., and Su, W. J. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *Journal of Machine Learning Research*, 26(113), 2025.

Zezulka, S. and Genin, K. Performativity and Prospective Fairness. *arXiv preprint arXiv:2310.08349*, 2023.

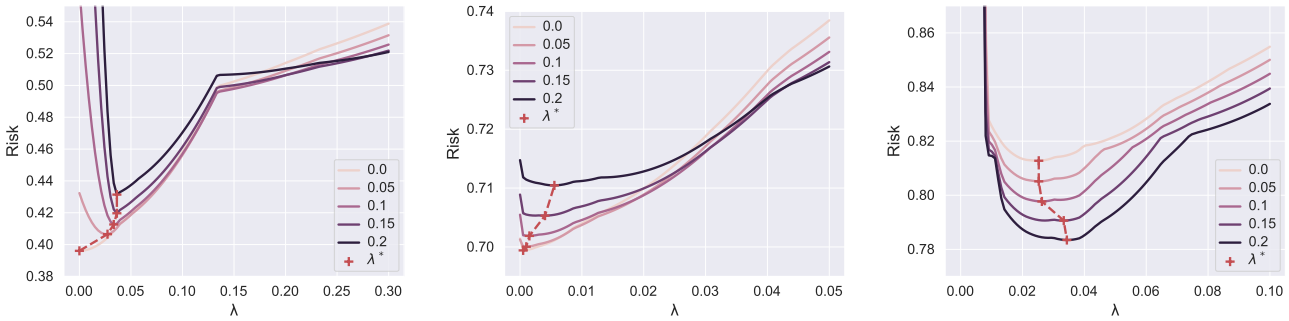
Zhu, Y., Zhang, J., Gangrade, A., and Scott, C. Label noise: ignorance is bliss. In *Conference on Neural Information Processing Systems (NeurIPS)*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

A. Additional figures



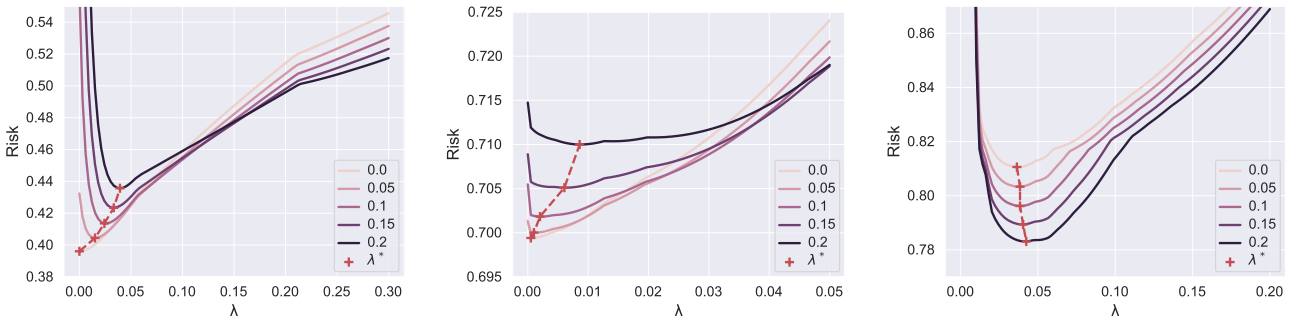
(a) Housing ($n = 4000$) *(b)* LSAC ($n = 4000$) *(c)* LSAC ($n = 100$)

Figure 6. Same experiments as in Figure 4, using dropout regularization instead of ridge. Although dropout induces a more complex form of regularization (even for linear regression (Clara et al., 2024)), the findings are similar. For the Housing dataset in the plot (a), dropout provides benefits only for large performative effects, which can be explained by the small dimension $d = 8$. For LSAC in the plots (b)-(c), the observed effects are the same as for ridge regularization. Finally, for LSAC in the plot (c), being in the proportional setting with large noise, the optimal dropout rate increases with the strength of the performative effect and the optimal risk get smaller, as it was the case for ridge regularization.



(a) Housing ($n = 4000$) *(b)* LSAC ($n = 4000$) *(c)* LSAC ($n = 100$)

Figure 7. Same experiments as in Figure 4, using Lasso regularization instead of ridge. Similar conclusions hold: the performative effect worsens performance in the population regime and helps in the proportional regime; the optimal regularizer continues to be non-decreasing with the strength of the performative effect in the population regime. The optimal regularizer seems close to constant in the proportional regime, potentially due to the number of features being too small to observe a dependency on the Lasso regularization. Indeed, with 22 features, the support of θ cannot change smoothly as the regularization increases.



(a) Housing ($n = 4000$) *(b)* LSAC ($n = 4000$) *(c)* LSAC ($n = 100$)

Figure 8. Same experiments as in Figure 4, using an elastic net regularization with an equal ratio between ℓ_1 and ℓ_2 penalties instead of ridge. As intuition suggests, the results lie between those obtained with ridge and Lasso regularization.

B. Additional proofs for Section 4

This appendix contains the missing proofs for Section 4. We start with the convergence at exponential rate to the fixed point θ^∞ in (5). Then, we prove Theorem 4.1 giving the first-order approximation of the risk, as well as the expression in (10) giving the higher-order approximation of the risk. Finally, we prove the upper bound on $\|F\|_{\text{op}}$ in (27).

Lemma B.1. *The sequence $(\theta_k)_k$ converges to the fixed point*

$$\theta^\infty = (I_p + \lambda\Sigma^{-1} - D)^{-1}\theta^*_{\text{pop}}.$$

Moreover, for any $\varepsilon \in (0,1)$, if we start at $\theta_0 = 0$, after at most

$$k_\varepsilon = \left\lceil \frac{\ln(1/\varepsilon)}{\ln\left(1/\left(\frac{\|\Sigma\|_{\text{op}}}{\|\Sigma\|_{\text{op}}+\lambda}\max\{\|b\|_\infty, \|c\|_\infty\}\right)\right)} \right\rceil$$

iterations, the relative error $\frac{\|\theta^{k_\varepsilon} - \theta^\infty\|_2}{\|\theta^\infty\|_2}$ is smaller than ε.

Proof. Denoting $T = (\Sigma + \lambda I_p)^{-1}\Sigma D$, the recurrence relation is

$$\theta^k = T\theta^{k-1} + (\Sigma + \lambda I_p)^{-1}\Sigma\theta^*_{\text{pop}}.$$

When going to the limit, $\sum_i T^i \to (I_p - T)^{-1}$. The convergence requires the matrix T to have smaller eigenvalues than one, which is guaranteed by $\|b\|_\infty$ and $\|c\|_\infty$ being smaller than one. Thus, we have

$$\theta^\infty = (I_p - T)^{-1}(\Sigma + \lambda I_p)^{-1}\Sigma\theta^*_{\text{pop}}.$$

Noticing that $I_p = (\Sigma + \lambda I_p)^{-1}(\Sigma + \lambda I_p)$ and using the definition of T gives the expression of θ_∞.

Let $e_k = \theta^k - \theta^\infty$. Using $\theta^\infty = T\theta^\infty + (\Sigma + \lambda I_p)^{-1}\Sigma\,\theta^*_{\text{pop}}$, we have

$$e_k = \theta^k - \theta^\infty = T\theta^{k-1} + (\Sigma + \lambda I_p)^{-1}\Sigma\theta^*_{\text{pop}} - \theta^\infty = T(\theta^{k-1} - \theta^\infty) = Te_{k-1}.$$

Thus,

$$e_k = T^k e_0 = -T^k\theta^\infty \implies \|e_k\|_2 \le \|T\|^k_{\text{op}}\|\theta^\infty\|_2 \implies \frac{\|e_k\|_2}{\|\theta^\infty\|_2} = \frac{\|\theta^k - \theta^\infty\|_2}{\|\theta^\infty\|_2} \le \|T\|^k_{\text{op}}.$$

Consequently, $\|T\|^k \le \varepsilon$ suffices, i.e., $k \ge \ln(1/\varepsilon)/\ln(1/\|T\|_{\text{op}})$. We finally note that

$$\|T\|_{\text{op}} \le \|(\Sigma + \lambda I_p)^{-1}\Sigma\|_{\text{op}}\|D\|_{\text{op}} = \frac{\|\Sigma\|_{\text{op}}}{\|\Sigma\|_{\text{op}} + \lambda}\max\{\|b\|_\infty, \|c\|_\infty\}.$$

Combining the last two inequalities gives the wanted convergence rate. □

Proof of Theorem 4.1 and of the higher-order approximation in (10). We start by computing the Taylor expansion:

$$A = (\Sigma + \lambda I_p - \Sigma D)^{-1}\Sigma - I_p = (I_p - (D - \lambda\Sigma^{-1}))^{-1} - I_p = \sum_{i=1}^{\infty}(D - \lambda\Sigma^{-1})^i = \sum_{i=1}^{\infty} F^i.$$

Let us define $A^{(k)} = \sum_{i=1}^{k}(D - \lambda\Sigma^{-1})^i$. For the two first orders, we have:

$$A^{(1)\top}\Sigma A^{(1)} = (D - \lambda\Sigma^{-1})\Sigma(D - \lambda\Sigma^{-1}) = D\Sigma D - 2\lambda D + \lambda^2\Sigma^{-1}.$$

This is independent of c and gives the simple formula

$$R^{(1)}(\lambda) = \frac{1}{d}\operatorname{Tr}(\operatorname{diag}(b^2)\Sigma_1) - 2\lambda\bar{b} + \frac{1}{d}\lambda^2\operatorname{Tr}(S_1),$$

where $\bar{b} := \frac{1}{d}\operatorname{Tr}[\operatorname{diag}(b)] = \frac{1}{d}\sum_{i=1}^{d} b_i$, $b^2 := [b_1^2, \dots, b_d^2] \in \mathbb{R}^d$ and $S_1 = (\Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21})^{-1}$ denotes the Schur complement of Σ. We go further in the expansion to recover (10):

$$\begin{aligned}
A^{(2)\top}\Sigma A^{(2)} &= A^{(1)\top}\Sigma A^{(1)} + A^{(1)\top}\Sigma\left(A^{(2)} - A^{(1)}\right) + \left(A^{(2)} - A^{(1)}\right)^\top \Sigma A^{(1)} \\
&\quad + \left(A^{(2)} - A^{(1)}\right)^\top \Sigma\left(A^{(2)} - A^{(1)}\right) \\
&= D\Sigma D + D^2\Sigma D + D\Sigma D^2 - \lambda\left[D\Sigma D\Sigma^{-1} + \Sigma^{-1}D\Sigma D + 2D + 4D^2\right] \\
&\quad + \lambda^2\left[\Sigma^{-1} + 3(\Sigma^{-1}D + D\Sigma^{-1})\right] - 2\lambda^3\Sigma^{-2} + O\left(\|F\|_{\mathrm{op}}^4\right).
\end{aligned}$$

The final formula results from taking the trace of the first block. We write the matrix product block per block to prove that

$$\operatorname{Tr}\left[(D\Sigma D\Sigma^{-1} + \Sigma^{-1}D\Sigma D)_1\right] = 2\operatorname{Tr}\left[\operatorname{diag}(b)\Sigma_1\operatorname{diag}(b)S_1\right] + 2\operatorname{Tr}\left[\operatorname{diag}(b)\Sigma_{12}\operatorname{diag}(c)S_{21}\right],$$

where $S_{21}^\top = -(\Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_2^{-1}$. This concludes the proof. □

Lemma B.2. *Let $F = D - \lambda\,\Sigma^{-1}$. Then, we have that*

$$\|F\|_{\mathrm{op}} \le \max\left(\left|\max_{1\le i\le d}\{b_i, c_i\} - \frac{\lambda}{\lambda_{\max}(\Sigma)}\right|, \left|\min_{1\le i\le d}\{b_i, c_i\} - \frac{\lambda}{\lambda_{\min}(\Sigma)}\right|\right). \tag{26}$$

Proof. By Weyl's inequalities for Hermitian matrices,

$$\begin{aligned}
\lambda_{\max}(F) &\le \lambda_{\max}(D) - \lambda\lambda_{\min}(\Sigma^{-1}) = \max_{1\le i\le d}\{b_i, c_i\} - \frac{\lambda}{\lambda_{\max}(\Sigma)}, \\
\lambda_{\min}(F) &\ge \lambda_{\min}(D) - \lambda\lambda_{\max}(\Sigma^{-1}) = \min_{1\le i\le d}\{b_i, c_i\} - \frac{\lambda}{\lambda_{\min}(\Sigma)}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|F\|_{\mathrm{op}} &= \max\{|\lambda_{\max}(F)|, |\lambda_{\min}(F)|\} \\
&\le \max\left(\left|\max_{1\le i\le d}\{b_i, c_i\} - \frac{\lambda}{\lambda_{\max}(\Sigma)}\right|, \left|\min_{1\le i\le d}\{b_i, c_i\} - \frac{\lambda}{\lambda_{\min}(\Sigma)}\right|\right),
\end{aligned}$$

and the equality happens if and only if Σ and D are simultaneously diagonalizable. □

Finally, we can rewrite this result as

$$\|F\|_{\mathrm{op}} \le \max\left(\begin{array}{c}\left|\max_{1\le i\le d}\{b_i, c_i\} - \frac{\lambda}{\|\Sigma\|_{\mathrm{op}}}\right|, \\ \left|\min_{1\le i\le d}\{b_i, c_i\} - \frac{\lambda}{\lambda_{\min}(\Sigma)}\right|\end{array}\right), \tag{27}$$

where $\lambda_{\min}(\Sigma)$ is the smallest eigenvalue of Σ, since $\lambda_{\max}(\Sigma) = \|\Sigma\|_{\mathrm{op}}$, due to Σ being a covariance matrix and, hence, positive semidefinite.

C. Proof of Theorem 5.1

Deterministic equivalent for $\mathcal{R}_1(\Sigma, \theta_1, \theta^*_{\text{pop}})$. Let $\mathcal{R}_k(\Sigma, \theta_k, \theta^*_{\text{pop}})$ be the excess risk of the estimator θ_k given by (12), i.e.,

$$\mathcal{R}_k(\Sigma, \theta_k, \theta^*_{\text{pop}}) = \left\|\theta_k - \theta^*_{\text{pop}}\right\|^2_\Sigma .$$

Having fixed the initialization θ_0, the only randomness in $\mathcal{R}_1(\Sigma, \theta_1, \theta^*_{\text{pop}})$ comes from $(X^{(0)}, y^{(0)})$. This corresponds to the setting in which one trains from the (deterministic) vector of regression coefficients $\theta^*_{\text{pop}} + D\theta_0$. The following lemma gives a deterministic equivalent for $\mathcal{R}_1(\Sigma, \theta_1, \theta^*_{\text{pop}})$, conditional on θ_0.

Lemma C.1. *Let Assumption 3.1 hold. Let $R > 0$ be a constant such that $\theta^*_{\text{pop}}, \theta_0 \in B_p(R)$. Assume that $\kappa, \sigma, \lambda \in (1/M, M)$ and $\|\Sigma\|_{\text{op}},\ \|\Sigma^{-1}\|_{\text{op}} \le M$ for some constant $M > 1$. Then, there exists a constant $C = C(M, R)$ such that, for any $\delta \in (0, 1/2]$, the following holds*

$$\sup_{\theta^*_{\text{pop}}, \theta_0 \in B_p(R)} \Pr\left(\left|\mathcal{R}_1(\Sigma, \theta_1, \theta^*_{\text{pop}}) - \mathcal{R}^{(1)}_{\text{eq}}\left(\Sigma, \theta_0, \theta^*_{\text{pop}}\right)\right| \ge \delta\right) \le Cpe^{-p\delta^4/C}, \tag{28}$$

with probability at least $1 - Cpe^{-p\delta^4/C}$, where

$$\begin{aligned}\mathcal{R}^{(1)}_{\text{eq}}\left(\Sigma, \theta_0, \theta^*_{\text{pop}}\right) &= \left\|\left(\Sigma + \tau I_p\right)^{-1}\Sigma(\theta^*_{\text{pop}} + D\theta_0) - \theta^*_{\text{pop}}\right\|^2_\Sigma \\ &\quad + \kappa \operatorname{Tr}\left[\Sigma^2\left(\Sigma + \tau I_p\right)^{-2}\right] \frac{\sigma^2 + \tau^2\left\|\left(\Sigma + \tau I_p\right)^{-1}(\theta^*_{\text{pop}} + D\theta_0)\right\|^2_\Sigma}{p - \kappa \operatorname{Tr}\left[\Sigma^2\left(\Sigma + \tau I_p\right)^{-2}\right]},\end{aligned} \tag{29}$$

and τ is the unique solution of (15).

Proof. Note that we are generating labels using $\theta_0^{\text{perfo}} := \theta^*_{\text{pop}} + D\theta_0$ as a vector of regression coefficients. Thus, we can apply Theorem 3 by (Ildiz et al., 2025) (which utilizes the non-asymptotic characterization of the minimum norm interpolator by (Han & Xu, 2023)), replacing β^s with θ_0^{perfo} in that statement. This gives that (28) holds with $\mathcal{R}^{(1)}_{\text{eq}}\left(\Sigma, \theta_0, \theta^*_{\text{pop}}\right)$ replaced by $\widetilde{\mathcal{R}}^{(1)}_{\text{eq}}\left(\Sigma, \theta^*_{\text{pop}}, \theta_0^{\text{perfo}}\right)$ defined as

$$\widetilde{\mathcal{R}}^{(1)}_{\text{eq}}\left(\Sigma, \theta^*_{\text{pop}}, \theta_0^{\text{perfo}}\right) = \mathbb{E}_{g^{(1)}}\left[\left\|X^{(1)}\left(\Sigma, \theta_0^{\text{perfo}}, g^{(1)}\right) - \theta^*_{\text{pop}}\right\|^2_\Sigma\right], \tag{30}$$

where

$$X^{(1)}\left(\Sigma, \theta_0^{\text{perfo}}, g^{(1)}\right) = (\Sigma + \tau I_p)^{-1}\Sigma\left[\theta_0^{\text{perfo}} + \frac{\Sigma^{-1/2}\gamma^{(1)}(\theta_0^{\text{perfo}})g^{(1)}}{\sqrt{p}}\right], \tag{31}$$

$$\left(\gamma^{(1)}(\theta_0^{\text{perfo}})\right)^2 = \kappa\left(\sigma^2 + \widetilde{\mathcal{R}}^{(1)}_{\text{eq}}\left(\Sigma, \theta_0^{\text{perfo}}, \theta_0^{\text{perfo}}\right)\right), \tag{32}$$

τ is the unique solution of (15) and $g^{(1)} \sim \mathcal{N}(0, I_p)$. By plugging (31) into (30) and computing the expectation with respect to $g^{(1)}$, we get

$$\widetilde{\mathcal{R}}^{(1)}_{\text{eq}}\left(\Sigma, \theta^*_{\text{pop}}, \theta_0^{\text{perfo}}\right) = \left\|\left(\Sigma + \tau I_p\right)^{-1}\Sigma\theta_0^{\text{perfo}} - \theta^*_{\text{pop}}\right\|^2_\Sigma + \frac{\left(\gamma^{(1)}(\theta_0^{\text{perfo}})\right)^2}{p}\operatorname{Tr}\left[\Sigma^2\left(\Sigma + \tau I_p\right)^{-2}\right]. \tag{33}$$

Next, we solve the fixed point equation in $\gamma^{(1)}(\theta_0^{\text{perfo}})$:

$$\begin{aligned}\left(\gamma^{(1)}(\theta_0^{\text{perfo}})\right)^2 &= \kappa\left(\sigma^2 + \widetilde{\mathcal{R}}^{(1)}_{\text{eq}}\left(\Sigma, \theta_0^{\text{perfo}}, \theta_0^{\text{perfo}}\right)\right) \\ &= \kappa\left(\sigma^2 + \left\|\left(\left(\Sigma + \tau I_p\right)^{-1}\Sigma - I_p\right)\theta_0^{\text{perfo}}\right\|^2_\Sigma + \frac{\left(\gamma^{(1)}(\theta_0^{\text{perfo}})\right)^2}{p}\operatorname{Tr}\left[\Sigma^2\left(\Sigma + \tau I_p\right)^{-2}\right]\right) \\ &= \kappa\left(\sigma^2 + \tau^2\left\|\left(\Sigma + \tau I_p\right)^{-1}\theta_0^{\text{perfo}}\right\|^2_\Sigma + \frac{\left(\gamma^{(1)}(\theta_0^{\text{perfo}})\right)^2}{p}\operatorname{Tr}\left[\Sigma^2\left(\Sigma + \tau I_p\right)^{-2}\right]\right).\end{aligned}$$

The last equality comes from

$$I_p - (\Sigma + \tau I_p)^{-1} \Sigma = (\Sigma + \tau I_p)^{-1} (\Sigma + \tau I_p) - (\Sigma + \tau I_p)^{-1} \Sigma = \tau (\Sigma + \tau I_p)^{-1} .$$

Rearranging gives that

$$\left(\gamma^{(1)}(\theta_0^{\text{perfo}})\right)^2 = \kappa \frac{\sigma^2 + \tau^2 \left\| (\Sigma + \tau I_p)^{-1} \theta_0^{\text{perfo}} \right\|_\Sigma^2}{1 - \frac{1}{n} \operatorname{Tr}\left[\Sigma^2 (\Sigma + \tau I_p)^{-2}\right]}, \tag{34}$$

which plugged into (33) gives the desired result. □

We note that the expression in (29) depends on θ_0 and, in fact, it keeps depending on it even after neglecting terms of order $O(\|D\|_{\text{op}}^2)$.

Deterministic equivalent for $\mathcal{R}_2(\Sigma, \theta_2, \theta_{\text{pop}}^*)$. Next, by iterating twice the strategy of Lemma C.1, we derive a deterministic equivalent for $\mathcal{R}_2(\Sigma, \theta_2, \theta_{\text{pop}}^*)$.

Lemma C.2. *Let Assumption 3.1 hold. Let $R > 0$ be a constant such that $\theta_{\text{pop}}^*, \theta_0 \in B_p(R)$. Assume that $\kappa, \sigma, \lambda \in (1/M, M)$ and $\|\Sigma\|_{\text{op}},\ \|\Sigma^{-1}\|_{\text{op}} \leq M$ for some constant $M > 1$. Then, there exists a constant $C = C(M, R)$ such that, for any $\delta \in (0, 1/2]$, the following holds*

$$\sup_{\theta_{\text{pop}}^*, \theta_0 \in B_p(R)} \Pr\left(\left| \mathcal{R}_2(\Sigma, \theta_2, \theta_{\text{pop}}^*) - \mathcal{R}_{\text{eq}}^{(2)}\left(\Sigma, \theta_0, \theta_{\text{pop}}^*\right) \right| \geq \delta \right) \leq C p e^{-p\delta^4/C}, \tag{35}$$

with probability at least $1 - Cpe^{-p\delta^4/C}$, where

$$\begin{aligned}
\mathcal{R}_{\text{eq}}^{(2)}\left(\Sigma, \theta_0, \theta_{\text{pop}}^*\right) &= \left\| (\Sigma + \tau I_p)^{-1} \Sigma D (\Sigma + \tau I_p)^{-1} \Sigma (\theta_{\text{pop}}^* + D\theta_0) - \tau (\Sigma + \tau I_p)^{-1} \theta_{\text{pop}}^* \right\|_\Sigma^2 \\
&+ \kappa \operatorname{Tr}\left[\Sigma (\Sigma + \tau I_p)^{-2} D \Sigma^3 (\Sigma + \tau I_p)^{-2} D\right] \frac{\sigma^2 + \tau^2 \| (\Sigma + \tau I_p)^{-1} (\theta_{\text{pop}}^* + D\theta_0) \|_\Sigma^2}{p - \kappa \operatorname{Tr}\left[\Sigma^2 (\Sigma + \tau I_p)^{-2}\right]} \\
&+ \kappa \operatorname{Tr}\left[\Sigma^2 (\Sigma + \tau I_p)^{-2}\right] \frac{\sigma^2 + \tau^2 \| (\Sigma + \tau I_p)^{-1} \left(\theta_{\text{pop}}^* + D (\Sigma + \tau I_p)^{-1} \Sigma (\theta_{\text{pop}}^* + D\theta_0)\right) \|_\Sigma^2}{p - \kappa \operatorname{Tr}\left[\Sigma^2 (\Sigma + \tau I_p)^{-2}\right]} \\
&+ \kappa^2 \tau^2 \operatorname{Tr}\left[\Sigma^2 (\Sigma + \tau I_p)^{-2}\right] \operatorname{Tr}\left[\Sigma (\Sigma + \tau I_p)^{-2} D \Sigma (\Sigma + \tau I_p)^{-2} D\right] \\
&\qquad \cdot \frac{\sigma^2 + \tau^2 \| (\Sigma + \tau I_p)^{-1} (\theta_{\text{pop}}^* + D\theta_0) \|_\Sigma^2}{\left(p - \kappa \operatorname{Tr}\left[\Sigma^2 (\Sigma + \tau I_p)^{-2}\right]\right)^2},
\end{aligned} \tag{36}$$

and τ is the unique solution of (15).

Proof. The proof extends the argument of Theorem 2 by (Ildiz et al., 2025) to the ridge regression case, and it applies the distributional characterization of the minimum norm interpolator by (Han & Xu, 2023) twice. First, note that $\|\theta_1\|_2$ is bounded by a constant $C_1 = C_1(R, M)$ independent of n, p, with probability at least $C_2 e^{-p/C_2}$, where $C_2 = C_2(R, M)$ is a constant independent of n, p. This follows from a direct adaptation of Proposition 11 by (Ildiz et al., 2025). Define $R' := \max(C_1, \|\theta_{\text{pop}}^*\|_2)$. Then, upon conditioning on θ_1, we can apply Lemma C.1 (after re-defining R to be R'), which gives that, for some constant $C_3 = C_3(R, M)$,

$$\sup_{\theta_{\text{pop}}^*, \theta_1 \in B_p(R')} \Pr\left(\left| \mathcal{R}_2(\Sigma, \theta_2, \theta_{\text{pop}}^*) - \mathcal{R}_{\text{eq}}^{(1)}\left(\Sigma, \theta_1, \theta_{\text{pop}}^*\right) \right| \geq \delta \right) \leq C_3 p e^{-p\delta^4/C_3}, \tag{37}$$

where $\mathcal{R}_{\text{eq}}^{(1)}\left(\Sigma, \theta_1, \theta_{\text{pop}}^*\right)$ is defined in (29) and τ is the unique solution of (15).

We now evaluate the first term in the expression for $\mathcal{R}_{\text{eq}}^{(1)}\left(\Sigma, \theta_1, \theta_{\text{pop}}^*\right)$:

$$\begin{aligned}\mathcal{R}_{\text{eq},1}^{(1)}\left(\Sigma, \theta_1, \theta_{\text{pop}}^*\right) :&= \left\|(\Sigma+\tau I_p)^{-1}\Sigma(\theta_{\text{pop}}^*+D\theta_1)-\theta_{\text{pop}}^*\right\|_\Sigma^2 \\ &= \left\|(\Sigma+\tau I_p)^{-1}\Sigma D\theta_1-\tau(\Sigma+\tau I_p)^{-1}\theta_{\text{pop}}^*\right\|_\Sigma^2.\end{aligned}$$

Let $M_1=\Sigma^{1/2}$, $M_2=(\Sigma+\tau I_p)^{-1}\Sigma D$ and $a=\tau(\Sigma+\tau I_p)^{-1}\theta_{\text{pop}}^*$. Then, the function $\theta_1 \mapsto \mathcal{R}_{\text{eq},1}^{(1)}\left(\Sigma, \theta_1, \theta_{\text{pop}}^*\right)$ can be expressed as

$$f(\theta_1)=\|M_1(M_2\theta_1-a)\|_2^2,$$

which has gradient

$$\nabla f(\theta_1)=2M_2^\top M_1^\top M_1(M_2\theta_1-a).$$

As $\|\theta_1\|_2 \leq C_1$, f is Lipschitz and its Lipschitz constant is $2\|M_1\|_{\text{op}}^2\|M_2\|_{\text{op}}(\|M_1\|_{\text{op}}C_1+\|a\|_2)$. As $\|M_1\|_{\text{op}}, \|M_2\|_{\text{op}}, C_1, \|a\|_2$ are all upper bounded by constants dependent only on R, M, the Lipschitz constant of f is also upper bounded by a constant dependent only on R, M. Thus, an application of the distributional characterization by (Han & Xu, 2023) (restated as Theorem 4 in (Ildiz et al., 2025)) gives that, for some constant $C_4=C_4(R,M)$,

$$\sup_{\theta_{\text{pop}}^*,\theta_1\in B_p(R')} \Pr\left(\left|\mathcal{R}_{\text{eq},1}^{(1)}\left(\Sigma, \theta_1, \theta_{\text{pop}}^*\right)-\widetilde{\mathcal{R}}_{\text{eq},1}^{(1)}\left(\Sigma, \theta_{\text{pop}}^*, \theta_0^{\text{perfo}}\right)\right| \geq \delta\right) \leq C_4 p e^{-p\delta^4/C_4}, \tag{38}$$

where

$$\widetilde{\mathcal{R}}_{\text{eq},1}^{(1)}\left(\Sigma, \theta_{\text{pop}}^*, \theta_0^{\text{perfo}}\right)=\mathbb{E}_{g^{(1)}}\left[\left\|(\Sigma+\tau I_p)^{-1}\Sigma D X^{(1)}\left(\Sigma, \theta_0^{\text{perfo}}, g^{(1)}\right)-\tau(\Sigma+\tau I_p)^{-1}\theta_{\text{pop}}^*\right\|_\Sigma^2\right]. \tag{39}$$

We recall from Lemma C.1 that $\theta_0^{\text{perfo}}=\theta_{\text{pop}}^*+D\theta_0$, $g^{(1)}\sim\mathcal{N}(0,I_p)$ and $X^{(1)}\left(\Sigma, \theta_0^{\text{perfo}}, g^{(1)}\right)$ is given by (31). By plugging (31) into the RHS of (39) and computing the expectation with respect to $g^{(1)}$, we have

$$\begin{aligned}\widetilde{\mathcal{R}}_{\text{eq},1}^{(1)}\left(\Sigma, \theta_{\text{pop}}^*, \theta_0^{\text{perfo}}\right) &= \mathbb{E}_{g^{(1)}}\Bigg[\Bigg\|(\Sigma+\tau I_p)^{-1}\Sigma D(\Sigma+\tau I_p)^{-1}\Sigma\theta_0^{\text{perfo}}-\tau(\Sigma+\tau I_p)^{-1}\theta_{\text{pop}}^* \\ &\qquad +(\Sigma+\tau I_p)^{-1}\Sigma D(\Sigma+\tau I_p)^{-1}\Sigma^{1/2}\frac{\gamma^{(1)}(\theta_0^{\text{perfo}})g^{(1)}}{\sqrt{p}}\Bigg\|_\Sigma^2\Bigg] \\ &= \left\|(\Sigma+\tau I_p)^{-1}\Sigma D(\Sigma+\tau I_p)^{-1}\Sigma\theta_0^{\text{perfo}}-\tau(\Sigma+\tau I_p)^{-1}\theta_{\text{pop}}^*\right\|_\Sigma^2 \\ &\qquad +\frac{\left(\gamma^{(1)}(\theta_0^{\text{perfo}})\right)^2}{p}\operatorname{Tr}\left[\Sigma\left(\Sigma+\tau I_p\right)^{-2}D\Sigma^3\left(\Sigma+\tau I_p\right)^{-2}D\right],\end{aligned} \tag{40}$$

where in the last step we have used the circulant property of the trace. By using the expression for $\gamma^{(1)}(\theta_0^{\text{perfo}})$ in (34) and recalling that $\theta_0^{\text{perfo}}=\theta_{\text{pop}}^*+D\theta_0$, one readily obtains that the RHS of (40) coincides with the first two lines of the RHS of (36).

Finally, we evaluate the second term in the expression for $\mathcal{R}_{\text{eq}}^{(1)}\left(\Sigma, \theta_1, \theta_{\text{pop}}^*\right)$:

$$\mathcal{R}_{\text{eq},2}^{(1)}\left(\Sigma, \theta_1, \theta_{\text{pop}}^*\right) :=\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]\frac{\sigma^2+\tau^2\left\|\left(\Sigma+\tau I_p\right)^{-1}\left(\theta_{\text{pop}}^*+D\theta_1\right)\right\|_\Sigma^2}{p-\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]}.$$

Let $M_1=\Sigma^{1/2}$, $M_2=(\Sigma+\tau I_p)^{-1}D$ and $a=(\Sigma+\tau I_p)^{-1}\theta_{\text{pop}}^*$. Then, the function $\theta_1\mapsto\left\|\left(\Sigma+\tau I_p\right)^{-1}\left(\theta_{\text{pop}}^*+D\theta_1\right)\right\|_\Sigma^2$ can be expressed as

$$f(\theta_1)=\|M_1(M_2\theta_1+a)\|_2^2,$$

which has gradient

$$\nabla f(\theta_1) = 2M_2^\top M_1^\top M_1(M_2\theta_1 + a).$$

As $\|\theta_1\|_2 \leq C_1$, f is Lipschitz and its Lipschitz constant is $2\|M_1\|_{\rm op}^2\|M_2\|_{\rm op}(\|M_1\|_{\rm op}C_1 + \|a\|_2)$. As $\|M_1\|_{\rm op}, \|M_2\|_{\rm op}, C_1, \|a\|_2$ are all upper bounded by constants dependent only on R, M, the Lipschitz constant of f is also upper bounded by a constant dependent only on R, M. Note that the quantity $|p - \kappa \operatorname{Tr}\left[\Sigma^2(\Sigma+\tau I_p)^{-2}\right]|$ is lower bounded by a constant dependent only on R, M, as a consequence of Proposition 2.1 in (Han & Xu, 2023). Thus, we have that the function $\theta_1 \mapsto \mathcal{R}_{\rm eq,2}^{(1)}\left(\Sigma, \theta_1, \theta_{\rm pop}^*\right)$ is Lipschitz and its Lipschitz constant is $C_5 = C_5(R, M)$. Hence, another application of the distributional characterization by (Han & Xu, 2023) (cf. Theorem 4 in (Ildiz et al., 2025)) gives that, for some constant $C_6 = C_6(R, M)$,

$$\sup_{\theta_{\rm pop}^*, \theta_1 \in B_p(R')} \Pr\left(\left|\mathcal{R}_{\rm eq,2}^{(1)}\left(\Sigma, \theta_1, \theta_{\rm pop}^*\right) - \widetilde{\mathcal{R}}_{\rm eq,2}^{(1)}\left(\Sigma, \theta_{\rm pop}^*, \theta_0^{\rm perfo}\right)\right| \geq \delta\right) \leq C_6 p e^{-p\delta^4/C_6}, \tag{41}$$

where

$$\begin{aligned}\widetilde{\mathcal{R}}_{\rm eq,2}^{(1)}\left(\Sigma, \theta_{\rm pop}^*, \theta_0^{\rm perfo}\right) &= \kappa \operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right] \\ &\cdot \frac{\sigma^2 + \tau^2 \mathbb{E}_{g^{(1)}}\left[\left\|(\Sigma+\tau I_p)^{-1}\left(\theta_{\rm pop}^* + DX^{(1)}\left(\Sigma, \theta_0^{\rm perfo}, g^{(1)}\right)\right)\right\|_\Sigma^2\right]}{p - \kappa \operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]}.\end{aligned} \tag{42}$$

By using (31) and computing the expectation with respect to $g^{(1)}$, we have

$$\begin{aligned}&\mathbb{E}_{g^{(1)}}\left[\left\|(\Sigma+\tau I_p)^{-1}\left(\theta_{\rm pop}^* + DX^{(1)}\left(\Sigma, \theta_0^{\rm perfo}, g^{(1)}\right)\right)\right\|_\Sigma^2\right] \\ &= \mathbb{E}_{g^{(1)}}\Bigg[\Bigg\|(\Sigma+\tau I_p)^{-1}\theta_{\rm pop}^* + (\Sigma+\tau I_p)^{-1}D(\Sigma+\tau I_p)^{-1}\Sigma\theta_0^{\rm perfo} \\ &\qquad + (\Sigma+\tau I_p)^{-1}D(\Sigma+\tau I_p)^{-1}\Sigma^{1/2}\frac{\gamma^{(1)}(\theta_0^{\rm perfo})g^{(1)}}{\sqrt{p}}\Bigg\|_\Sigma^2\Bigg] \\ &= \left\|(\Sigma+\tau I_p)^{-1}\theta_{\rm pop}^* + (\Sigma+\tau I_p)^{-1}D(\Sigma+\tau I_p)^{-1}\Sigma\theta_0^{\rm perfo}\right\|_\Sigma^2 \\ &\qquad + \frac{\left(\gamma^{(1)}(\theta_0^{\rm perfo})\right)^2}{p}\operatorname{Tr}\left[\Sigma(\Sigma+\tau I_p)^{-2}D\Sigma(\Sigma+\tau I_p)^{-2}D\right],\end{aligned} \tag{43}$$

where in the last step we have used the circulant property of the trace. By plugging (43) into (42), using the expression for $\gamma^{(1)}(\theta_0^{\rm perfo})$ in (34) and recalling that $\theta_0^{\rm perfo} = \theta_{\rm pop}^* + D\theta_0$, one readily obtains that the RHS of (42) coincides with the last two lines of the RHS of (36). As $\mathcal{R}_{\rm eq}^{(1)}\left(\Sigma, \theta_1, \theta_{\rm pop}^*\right) = \mathcal{R}_{\rm eq,1}^{(1)}\left(\Sigma, \theta_1, \theta_{\rm pop}^*\right) + \mathcal{R}_{\rm eq,2}^{(1)}\left(\Sigma, \theta_1, \theta_{\rm pop}^*\right)$, the desired result readily follows by combining (37), (38) and (41). □

Concluding the argument. Note that

$$\begin{aligned}\operatorname{Tr}\left[\Sigma(\Sigma+\tau I_p)^{-2}D\Sigma^3(\Sigma+\tau I_p)^{-2}D\right] &= O(\|D\|_{\rm op}^2), \\ \operatorname{Tr}\left[\Sigma(\Sigma+\tau I_p)^{-2}D\Sigma(\Sigma+\tau I_p)^{-2}D\right] &= O(\|D\|_{\rm op}^2).\end{aligned}$$

Furthermore, we have

$$\begin{aligned}&\left\|(\Sigma+\tau I_p)^{-1}\Sigma D(\Sigma+\tau I_p)^{-1}\Sigma(\theta_{\rm pop}^* + D\theta_0) - \tau(\Sigma+\tau I_p)^{-1}\theta_{\rm pop}^*\right\|_\Sigma^2 \\ &= \left\|(\Sigma+\tau I_p)^{-1}\Sigma D(\Sigma+\tau I_p)^{-1}\Sigma\theta_{\rm pop}^* - \tau(\Sigma+\tau I_p)^{-1}\theta_{\rm pop}^*\right\|_\Sigma^2 + O(\|D\|_{\rm op}^2) \\ &= \left\|\tau(\Sigma+\tau I_p)^{-1}\theta_{\rm pop}^*\right\|_\Sigma^2 \\ &\qquad - 2\tau\langle(\Sigma+\tau I_p)^{-1}\theta_{\rm pop}^*, \Sigma(\Sigma+\tau I_p)^{-1}\Sigma D(\Sigma+\tau I_p)^{-1}\Sigma\theta_{\rm pop}^*\rangle + O(\|D\|_{\rm op}^2) \\ &= \tau\langle\theta_{\rm pop}^*, (\Sigma+\tau I_p)^{-1}\left(\tau I_p - 2(\Sigma+\tau I_p)^{-1}\Sigma^2 D\right)\Sigma(\Sigma+\tau I_p)^{-1}\theta_{\rm pop}^*\rangle + O(\|D\|_{\rm op}^2).\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\left\|(\Sigma+\tau I_p)^{-1}\left(\theta^*_{\text{pop}}+D\left(\Sigma+\tau I_p\right)^{-1}\Sigma(\theta^*_{\text{pop}}+D\theta_0)\right)\right\|_\Sigma^2\\
&=\left\|(\Sigma+\tau I_p)^{-1}\left(\theta^*_{\text{pop}}+D\left(\Sigma+\tau I_p\right)^{-1}\Sigma\theta^*_{\text{pop}}\right)\right\|_\Sigma^2+O(\|D\|_{\text{op}}^2)\\
&=\left\|(\Sigma+\tau I_p)^{-1}\theta^*_{\text{pop}}\right\|_\Sigma^2+2\langle(\Sigma+\tau I_p)^{-1}\theta^*_{\text{pop}},\Sigma\left(\Sigma+\tau I_p\right)^{-1}D\left(\Sigma+\tau I_p\right)^{-1}\Sigma\theta^*_{\text{pop}}\rangle+O(\|D\|_{\text{op}}^2)\\
&=\langle\theta^*_{\text{pop}},(\Sigma+\tau I_p)^{-1}\left(I_p+2\left(\Sigma+\tau I_p\right)^{-1}\Sigma D\right)\Sigma\left(\Sigma+\tau I_p\right)^{-1}\theta^*_{\text{pop}}\rangle+O(\|D\|_{\text{op}}^2).
\end{aligned}$$

Recalling the definitions (14) and (36), we conclude that

$$\mathcal{R}_{\text{eq}}^{(2)}\left(\Sigma,\theta_0,\theta^*_{\text{pop}}\right)=\mathcal{R}_{\text{eq}}\left(\Sigma,\theta^*_{\text{pop}},D,\lambda\right)+O(\|D\|_{\text{op}}^2). \tag{44}$$

Thus, the desired result follows from (44) and Lemma C.2.

D. Extension to sub-Gaussian data

Throughout this appendix, we relax Assumption 3.1 as follows.

Assumption D.1 (Regression performative model – relaxed assumption)**.** For $\theta\in\mathbb{R}^p$, samples from $\mathcal{D}(\theta)$ are taken i.i.d. with features x drawn independently of θ and such that $\Sigma^{-1/2}x$ has independent, zero mean, unit variance and uniformly sub-Gaussian entries. The label y is given by

$$y=x^\top\theta^*_{\text{pop}}+x^\top D\theta+w,\quad w\sim\mathcal{N}(0,\sigma^2). \tag{45}$$

We assume $p=2d$, $(\theta^*_{\text{pop}})^\top=(a^\top,0)$ with a having zero mean and covariance I_d/d, and $D=\text{diag}(b,c)$ where $b,c\in\mathbb{R}^d$ with $\|b\|_\infty,\|c\|_\infty<1$. We further assume that $a\sqrt{d}$ has sub-Gaussian norm upper bounded by a universal constant (independent of d).

Theorem D.2 (Excess risk – over-parameterized, relaxed assumptions)**.** *Let Assumption D.1 hold. Let $R>0$ be a constant s.t. $\theta^*_{\text{pop}}\in B_p(R)$ and let θ_0 be sampled uniformly on the unit sphere. Assume that $\kappa,\sigma,\lambda\in(1/M,M)$ and $\|\Sigma\|_{\text{op}},\ \|\Sigma^{-1}\|_{\text{op}}\le M$ for some constant $M>1$. Then, there exists a constant $C=C\left(M,R\right)$ such that for any $\delta\in(0,1/2]$, with probability at least $1-C\delta^{-7}p^{-1/8}$,*

$$\left|\mathcal{R}(\Sigma,\theta_2,\theta^*_{\text{pop}})-\mathcal{R}_{\text{eq}}\left(\Sigma,\theta^*_{\text{pop}},D,\lambda\right)\right|\le\delta+O(\|D\|_{\text{op}}^2), \tag{46}$$

where $\mathcal{R}_{\text{eq}}\left(\Sigma,\theta^*_{\text{pop}},D,\lambda\right)$ is given by (14).

Lemma D.3 (Norm control)**.** *In the setting of Theorem D.2, we have that*

$$\|\theta_1\|_2\le C, \tag{47}$$

$$\|\theta_2\|_2\le C, \tag{48}$$

$$\|\theta_0\|_\infty\le C\frac{\log p}{\sqrt{p}}, \tag{49}$$

$$\|\theta^*_{\text{pop}}\|_\infty\le C\frac{\log p}{\sqrt{p}}, \tag{50}$$

$$\|\theta_1\|_\infty\le C\frac{\log p}{\sqrt{p}}, \tag{51}$$

with probability at least $1-Ce^{-\log^2 p/C}$, where $C=C(R,M)$ is a constant depending only on R,M (and not on n,p).

Proof. We start by proving (47). The claim follows by extending the argument of Proposition 11 in (Ildiz et al., 2025) and we repeat it here for completeness. Recall that

$$\theta_1 = \frac{1}{p}\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-1} X^{0\top}X^0(\theta^*_{\text{pop}} + D\theta_0) + \frac{1}{p}\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-1} X^{0\top}w.$$

Note that

$$\left\|\frac{1}{p}\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-1} X^{0\top}X^0\right\|_{\text{op}} \leq 1,$$

which implies that

$$\left\|\frac{1}{p}\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-1} X^{0\top}X^0(\theta^*_{\text{pop}} + D\theta_0)\right\|_2 \leq C_1,$$

for some constant $C_1 = C_1(R, M)$. Next, we can write

$$\begin{aligned}\left\|\frac{1}{p}\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-1} X^{0\top}w\right\|_2^2 &= \frac{w^\top X^0}{p}\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-2}\frac{X^{0\top}w}{p} \\ &\leq \frac{w^\top w}{p}\left\|\frac{1}{p}X^0\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-2}X^{0\top}\right\|_{\text{op}}.\end{aligned}$$

Using Bernstein's inequality, we have that $w^\top w/p$ is upper bounded by $C_2 = C_2(R, M)$ with probability at least $1 - e^{-p/C_2}$. Furthermore, $\left\|\frac{1}{p}X^0\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-2}X^{0\top}\right\|_{\text{op}} \leq \left\|\frac{1}{p}X^0\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-2}X^{0\top}\right\|_{\text{op}}$ is also upper bounded by a universal constant. Thus, an application of the triangle inequality gives (47). Repeating the same argument with θ_1 in place of θ_0 and X^1 in place of X^0 readily gives (48).

Let $v \in \mathbb{R}^p$ be a vector such that $v\sqrt{p}$ has sub-Gaussian norm upper bounded by a universal constant (independent of p). We will now show that

$$\|v\|_\infty \leq C\frac{\log p}{\sqrt{p}}, \tag{52}$$

with probability at least $1 - e^{-\log^2 p}$. To see this, it suffices to note that the j-th coordinate $v_j\sqrt{p}$ is sub-Gaussian with sub-Gaussian norm upper bounded by a universal constant. Thus,

$$\mathbb{P}(|v_j\sqrt{p}| > t) \leq 2e^{-t^2/C_3},$$

for some universal constant C_3. Taking $t = C\log p$ and doing a union bound over $j \in \{1, \dots, p\}$ gives (52).

Since θ_0 is sampled uniformly on the sphere, (49) is implied by (52). Therefore, $\theta_0\sqrt{p}$ has sub-Gaussian norm upper bounded by a universal constant (independent of p). Furthermore, (50) is implied by (52) since θ^*_{pop} satisfies Assumption D.1. Finally, letting $\|\cdot\|_{\psi_2}$ denote the sub-Gaussian norm of a vector, we have

$$\begin{aligned}\|\theta_1\sqrt{p}\|_{\psi_2} &\leq \left\|\frac{1}{p}\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-1} X^{0\top}X^0\right\|_{\text{op}}(\|\theta^*_{\text{pop}}\sqrt{p}\|_{\psi_2} + \|\theta_0\sqrt{p}\|_{\psi_2}) \\ &\quad + \left\|\frac{1}{\sqrt{p}}\left(\frac{1}{p}X^{0\top}X^0 + \lambda I_p\right)^{-1} X^{0\top}\right\|_{\text{op}}\|w\|_{\psi_2} \\ &\leq \|\theta^*_{\text{pop}}\sqrt{p}\|_{\psi_2} + \|\theta_0\sqrt{p}\|_{\psi_2} + \|w\|_{\psi_2},\end{aligned} \tag{53}$$

which is upper bounded by a universal constant. Thus, (51) is also implied by (52) and the proof is complete. □

Lemma D.4. *Let Assumption D.1 hold. Let $R > 0$ be a constant such that $\theta^*_{\text{pop}} \in B_p(R)$ and let θ_0 be sampled uniformly on the unit sphere. Assume that $\kappa, \sigma, \lambda \in (1/M, M)$ and $\|\Sigma\|_{\text{op}},\ \|\Sigma^{-1}\|_{\text{op}} \leq M$ for some constant $M > 1$. Then, there exists a constant $C = C(M, R)$ such that, for any $\delta \in (0, 1/2]$, the following holds*

$$\sup_{\theta^*_{\text{pop}}, \theta_0 \in B_p(R)} \Pr\left(\left|\mathcal{R}_1(\Sigma, \theta_1, \theta^*_{\text{pop}}) - \mathcal{R}^{(1)}_{\text{eq}}\left(\Sigma, \theta_0, \theta^*_{\text{pop}}\right)\right| \geq \delta\right) \leq Cpe^{-p\delta^4/C}, \tag{54}$$

with probability at least $1 - C\delta^{-7}p^{-1/8}$, where $\mathcal{R}^{(1)}_{\text{eq}}$ is given by (29).

Proof. By Lemma D.3, we have that $\theta^*_{\text{pop}} + D\theta_0$ satisfies the delocalization condition of Proposition 10.3 by (Han & Xu, 2023). This implies that the hypotheses of Theorem 2.4 by (Han & Xu, 2023) are satisfied when we train using $\theta^*_{\text{pop}} + D\theta_0$. Thus, we can now follow the same steps as in Lemma C.1 which invokes Theorem 3 by (Ildiz et al., 2025). In particular, Theorem 3 by (Ildiz et al., 2025) uses Theorem 4 therein plus the bound on $\|\theta_1\|_2$ given by Lemma D.3. Thus, it suffices to replace the application of Theorem 4 by (Ildiz et al., 2025) with the application of Theorem 2.4 by (Han & Xu, 2023), and the desired result readily holds. □

Proof of Theorem D.2. By Lemma D.3, we have that $\theta^*_{\text{pop}} + D\theta_0$ and $\theta^*_{\text{pop}} + D\theta_1$ satisfy the delocalization condition of Proposition 10.3 by (Han & Xu, 2023). This implies that the hypotheses of Theorem 2.4 by (Han & Xu, 2023) are satisfied when we train using either $\theta^*_{\text{pop}} + D\theta_0$ or $\theta^*_{\text{pop}} + D\theta_1$ as vector of regression coefficients. Consequently, the desired result is obtained by following the same steps as in the proof of Theorem 5.1, the only differences being that *(i)* we apply Lemma D.4 in place of Lemma C.1, and *(ii)* we apply Theorem 2.4 by (Han & Xu, 2023) in place of Theorem 4 by (Ildiz et al., 2025). This requires an upper bound on $\|\theta_1\|_2, \|\theta_2\|_2$ which is provided by Lemma D.3. □

E. Proof of Theorem 5.2

We start by computing explicitly $\mathbb{E}_{\theta^*_{\text{pop}}} \mathcal{R}_{\text{eq}}\left(\Sigma, \theta^*_{\text{pop}}, D, \lambda\right)$.

Lemma E.1. *Consider the setting of Theorem 5.1, assume that a has covariance I_d/d, and let $\Sigma = \begin{bmatrix} I_d & \rho I_d \\ \rho I_d & I_d \end{bmatrix}$. Then, we have that*

$$\begin{aligned} \mathbb{E}_{\theta^*_{\text{pop}}} \mathcal{R}_{\text{eq}}\left(\Sigma, \theta^*_{\text{pop}}, D, \lambda\right) &= \widetilde{\mathcal{R}}(D, \lambda, \rho) + O(\bar{b}\rho^2 + \rho^4), \\ \widetilde{\mathcal{R}}(D, \lambda, \rho) &:= \mathcal{R}_0(\lambda, \rho) + \bar{b}A_1(\lambda) + \bar{c}\rho^2 A_2(\lambda), \end{aligned}$$

where $\bar{b} = \text{Tr}[\text{diag}(b)]/d, \bar{c} = \text{Tr}[\text{diag}(c)]/d$ and the auxiliary functions $\mathcal{R}_0(\lambda, \rho)$, $A_1(\lambda)$, and $A_2(\lambda)$ are given by

$$\begin{aligned} \mathcal{R}_0(\lambda, \rho) &= \frac{\tau^2}{(1+\tau)^2} + \frac{\kappa}{(1+\tau)^2 - \kappa}\left(\sigma^2 + \frac{\tau^2}{(1+\tau)^2}\right) \\ &\quad + \rho^2\left(\frac{\tau^2(1-2\tau)}{(1+\tau)^4} + \frac{\kappa\tau^2(1-2\tau)}{(1+\tau)^4\left((1+\tau)^2 - \kappa\right)} + \frac{\kappa\tau(\tau-2)}{\left((1+\tau)^2-\kappa\right)^2}\left(\sigma^2 + \frac{\tau^2}{(1+\tau)^2}\right)\right), \\ A_1(\lambda) &= -\frac{2\tau}{(1+\tau)^3} + \frac{2\kappa\tau^2}{(1+\tau)^3\left((1+\tau)^2 - \kappa\right)}, \\ A_2(\lambda) &= -\frac{4\tau^3}{(1+\tau)^5} + \frac{2\kappa\tau^3(\tau^2-1)}{(1+\tau)^6\left((1+\tau)^2 - \kappa\right)}. \end{aligned} \tag{55}$$

Proof. Given a $p \times p$ matrix M, let us denote by $(M)_1$ its top-left $d \times d$ block. For any $M \in \mathbb{R}^{p\times p}$, we have

$$\begin{aligned} \mathbb{E}_{\theta^*_{\text{pop}}}\left[\langle \theta^*_{\text{pop}}, M\theta^*_{\text{pop}}\rangle\right] &= \mathbb{E}_{\theta^*_{\text{pop}}}\left[(\theta^*_{\text{pop}})^\top M\theta^*_{\text{pop}}\right] = \mathbb{E}_{\theta^*_{\text{pop}}}\left[\text{Tr}\left[(\theta^*_{\text{pop}})^\top M\theta^*_{\text{pop}}\right]\right] \\ &= \mathbb{E}_{\theta^*_{\text{pop}}}\left[\text{Tr}\left[M\theta^*_{\text{pop}}(\theta^*_{\text{pop}})^\top\right]\right] = \text{Tr}\left[(M)_1\right]/d, \end{aligned}$$

where the third equality uses the circulant property of the trace and the last one that $(\theta^*_{\text{pop}})^\top = (a^\top, 0)$ with a having covariance I_d/d. Thus, from (14), we have

$$\begin{aligned}\mathbb{E}_{\theta^*_{\text{pop}}}\mathcal{R}_{\text{eq}}\left(\Sigma, \theta^*_{\text{pop}}, D, \lambda\right) = &\frac{\tau^2}{d}\operatorname{Tr}\left[\left(\Sigma\left(\Sigma+\tau I_p\right)^{-2}\right)_1\right] \\ &+\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]\frac{\sigma^2+\frac{\tau^2}{d}\operatorname{Tr}\left[\left(\Sigma\left(\Sigma+\tau I_p\right)^{-2}\right)_1\right]}{p-\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]} \\ &-\frac{2\tau}{d}\operatorname{Tr}\left[\left(\left(\Sigma+\tau I_p\right)^{-2}\Sigma^2 D\left(\Sigma+\tau I_p\right)^{-1}\Sigma\right)_1\right] \\ &+\frac{2\kappa\tau^2}{d}\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]\frac{\operatorname{Tr}\left[\left(\left(\Sigma+\tau I_p\right)^{-2}\Sigma D\left(\Sigma+\tau I_p\right)^{-1}\Sigma\right)_1\right]}{p-\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]}.\end{aligned} \tag{56}$$

Note that $\Sigma+\tau I_p = \begin{bmatrix}1+\tau & \rho \\ \rho & 1+\tau\end{bmatrix}\otimes I_d$ has inverse $(\Sigma+\tau I_p)^{-1} = \frac{1}{(1+\tau)^2-\rho^2}\begin{bmatrix}1+\tau & -\rho \\ -\rho & 1+\tau\end{bmatrix}\otimes I_d$. Furthermore,

$$\begin{aligned}(\Sigma+\tau I_p)^{-2} &= \frac{1}{\left((1+\tau)^2-\rho^2\right)^2}\begin{bmatrix}(1+\tau)^2+\rho^2 & -2\rho(1+\tau) \\ -2\rho(1+\tau) & (1+\tau)^2+\rho^2\end{bmatrix}\otimes I_d, \\ \Sigma^2 &= \begin{bmatrix}1+\rho^2 & 2\rho \\ 2\rho & 1+\rho^2\end{bmatrix}\otimes I_d.\end{aligned}$$

A direct block multiplication gives

$$\begin{aligned}\operatorname{Tr}\left[\left(\Sigma(\Sigma+\tau I_p)^{-2}\right)_1\right] &= d\,\frac{(1+\tau)^2-\rho^2(1+2\tau)}{\left((1+\tau)^2-\rho^2\right)^2}, \\ \operatorname{Tr}\left[\Sigma^2(\Sigma+\tau I_p)^{-2}\right] &= 2d\,\frac{(1+\tau-\rho^2)^2+\rho^2\tau^2}{\left((1+\tau)^2-\rho^2\right)^2}, \\ \operatorname{Tr}\left[\left((\Sigma+\tau I_p)^{-2}\Sigma^2 D(\Sigma+\tau I_p)^{-1}\Sigma\right)_1\right] &= \frac{(1+\tau-\rho^2)\left((1+\tau-\rho^2)^2+\rho^2\tau^2\right)}{\left((1+\tau)^2-\rho^2\right)^3}\operatorname{Tr}[\operatorname{diag}(b)] \\ &\quad+\frac{2(1+\tau-\rho^2)\rho^2\tau^2}{\left((1+\tau)^2-\rho^2\right)^3}\operatorname{Tr}[\operatorname{diag}(c)], \\ \operatorname{Tr}\left[\left((\Sigma+\tau I_p)^{-2}\Sigma D(\Sigma+\tau I_p)^{-1}\Sigma\right)_1\right] &= \frac{(1+\tau-\rho^2)\left((1+\tau)(1+\tau-\rho^2)-\rho^2\tau\right)}{\left((1+\tau)^2-\rho^2\right)^3}\operatorname{Tr}[\operatorname{diag}(b)] \\ &\quad+\frac{\rho^2\tau\left(\rho^2+\tau^2-1\right)}{\left((1+\tau)^2-\rho^2\right)^3}\operatorname{Tr}[\operatorname{diag}(c)].\end{aligned}$$

Expanding each rational function at $\rho = 0$ using

$$\begin{aligned}\frac{1}{((1+\tau)^2-\rho^2)^2} &= \frac{1}{(1+\tau)^4}\left(1+\frac{2\rho^2}{(1+\tau)^2}\right)+O(\rho^4), \\ \frac{1}{((1+\tau)^2-\rho^2)^3} &= \frac{1}{(1+\tau)^6}\left(1+\frac{3\rho^2}{(1+\tau)^2}\right)+O(\rho^4),\end{aligned}$$

yields, to order ρ^2,

$$\begin{aligned}
\frac{1}{d}\operatorname{Tr}\left[\left(\Sigma(\Sigma+\tau I_p)^{-2}\right)_1\right] &= \left(\frac{1}{(1+\tau)^2}+\rho^2\frac{1-2\tau}{(1+\tau)^4}\right)+O(\rho^4),\\
\frac{1}{d}\operatorname{Tr}\left[\Sigma^2(\Sigma+\tau I_p)^{-2}\right] &= \frac{2}{(1+\tau)^2}+2\rho^2\frac{\tau^2-2\tau}{(1+\tau)^4}+O(\rho^4),\\
\frac{1}{d}\operatorname{Tr}\left[\left((\Sigma+\tau I_p)^{-2}\Sigma^2 D(\Sigma+\tau I_p)^{-1}\Sigma\right)_1\right] &= \frac{\bar b}{(1+\tau)^3}+\rho^2\left(\frac{\tau^2-3\tau}{(1+\tau)^5}\bar b+\frac{2\tau^2}{(1+\tau)^5}\bar c\right)+O(\rho^4),\\
\frac{1}{d}\operatorname{Tr}\left[\left((\Sigma+\tau I_p)^{-2}\Sigma D(\Sigma+\tau I_p)^{-1}\Sigma\right)_1\right] &= \frac{\bar b}{(1+\tau)^3}+\rho^2\left(\frac{1-3\tau}{(1+\tau)^5}\bar b+\frac{\tau(\tau^2-1)}{(1+\tau)^6}\bar c\right)+O(\rho^4).
\end{aligned}$$

Moreover, we have that

$$\frac{\kappa\operatorname{tr}\left[\Sigma^2(\Sigma+\tau I_p)^{-2}\right]}{p-\kappa\operatorname{tr}\left[\Sigma^2(\Sigma+\tau I_p)^{-2}\right]} = \frac{\kappa}{(1+\tau)^2-\kappa}+\rho^2\frac{\kappa\tau(\tau-2)}{\left((1+\tau)^2-\kappa\right)^2}+O(\rho^4).$$

Plugging these into (56) gives the claimed result. □

Let us further define

$$\tau^*(D,\rho) := \arg\min_{\tau\geq 0}\widetilde{\mathcal{R}}(D,\lambda,\rho), \qquad \tau_0^*(\rho) := \arg\min_{\tau\geq 0}\mathcal{R}_0(\lambda,\rho), \qquad \tau_0 := \tau_0^*(0). \tag{57}$$

Then, the following result proves an expression for $\tau^*(D,\rho)$, up to order ρ^2.

Lemma E.2. *In the setting of Theorem 5.2, we have that*

$$\tau^*(D,\rho) = \tau_0^*(\rho)+\bar b\left(B_3(\sigma,\kappa)+O(\rho^2)\right)+\bar c\left(\rho^2 C_3(\sigma,\kappa)+O(\rho^4)\right)+O(\bar b^2+\bar c^2),$$

where

$$\begin{aligned}
\tau_0 &= \frac{1+\kappa+\kappa\sigma^2+\sqrt{(1+\kappa+\kappa\sigma^2)^2-4\kappa}}{2}-1,\\
\tau_0^*(\rho) &= \tau_0-\rho^2\frac{\kappa\tau_0^2}{(1+\tau_0)\left((1+\tau_0)^2-\kappa\right)}+O(\rho^4),\\
B_3(\sigma,\kappa) &= -\frac{2(1+\tau_0)^4-3(\kappa+1)(1+\tau_0)^3+4\kappa(1+\tau_0)^2+\kappa(\kappa+1)(1+\tau_0)-2\kappa^2}{(1+\tau_0)^2\left((1+\tau_0)^2-\kappa\right)},\\
C_3(\sigma,\kappa) &= -\frac{\tau_0^2\left(4\tau_0^4+(6-3\kappa)\tau_0^3-(6+3\kappa)\tau_0^2+(\kappa^2+9\kappa-14)\tau_0-3\kappa^2+9\kappa-6\right)}{(1+\tau_0)^4\left((1+\tau_0)^2-\kappa\right)}.
\end{aligned} \tag{58}$$

Proof. A direct differentiation gives

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\tau}\widetilde{\mathcal{R}}(D,\lambda,\rho) = {}& \frac{2}{1+\tau}\left(\frac{\tau}{(1+\tau)^2-\kappa}-\frac{\kappa\left(\sigma^2(1+\tau)^2+\tau^2\right)}{\left((1+\tau)^2-\kappa\right)^2}\right)\\
&+\rho^2\left(\frac{2\tau\left(\tau^2-4\tau+1\right)}{(1+\tau)^5}+\frac{2\kappa\tau\left((1+\tau)^2\left(3\tau^2-5\tau+1\right)-\kappa\left(\tau^2-4\tau+1\right)\right)}{(1+\tau)^5\left((1+\tau)^2-\kappa\right)^2}\right.\\
&\qquad-\frac{2\kappa^2\left(\kappa\sigma^2(\tau-1)(1+\tau)^3+\kappa\tau^2(\tau^2+\tau-3)\right)}{(1+\tau)^3\left((1+\tau)^2-\kappa\right)^3}\\
&\qquad\left.-\frac{2\kappa\left(\sigma^2(1+\tau)^4(\tau^2-4\tau+1)+\tau^2(1+\tau)^2(\tau^2-5\tau+3)\right)}{(1+\tau)^3\left((1+\tau)^2-\kappa\right)^3}\right)\\
&+\bar b\left(\frac{4\tau-2}{(1+\tau)^4}+\frac{\kappa\tau(4-6\tau)}{(1+\tau)^4\left((1+\tau)^2-\kappa\right)}-\frac{4\kappa^2\tau^2}{(1+\tau)^4\left((1+\tau)^2-\kappa\right)^2}\right)\\
&+\bar c\rho^2\left(\frac{4\tau^2(2\tau-3)}{(1+\tau)^6}+\frac{2\kappa\tau^2(1+\tau)\left(\kappa(\tau^2-6\tau+3)-(1+\tau)^2(3\tau^2-8\tau+3)\right)}{(1+\tau)^7\left((1+\tau)^2-\kappa\right)^2}\right).
\end{aligned}$$

With this explicit derivatives, the stationarity equation $\frac{\mathrm{d}}{\mathrm{d}\tau}\widetilde{\mathcal{R}}(D,\lambda,\rho)=0$ is equivalent to $\frac{F\left(\tau,\bar{b},\bar{c},\rho^2\right)}{(1+\tau)^7((1+\tau)^2-\kappa)^3}=0$, where

$$\begin{aligned}
F\left(\tau,\bar{b},\bar{c},\rho^2\right) &= F_0(\tau)\rho^2F_\rho(\tau)+\bar{b}F_b(\tau)+\bar{c}\rho^2F_{\rho c}(\tau),\\
F_0(\tau) &= 2(1+\tau)^7\left(\kappa-(1+\tau)^2\right)\left(\kappa\sigma^2\tau+\kappa\sigma^2+\kappa\tau-\tau^2-\tau\right),\\
F_b(\tau) &= 2(1+\tau)^4\left((1+\tau)^2-\kappa\right)\left((\tau-1)\kappa^2+(\tau+1)(2-2\tau-3\tau^2)\kappa+(1+\tau)^3(2\tau-1)\right),\\
F_\rho(\tau) &= -2(1+\tau)^5\left(\kappa^2\left(\tau(\tau^2+\tau-1)+\sigma^2(1+\tau)^2(\tau-1)\right)\right.\\
&\qquad +\kappa(1+\tau)^2\left(\tau(\tau^2-6\tau+2)+\sigma^2(1+\tau)(\tau^2-4\tau+1)\right)\\
&\qquad \left.-\tau(1+\tau)^3(\tau^2-4\tau+1)\right),\\
F_{\rho c}(\tau) &= 2\tau^2(1+\tau)^2\left((1+\tau)^2-\kappa\right)\left((\tau-3)\kappa^2-3(\tau+1)(\tau^2-3)\kappa+2(\tau+1)^3(2\tau-3)\right).
\end{aligned}$$

Setting $\bar{b}=\bar{c}=\rho^2=0$ yields

$$(1+\tau)^2-(1+\kappa+\kappa\sigma^2)(1+\tau)+\kappa=0,$$

and the desired minimum corresponds to its largest solution, which is given by τ_0 as expressed in the statement. It is easy to see that

$$\begin{aligned}
\partial_\tau F\left(\tau_0,0,0,0)\right) &= -2(1+\tau_0)^7\left((1+\tau_0)^2-\kappa\right)\left(2(1+\tau_0)-(1+\kappa+\kappa\sigma^2)\right)\\
&= -2(1+\tau_0)^7\left((1+\tau_0)^2-\kappa\right)\sqrt{(1+\kappa+\kappa\sigma^2)^2-4\kappa}\neq 0.
\end{aligned}$$

Therefore, the implicit function theorem gives a smooth map $\tau^*(\bar{b},\bar{c},\rho^2)$ with $\tau^*(0,0,0)=\tau_0$ and $F(\tau^*,\cdot)=0$. Differentiating $F=0$ at $(\tau_0,0,0,0)$ in each small parameter and dividing by $\partial_\tau F(\tau_0,0,0,0)$ yields the linear expansion for $\tau^*-\tau_0$. The coefficient for ρ^2 in $\tau_0^*(\rho)$ is given by

$$\partial_{\rho^2}\tau^*(0,0,0)=-\left.\frac{\partial_{\rho^2}F}{\partial_\tau F}\right|_{(\tau_0,0,0,0)}=-\frac{F_\rho(\tau_0)}{\partial_\tau F_0(\tau_0,0,0,0)}.$$

The $\bar{b}$ coefficient, B_3, is

$$B_3(\sigma,\kappa)=\partial_{\bar{b}}\tau^*(0,0,0)=-\left.\frac{\partial_{\bar{b}}F}{\partial_\tau F}\right|_{(\tau_0,0,0,0)}=-\frac{F_b(\tau_0)}{\partial_\tau F_0(\tau_0,0,0,0)}.$$

The $\bar{c}\rho^2$ coefficient, C_3, is found from the mixed partial derivative:

$$C_3(\sigma,\kappa)=\partial_{\rho^2}\partial_{\bar{c}}\tau^*(0,0,0)=-\left.\frac{\partial_{\rho^2}\partial_{\bar{c}}F}{\partial_\tau F}\right|_{(\tau_0,0,0,0)}=-\frac{F_{\rho c}(\tau_0)}{\partial_\tau F_0(\tau_0,0,0,0)}.$$

Substituting the expressions for $\partial_\tau F_0(\tau_0)$, $F_\rho(\tau_0)$, $F_b(\tau_0)$, and $F_{\rho c}(\tau_0)$ and cancelling common factors gives the coefficients as stated in (58). □

As λ and τ are linked by the fixed point equation (15), an application of Lemma E.2 readily gives that

$$\lambda^*_{\mathrm{eq}}(D,\rho)=\lambda^*_{\mathrm{eq},D=0}(\rho)+\bar{b}(B_1(\sigma,\kappa)+O(\rho^2))+\bar{c}\rho^2(C_1(\sigma,\kappa)+O(\rho^2))+O(\bar{b}^2+\bar{c}^2), \tag{59}$$

where

$$\begin{aligned}
\lambda^*_{\mathrm{eq},D=0}(\rho) &= \tau_0\left(\kappa^{-1}-\frac{1}{1+\tau_0}\right)+\rho^2\left(\tau_0\left(\frac{1}{(1+\tau_0)^2}-\frac{1}{(1+\tau_0)^3}\right)\right.\\
&\qquad \left.-\left(\kappa^{-1}-\frac{1}{(1+\tau_0)^2}\right)\frac{\kappa\tau_0^2}{(1+\tau_0)\left((1+\tau_0)^2-\kappa\right)}\right)+O(\rho^4),\\
B_1(\sigma,\kappa) &= -\frac{2(1+\tau_0)^4-3(\kappa+1)(1+\tau_0)^3+4\kappa(1+\tau_0)^2+\kappa(\kappa+1)(1+\tau_0)-2\kappa^2}{\kappa(1+\tau_0)^4},\\
C_1(\sigma,\kappa) &= -\frac{\tau_0^2\left(4\tau_0^4+(6-3\kappa)\tau_0^3-(6+3\kappa)\tau_0^2+(\kappa^2+9\kappa-14)\tau_0-3\kappa^2+9\kappa-6\right)}{\kappa(1+\tau_0)^6}.
\end{aligned} \tag{60}$$

This proves (19). Next, the corollary below proves (20).

Corollary E.3. *Consider the setting of Theorem 5.2 and let τ_0 be given by (58). Then, we have that*

$$\mathcal{R}^*_{\text{eq}}(D,\rho) = \mathcal{R}^*_{\text{eq}}(\rho) + \bar{b}(B_2(\sigma,\kappa) + O(\rho^2)) + \bar{c}\rho^2(C_2(\sigma,\kappa) + O(\rho^2)) + O(\bar{b}^2 + \bar{c}^2), \tag{61}$$

where

$$\begin{aligned}\mathcal{R}^*_{\text{eq}}(\rho) = \frac{\tau_0^2}{(1+\tau_0)^2} + \frac{\kappa}{(1+\tau_0)^2-\kappa}\left(\sigma^2 + \frac{\tau_0^2}{(1+\tau_0)^2}\right) \\ + \rho^2\left(\frac{\tau_0^2(1-2\tau_0)}{(1+\tau_0)^4} + \frac{\kappa\tau_0^2(1-2\tau_0)}{(1+\tau_0)^4\left((1+\tau_0)^2-\kappa\right)}\right. \\ \left. + \frac{\kappa\tau_0(\tau_0-2)}{\left((1+\tau_0)^2-\kappa\right)^2}\left(\sigma^2 + \frac{\tau_0^2}{(1+\tau_0)^2}\right)\right) + O(\rho^4), \end{aligned} \tag{62}$$

$$B_2(\sigma,\kappa) = -\frac{2\tau_0}{(1+\tau_0)^3} + \frac{2\kappa\tau_0^2}{(1+\tau_0)^3\left((1+\tau_0)^2-\kappa\right)},$$

$$C_2(\sigma,\kappa) = -\frac{4\tau_0^3}{(1+\tau_0)^5} + \frac{2\kappa\tau_0^3(\tau_0^2-1)}{(1+\tau_0)^6\left((1+\tau_0)^2-\kappa\right)}.$$

Proof. Let us re-define $\widetilde{\mathcal{R}}(D,\lambda,\rho)$ given in (55) as $\widetilde{R}(\tau,\bar{b},\bar{c},\rho^2)$ to emphasize its dependence on $\tau,\bar{b},\bar{c}$. By definition of τ_0, we have $\partial_\tau \widetilde{R}(\tau_0,0,0,0) = 0$. Furthermore, from Lemma E.2, we have

$$\tau^*(D,\rho) = \tau_0 + O(\bar{b} + (1+\bar{c})\rho^2).$$

A first–order Taylor expansion of $\widetilde{R}(\tau^*(D,\rho),\bar{b},\bar{c},\rho^2)$ around $(\tau;\bar{b},\bar{c},\rho^2) = (\tau_0;0,0,0)$ gives

$$\begin{aligned}\widetilde{R}(\tau^*(D,\rho),\bar{b},\bar{c},\rho^2) = \widetilde{R}(\tau_0,\bar{b},\bar{c},\rho^2) + \partial_\tau\widetilde{R}(\tau_0,0,0,0)(\tau^*(D,\rho)-\tau_0) \\ + \ O\left((\tau^*(D,\rho)-\tau_0)\bar{b}\right) + O\left((\tau^*(D,\rho)-\tau_0)\rho^2\right) + O(\bar{b}^2+\bar{c}^2+\rho^4).\end{aligned}$$

As $\partial_\tau\widetilde{R}(\tau_0,0,0,0) = 0$, we conclude that

$$\widetilde{R}(\tau^*,\bar{b},\bar{c},\rho^2) = \widetilde{R}(\tau_0,\bar{b},\bar{c},\rho^2) + O(\bar{b}^2+\bar{c}^2+\rho^4),$$

and substituting $\tau = \tau_0$ in (55) gives the claimed expansion. □

We now move to the proof of (**??**), which follows from the lemma below.

Lemma E.4. *Let $B_1(\sigma,\kappa)$ be given by (60). Then, for any $\kappa > 1$, $B_1(\kappa,\cdot)$ has exactly one zero $\sigma_{B_1}(\kappa) > 0$, with*

$$B_1(\kappa,\sigma) \geq 0 \textit{ for } 0 \leq \sigma \leq \sigma_{B_1}(\kappa), \qquad B_1(\kappa,\sigma) \leq 0 \textit{ for } \sigma \geq \sigma_{B_1}(\kappa).$$

Moreover, as $\kappa \to \infty$,

$$\sigma^2_{B_1}(\kappa) = \frac{1}{2} - \frac{7}{18}\kappa^{-1} + O\left(\kappa^{-2}\right).$$

Proof. Let us define the shorthands

$$s(\sigma) := 1+\tau_0, \qquad N_{B_1}(s,\kappa) := 2s^4 - 3(\kappa+1)s^3 + 4\kappa s^2 + \kappa(\kappa+1)s - 2\kappa^2, \tag{63}$$

with τ_0 given by (58). Note that

$$s(\sigma) = \frac{1+\kappa+\kappa\sigma^2 + \sqrt{(1+\kappa+\kappa\sigma^2)^2 - 4\kappa}}{2} \geq \frac{1+\kappa+\sqrt{(1+\kappa)^2-4\kappa}}{2} = \kappa.$$

Now let us also define $\Phi(s) := -N_{B_1}(s,\kappa)/\left(\kappa s^4\right)$ for $s \geq \kappa$. A direct calculation gives the factorization

$$\frac{\mathrm{d}}{\mathrm{d}s}\Phi(s) = \frac{-N'_{B_1}(s)s + 4N_{B_1}(s)}{\kappa s^5} = \frac{-(s^2-\kappa)\left(3(\kappa+1)s - 8\kappa\right)}{\kappa s^5}.$$

For $s \geq \kappa$ we have $s^2 - \kappa > 0$, hence $\Phi'(s)$ changes sign only once at $s_* := \frac{8\kappa}{3(\kappa+1)}$. If $\kappa \geq 5/3$, then $s_* \leq \kappa$ and Φ is strictly decreasing on $[\kappa, \infty)$. If $1 < \kappa < 5/3$, then $\kappa < s_*$ and Φ is increasing on $[\kappa, s_*)$ and strictly decreasing on (s_*, ∞).

Note that $B_1(\kappa, \sigma) = -N_{B_1}(s(\sigma), \kappa)/(\kappa s(\sigma)^4) = \Phi(s(\sigma))$, $s(\sigma)$ is strictly increasing in σ, and $\Phi(s(\sigma)) \to -2/\kappa$ as $\sigma \to \infty$ (since $s(\sigma) \to \kappa\sigma^2$ and $N_{B_1}(s, \kappa) \to 2s^4$). Furthermore, $s(0) = \kappa$ and

$$N_{B_1}(\kappa, \kappa) = -\kappa^2(\kappa - 1)^2 < 0 \implies B_1(\kappa, 0) = -\frac{N_{B_1}(\kappa, \kappa)}{\kappa^5} > 0.$$

Therefore, $B_1(\kappa, \sigma)$ is strictly decreasing on $[0, \infty)$ if $\kappa \geq 5/3$, and for $1 < \kappa < 5/3$ it increases for small σ and then strictly decreases; in either case, since $B_1(\kappa, 0) > 0$ and $\lim_{\sigma\to\infty} B_1(\kappa, \sigma) = -2/\kappa < 0$, it crosses 0 exactly once, which proves the existence and uniqueness of $\sigma_{B_1}(\kappa)$. At the crossing $B_1(\kappa, \sigma_{B_1}) = 0$, hence $N_{B_1}(s(\sigma_{B_1}), \kappa) = 0$.

Now let $\varepsilon := \kappa^{-1}$ and write $s = \kappa c$. Dividing $N_{B_1}(\kappa c, \kappa) = 0$ by κ^4 yields the analytic equation

$$F(\varepsilon, c) = 0, \qquad F(\varepsilon, c) := 2c^4 - 3(1 + \varepsilon)c^3 + 4\varepsilon c^2 + (\varepsilon + \varepsilon^2)c - 2\varepsilon^2.$$

At $\varepsilon = 0$, $F(0, c) = 2c^4 - 3c^3$ has the positive root $c_0 = \frac{3}{2}$, and $\partial_c F(0, c_0) = 8c_0^3 - 9c_0^2 = \frac{27}{4} \neq 0$. By the implicit function theorem there exists a unique analytic branch $c(\varepsilon)$ with $c(0) = \frac{3}{2}$, having the expansion $c(\varepsilon) = \frac{3}{2} + c_1\varepsilon + O(\varepsilon^2)$. Substituting into $F(\varepsilon, c) = 0$ gives that, up to first order,

$$\frac{27}{4}c_1 + \frac{3}{8} = 0 \implies c_1 = -\frac{1}{18}.$$

Thus,

$$\sigma_{B_1}^2 = \frac{(s_c - 1)(s_c - \kappa)}{\kappa s_c} = \left(1 - \frac{1}{\kappa c(\varepsilon)}\right)(c(\varepsilon) - 1),$$

with $s_c = \kappa c(\varepsilon) = \frac{3}{2}\kappa - \frac{1}{18} + O(\kappa^{-1})$. Substituting $c(\varepsilon) = \frac{3}{2} - \frac{1}{18}\varepsilon + O(\varepsilon^2)$ and expanding yields

$$\sigma_{B_1}^2 = \frac{1}{2} - \frac{7}{18}\kappa^{-1} + O(\kappa^{-2}).$$

The analyticity of $c(\varepsilon)$ implies the remainder $O(\varepsilon^2)$ in c and, consequently, the remainder $O(\kappa^{-2})$ in the displayed expansion. □

Next, we move to the proof of (23), which follows from the lemma below.

Lemma E.5. *Let $C_1(\sigma, \kappa)$ be given by (60). Then, for every $\kappa \geq 2$ and all $\sigma \geq 0$, $C_1(\kappa, \sigma) \leq 0$.*

Proof. Let $s(\sigma) = 1 + \tau_0$, with τ_0 given by (58), and note that $\tau_0^2/(\kappa s(\sigma)^6) > 0$. Thus, the sign of C_1 is the opposite of the sign of $N_{C_1}(s(\sigma) - 1, \kappa)$, where

$$N_{C_1}(t, \kappa) = 4t^4 + (6 - 3\kappa)t^3 - (6 + 3\kappa)t^2 + (\kappa^2 + 9\kappa - 14)t - 3\kappa^2 + 9\kappa - 6.$$

At $\sigma = 0$, one has $s(0) - 1 = \kappa - 1$, and a direct substitution yields

$$N_{C_1}(\kappa - 1, \kappa) := \kappa^4 - 3\kappa^3 + 2\kappa^2 = \kappa^2(\kappa - 1)(\kappa - 2).$$

Moreover, differentiating in t gives

$$N''_{C_1}(t, \kappa) = 6(-3\kappa t - \kappa + 8t^2 + 6t - 2) > 0,$$

for all $t \geq \kappa - 1$ and $\kappa \geq 2$, so $N'_{C_1}(\cdot, \kappa)$ is increasing on $[\kappa - 1, \infty)$. As

$$N'_{C_1}(\kappa - 1, \kappa) = \kappa(\kappa - 2)(7\kappa - 3) \geq 0,$$

for $\kappa \geq 2$, it follows that $N_{C_1}(\cdot, \kappa)$ is increasing on $[\kappa - 1, \infty)$. Therefore, for every $\sigma \geq 0$,

$$N_{C_1}(s(\sigma) - 1, \kappa) \;\geq\; N_{C_1}(\kappa - 1, \kappa) = \kappa^2(\kappa - 1)(\kappa - 2) \geq 0 \quad \text{for } \kappa \geq 2.$$

Since the prefactor is positive, $C_1(\kappa, \sigma) \leq 0$ for all σ as soon as $\kappa \geq 2$. □

Lemma E.6. *Let $B_2(\sigma,\kappa)$ be given by (62). Then, for every $\kappa > 1$ and all $\sigma \geq 0$, $B_2(\kappa,\sigma) \leq 0$.*

Proof. We have

$$\begin{aligned} B_2(\sigma,\kappa) &= -\frac{2\tau_0}{(1+\tau_0)^3} + \frac{2\kappa\tau_0^2}{(1+\tau_0)^3\big((1+\tau_0)^2-\kappa\big)} \\ &= \frac{2\tau_0}{(1+\tau_0)^3}\left[-1+\frac{\kappa\tau_0}{(1+\tau_0)^2-\kappa}\right] \\ &= \frac{2\tau_0}{(1+\tau_0)^3}\frac{(1+\tau_0)\big(\kappa-(1+\tau_0)\big)}{(1+\tau_0)^2-\kappa} \\ &= \frac{2\tau_0}{(1+\tau_0)^2}\frac{\kappa-(1+\tau_0)}{(1+\tau_0)^2-\kappa}. \end{aligned}$$

Since $(1+\tau_0)^2 > (1+\tau_0) \geq \kappa$, we have $B_2(\sigma,\kappa) \leq 0$. □

Lemma E.7. *Let $C_2(\sigma,\kappa)$ be given by (62). Then, for every $\kappa > 1$ and all $\sigma \geq 0$, $C_2(\kappa,\sigma) \leq 0$.*

Proof. Let us again write $s(\sigma) = 1+\tau_0 > 0$ and note $\tau_0^2 - 1 = (s(\sigma)-1)^2 - 1 = s(\sigma)^2 - 2s(\sigma)$. Then, substituting and simplifying, we have

$$C_2(\kappa,\sigma) = -\frac{4\tau_0^3}{s(\sigma)^5} + \frac{2\tau_0^3}{s(\sigma)^6}\frac{\kappa(s(\sigma)^2-2s(\sigma))}{s(\sigma)^2-\kappa} = \frac{2\tau_0^3}{s(\sigma)^6}\left(\frac{\kappa s(\sigma)^2 - 2s(\sigma)^3}{s(\sigma)^2-\kappa}\right) = \frac{2\tau_0^3}{s(\sigma)^4}\frac{\kappa-2s(\sigma)}{s(\sigma)^2-\kappa}.$$

Note that $s(\sigma) > \kappa$, and therefore $s(\sigma)^2 - \kappa > 0$ and $\kappa - 2s(\sigma) \leq \kappa - 2\kappa = -\kappa < 0$. Because $\tau_0 > 0$ for $\kappa > 1$, the prefactor $2\tau_0^3/s(\sigma)^4 > 0$. Therefore $C_2(\kappa,\sigma) \leq 0$. □

Combining the results from Lemmas E.4, E.5, E.6 and E.7 concludes the proof of Theorem 5.2.

F. Test risk evaluated on $\mathcal{D}(\theta)$ in the over-parameterized setting

Let $\overline{\mathcal{R}}_k(\Sigma,\theta_k,\theta^*_{\text{pop}})$ be the excess risk of the estimator θ_k given by (12) evaluated on $\mathcal{D}(\theta)$, i.e.,

$$\overline{\mathcal{R}}_k(\Sigma,\theta_k,\theta^*_{\text{pop}}) = \left\|\theta_k - (\theta^*_{\text{pop}} + D\theta_{k-1})\right\|_\Sigma^2.$$

Theorem F.1 (Excess risk on $\mathcal{D}(\theta)$– over-parameterized)**.** *Let $R > 0$ be a constant s.t. $\theta^*_{\text{pop}}, \theta_0 \in B_p(R)$. Assume that $\kappa,\sigma,\lambda \in (1/M, M)$ and $\|\Sigma\|_{\text{op}},\ \|\Sigma^{-1}\|_{\text{op}} \leq M$ for some constant $M > 1$. Then, there exists a constant $C = C(M,R)$ such that for any $\delta \in (0,1/2]$, with probability at least $1 - Cpe^{-p\delta^4/C}$,*

$$\left|\overline{\mathcal{R}}_2(\Sigma,\theta_2,\theta^*_{\text{pop}}) - \overline{\mathcal{R}}_{\text{eq}}\left(\Sigma,\theta^*_{\text{pop}},D,\lambda\right)\right| \leq \delta + O(\|D\|_{\text{op}}^2), \tag{64}$$

where

$$\begin{aligned} &\overline{\mathcal{R}}_{\text{eq}}\left(\Sigma,\theta^*_{\text{pop}},D,\lambda\right) \\ &= \frac{\sigma^2\kappa\,\text{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right] + p\tau^2\langle\theta^*_{\text{pop}},(\Sigma+\tau I_p)^{-1}\Big(I_p + 2\left(\Sigma+\tau I_p\right)^{-1}\Sigma D\Big)\Sigma\left(\Sigma+\tau I_p\right)^{-1}\theta^*_{\text{pop}}\rangle}{p - \kappa\,\text{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]}, \end{aligned} \tag{65}$$

and τ is the unique solution of (15).

Proof. The argument is analogous to that used to prove Theorem 5.1, and we only report differences. Using the same approach as Lemma C.1, we have

$$\sup_{\theta^*_{\text{pop}},\theta_0\in B_p(R)} \Pr\left(\left|\overline{\mathcal{R}}_1(\Sigma,\theta_1,\theta^*_{\text{pop}}) - \overline{\mathcal{R}}^{(1)}_{\text{eq}}\left(\Sigma,\theta_0,\theta^*_{\text{pop}}\right)\right| \geq \delta\right) \leq Cpe^{-p\delta^4/C}, \tag{66}$$

where

$$\overline{\mathcal{R}}_{\text{eq}}^{(1)}\left(\Sigma,\theta_0,\theta_{\text{pop}}^*\right)=\frac{\sigma^2\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]+p\tau^2\left\|\left(\Sigma+\tau I_p\right)^{-1}\left(\theta_{\text{pop}}^*+D\theta_0\right)\right\|_\Sigma^2}{p-\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]}. \tag{67}$$

Next, using the same approach[3] as Lemma C.2, we have

$$\sup_{\theta_{\text{pop}}^*,\theta_0\in B_p(R)}\Pr\left(\left|\overline{\mathcal{R}}_2(\Sigma,\theta_2,\theta_{\text{pop}}^*)-\overline{\mathcal{R}}_{\text{eq}}^{(2)}\left(\Sigma,\theta_0,\theta_{\text{pop}}^*\right)\right|\geq\delta\right)\leq Cpe^{-p\delta^4/C}, \tag{68}$$

where

$$\begin{aligned}&\overline{\mathcal{R}}_{\text{eq}}^{(2)}\left(\Sigma,\theta_0,\theta_{\text{pop}}^*\right)\\&=\frac{\sigma^2\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]+p\tau^2\left\|\left(\Sigma+\tau I_p\right)^{-1}\left(\theta_{\text{pop}}^*+D\left(\Sigma+\tau I_p\right)^{-1}\Sigma(\theta_{\text{pop}}^*+D\theta_0)\right)\right\|_\Sigma^2}{p-\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]}\\&+p\kappa\tau^2\operatorname{Tr}\left[\Sigma\left(\Sigma+\tau I_p\right)^{-2}D\Sigma\left(\Sigma+\tau I_p\right)^{-2}D\right]\cdot\frac{\sigma^2+\tau^2\left\|\left(\Sigma+\tau I_p\right)^{-1}\left(\theta_{\text{pop}}^*+D\theta_0\right)\right\|_\Sigma^2}{\left(p-\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]\right)^2}.\end{aligned} \tag{69}$$

Noting that the quantity in the second line is $O(\|D\|_{\text{op}}^2)$ and that

$$\begin{aligned}&\left\|\left(\Sigma+\tau I_p\right)^{-1}\left(\theta_{\text{pop}}^*+D\left(\Sigma+\tau I_p\right)^{-1}\Sigma(\theta_{\text{pop}}^*+D\theta_0)\right)\right\|_\Sigma^2\\&=\langle\theta_{\text{pop}}^*,\left(\Sigma+\tau I_p\right)^{-1}\left(I_p+2\left(\Sigma+\tau I_p\right)^{-1}\Sigma D\right)\Sigma\left(\Sigma+\tau I_p\right)^{-1}\theta_{\text{pop}}^*\rangle+O(\|D\|_{\text{op}}^2)\end{aligned} \tag{70}$$

concludes the argument. □

Lemma F.2. *Consider the setting of Theorem F.1, assume that a has covariance I_d/d, and let $\Sigma=\begin{bmatrix}I_d & \rho I_d\\ \rho I_d & I_d\end{bmatrix}$. Then, we have that*

$$\begin{aligned}\mathbb{E}_{\theta_{\text{pop}}^*}\overline{\mathcal{R}}_{\text{eq}}\left(\Sigma,\theta_{\text{pop}}^*,D,\lambda\right)&=\overline{\mathcal{R}}(D,\lambda,\rho)+O(\bar{b}\rho^2+\rho^4),\\\overline{\mathcal{R}}(D,\lambda,\rho)&:=\mathcal{R}_0(\lambda,\rho)+\bar{b}\overline{A}_1(\lambda)+\bar{c}\rho^2\overline{A}_2(\lambda),\end{aligned}$$

where $\bar{b}=\operatorname{Tr}[\operatorname{diag}(b)]/d,\bar{c}=\operatorname{Tr}[\operatorname{diag}(c)]/d$, the auxiliary function $\mathcal{R}_0(\lambda,\rho)$ is given by (55), and the new auxiliary functions $\overline{A}_1(\lambda)$ and $\overline{A}_2(\lambda)$ are given by

$$\begin{aligned}\overline{A}_1(\lambda)&=\frac{2\tau^2}{(1+\tau)((1+\tau)^2-\kappa)},\\\overline{A}_2(\lambda)&=\frac{2\tau^3(\tau^2-1)}{(1+\tau)^4((1+\tau)^2-\kappa)}.\end{aligned} \tag{71}$$

[3] In fact, the derivation is simpler since the term corresponding to $\mathcal{R}_{\text{eq},1}^{(1)}\left(\Sigma,\theta_1,\theta_{\text{pop}}^*\right)$ here is absent.

Proof. Using $\mathbb{E}_{\theta^*_{\text{pop}}}\left[\langle \theta^*_{\text{pop}}, M\theta^*_{\text{pop}}\rangle\right] = \operatorname{Tr}\left[(M)_1\right]/d$ from the proof of Lemma E.1, we take the expectation of (65):

$$\begin{aligned}
&\mathbb{E}_{\theta^*_{\text{pop}}}\overline{\mathcal{R}}_{\text{eq}}\left(\Sigma, \theta^*_{\text{pop}}, D, \lambda\right) \\
&= \frac{\sigma^2\kappa \operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]}{p-\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]} \\
&\quad + \frac{p\tau^2}{p-\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]}\frac{1}{d}\operatorname{Tr}\left[\left(\Sigma\left(\Sigma+\tau I_p\right)^{-2}\right)_1\right] \\
&\quad + \frac{p\tau^2}{p-\kappa\operatorname{Tr}\left[\Sigma^2\left(\Sigma+\tau I_p\right)^{-2}\right]}\frac{2}{d}\operatorname{Tr}\left[\left(\left(\Sigma+\tau I_p\right)^{-2}\Sigma D\left(\Sigma+\tau I_p\right)^{-1}\Sigma\right)_1\right].
\end{aligned} \tag{72}$$

The first two terms correspond to the risk with $D = 0$ (i.e., $\bar{b} = \bar{c} = 0$). By the same computations as in the proof of Lemma E.1, these terms combine to $\mathcal{R}_0(\lambda, \rho) + O(\rho^4)$. The third term, which depends on D, requires approximations for its two factors. The first factor is new:

$$\begin{aligned}
\frac{p}{p-\kappa\operatorname{Tr}\left[\Sigma^2(\Sigma+\tau I_p)^{-2}\right]} &= 1 + \frac{\kappa\operatorname{tr}\left[\Sigma^2(\Sigma+\tau I_p)^{-2}\right]}{p-\kappa\operatorname{tr}\left[\Sigma^2(\Sigma+\tau I_p)^{-2}\right]} \\
&= 1 + \left(\frac{\kappa}{(1+\tau)^2-\kappa} + O(\rho^2)\right) = \frac{(1+\tau)^2}{(1+\tau)^2-\kappa} + O(\rho^2).
\end{aligned}$$

For the second factor, we use the trace expansion from Lemma E.1:

$$\frac{1}{d}\operatorname{Tr}\left[\left((\Sigma+\tau I_p)^{-2}\Sigma D(\Sigma+\tau I_p)^{-1}\Sigma\right)_1\right] = \frac{\bar{b}}{(1+\tau)^3} + \rho^2\left(\frac{1-3\tau}{(1+\tau)^5}\bar{b} + \frac{\tau(\tau^2-1)}{(1+\tau)^6}\bar{c}\right) + O(\rho^4).$$

We multiply these two factors by $2\tau^2$ (from (72)) and keep only the terms of order $O(\bar{b})$ and $O(\bar{c}\rho^2)$:

$$\begin{aligned}
\bar{b}\overline{A}_1(\lambda) &= \left(\frac{(1+\tau)^2}{(1+\tau)^2-\kappa}\right)\left(2\tau^2\frac{\bar{b}}{(1+\tau)^3}\right) = \bar{b}\frac{2\tau^2}{(1+\tau)((1+\tau)^2-\kappa)}, \\
\bar{c}\rho^2\overline{A}_2(\lambda) &= \left(\frac{(1+\tau)^2}{(1+\tau)^2-\kappa}\right)\left(2\tau^2\rho^2\bar{c}\frac{\tau(\tau^2-1)}{(1+\tau)^6}\right) = \bar{c}\rho^2\frac{2\tau^3(\tau^2-1)}{(1+\tau)^4((1+\tau)^2-\kappa)}.
\end{aligned}$$

Adding these terms to $\mathcal{R}_0(\lambda, \rho)$ yields the claimed expansion. □

Lemma F.3. *In the setting of Lemma F.2, we have that*

$$\tau^*(D, \rho) = \tau_0^*(\rho) + \bar{b}\left(\overline{B}_3(\sigma, \kappa) + O(\rho^2)\right) + \bar{c}\left(\rho^2\overline{C}_3(\sigma, \kappa) + O(\rho^4)\right) + O(\bar{b}^2 + \bar{c}^2),$$

where τ_0 and $\tau_0^*(\rho)$ are given by (58), and

$$\begin{aligned}
\overline{B}_3(\sigma, \kappa) &= -\frac{\tau_0\left((1+\tau_0)^2(2-\tau_0) - \kappa(\tau_0+2)\right)}{(1+\tau_0)\left((1+\tau_0)^2-\kappa\right)}, \\
\overline{C}_3(\sigma, \kappa) &= -\frac{\tau_0^2\left((4\tau_0-3)(1+\tau_0)((1+\tau_0)^2-\kappa) - \tau_0(\tau_0-1)(5(1+\tau_0)^2-3\kappa)\right)}{(1+\tau_0)^3\left((1+\tau_0)^2-\kappa\right)}.
\end{aligned} \tag{73}$$

Proof. A direct differentiation of $\overline{\mathcal{R}}(D, \lambda, \rho)$ from Lemma F.2 gives

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\overline{\mathcal{R}}(D, \lambda, \rho) = \frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{R}_0(\lambda, \rho) + \bar{b}\frac{\mathrm{d}}{\mathrm{d}\tau}\overline{A}_1(\lambda) + \bar{c}\rho^2\frac{\mathrm{d}}{\mathrm{d}\tau}\overline{A}_2(\lambda).$$

The first term is identical to that in the proof of Lemma E.2. The new derivatives are:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\overline{A}_1(\lambda) = \frac{2\tau(1+\tau)^2(2-\tau) - 2\kappa\tau(2+\tau)}{(1+\tau)^2((1+\tau)^2-\kappa)^2},$$
$$\frac{\mathrm{d}}{\mathrm{d}\tau}\overline{A}_2(\lambda) = \frac{2\tau^2\left((4\tau-3)(1+\tau)((1+\tau)^2-\kappa) - \tau(\tau-1)(5(1+\tau)^2-3\kappa)\right)}{(1+\tau)^4((1+\tau)^2-\kappa)^2}.$$

With these explicit derivatives, the stationarity equation $\frac{\mathrm{d}}{\mathrm{d}\tau}\overline{\mathcal{R}}(D,\lambda,\rho) = 0$ is equivalent to $\frac{\overline{F}\left(\tau,\bar{b},\bar{c},\rho^2\right)}{(1+\tau)^7((1+\tau)^2-\kappa)^3} = 0$, where

$$\overline{F}\left(\tau,\bar{b},\bar{c},\rho^2\right) = F_0(\tau) + \rho^2 F_\rho(\tau) + \bar{b}\overline{F}_b(\tau) + \bar{c}\rho^2\overline{F}_{\rho c}(\tau).$$

The functions $F_0(\tau)$ and $F_\rho(\tau)$ are identical to those defined in the proof of Lemma E.2. The new functions are

$$\overline{F}_b(\tau) = 2\tau(1+\tau)^5((1+\tau)^2-\kappa)\left((1+\tau)^2(2-\tau) - \kappa(2+\tau)\right),$$

$$\overline{F}_{\rho c}(\tau) = 2\tau^2(1+\tau)^3((1+\tau)^2-\kappa)\left((4\tau-3)(1+\tau)((1+\tau)^2-\kappa) - \tau(\tau-1)(5(1+\tau)^2-3\kappa)\right).$$

Setting $\bar{b} = \bar{c} = \rho^2 = 0$ yields the same equation for τ_0 as in Lemma E.2. The partial derivative $\partial_\tau\overline{F}\left(\tau_0,0,0,0\right)$ is also unchanged:

$$\partial_\tau\overline{F}\left(\tau_0,0,0,0)\right) = -2(1+\tau_0)^7\left((1+\tau_0)^2-\kappa\right)\sqrt{(1+\kappa+\kappa\sigma^2)^2-4\kappa} \neq 0.$$

Therefore, the implicit function theorem gives a smooth map $\tau^*(\bar{b},\bar{c},\rho^2)$ with $\tau^*(0,0,0) = \tau_0$ and $\overline{F}(\tau^*,\cdot) = 0$. Differentiating $\overline{F} = 0$ at $(\tau_0,0,0,0)$ and dividing by $\partial_\tau\overline{F}(\tau_0,0,0,0)$ yields

$$\overline{B}_3(\sigma,\kappa) = -\frac{\overline{F_b}(\tau_0)}{\partial_\tau F(\tau_0,0,0,0)}, \qquad \overline{C}_3(\sigma,\kappa) = -\frac{\overline{F}_{\rho c}(\tau_0)}{\partial_\tau F(\tau_0,0,0,0)}.$$

Substituting the expressions for $\overline{F_b}(\tau_0)$, $\overline{F}_{\rho c}(\tau_0)$, and $\partial_\tau F(\tau_0,0,0,0)$ and cancelling common factors gives the coefficients as stated in (73). □

As λ and τ are linked by the fixed point equation (15), an application of Lemma F.3 readily gives that

$$\lambda^*_{\mathrm{eq}}(D,\rho) = \lambda^*_{\mathrm{eq},D=0}(\rho) + \bar{b}(\overline{B}_1(\sigma,\kappa) + O(\rho^2)) + \bar{c}\rho^2(\overline{C}_1(\sigma,\kappa) + O(\rho^2)) + O(\bar{b}^2+\bar{c}^2), \tag{74}$$

where $\lambda^*_{\mathrm{eq},D=0}(\rho)$ is unchanged from (60), and the new coefficients are

$$\begin{aligned}\overline{B}_1(\sigma,\kappa) &= -\frac{\tau_0\left((1+\tau_0)^2(2-\tau_0) - \kappa(\tau_0+2)\right)}{\kappa(1+\tau_0)^3},\\ \overline{C}_1(\sigma,\kappa) &= -\frac{\tau_0^2\left((4\tau_0-3)(1+\tau_0)((1+\tau_0)^2-\kappa) - \tau_0(\tau_0-1)(5(1+\tau_0)^2-3\kappa)\right)}{\kappa(1+\tau_0)^5}.\end{aligned} \tag{75}$$

Next, the corollary below provides the expansion for the optimal equilibrium risk.

Corollary F.4. *Consider the setting of Lemma F.2 and let τ_0 be given by (58). Then, we have that*

$$\overline{\mathcal{R}^*_{\mathrm{eq}}}(D,\rho) = \mathcal{R}^*_{\mathrm{eq}}(\rho) + \bar{b}(\overline{B}_2(\sigma,\kappa) + O(\rho^2)) + \bar{c}\rho^2(\overline{C}_2(\sigma,\kappa) + O(\rho^2)) + O(\bar{b}^2+\bar{c}^2), \tag{76}$$

where $\mathcal{R}^*_{\mathrm{eq}}(\rho)$ is given by (62), and

$$\begin{aligned}\overline{B}_2(\sigma,\kappa) &= \frac{2\tau_0^2}{(1+\tau_0)((1+\tau_0)^2-\kappa)},\\ \overline{C}_2(\sigma,\kappa) &= \frac{2\tau_0^3(\tau_0^2-1)}{(1+\tau_0)^4((1+\tau_0)^2-\kappa)}.\end{aligned} \tag{77}$$

Lemma F.5. *Let $\overline{B}_1(\sigma,\kappa)$ be given by (75). Then, for any $\kappa \geq 2$ and all $\sigma \geq 0$, $\overline{B}_1(\sigma,\kappa) \geq 0$.*

Proof. Let $s(\sigma) = 1 + \tau_0$. By the definition (75), we can write

$$\overline{B}_1(\sigma,\kappa) = -\frac{\tau_0 N_{\overline{B}_1}(s(\sigma),\kappa)}{\kappa s(\sigma)^3}, \qquad N_{\overline{B}_1}(s,\kappa) := s^2(3-s) - \kappa(s+1).$$

From the construction we have $s(\sigma) \geq \kappa$ and here we assume $\kappa \geq 2$, so in particular $s(\sigma) \geq 2$. Since $\tau_0 > 0$, $\kappa > 0$ and $s(\sigma) > 0$, the prefactor $-\frac{\tau_0}{\kappa s(\sigma)^3} < 0$. Therefore, to prove $\overline{B}_1(\sigma,\kappa) > 0$ it suffices to show

$$N_{\overline{B}_1}(s,\kappa) < 0 \qquad \text{for all } s \geq \kappa \geq 2.$$

We analyse $N_{\overline{B}_1}$ as a function of s (with κ fixed). Its derivatives are

$$N'_{\overline{B}_1}(s,\kappa) = 6s - 3s^2 - \kappa, \qquad N''_{\overline{B}_1}(s,\kappa) = 6 - 6s.$$

For $s \geq 2$ we have $N''_{\overline{B}_1}(s,\kappa) < 0$, so $N_{\overline{B}_1}$ is concave on $[2,\infty)$, and hence on $[\kappa,\infty)$ since $\kappa \geq 2$. We first evaluate $N_{\overline{B}_1}$ and its derivative at the boundary point $s = \kappa$:

$$N_{\overline{B}_1}(\kappa,\kappa) = \kappa^2(3-\kappa) - \kappa(\kappa+1) = -\kappa(\kappa-1)^2 < 0,$$

and

$$N'_{\overline{B}_1}(\kappa,\kappa) = 6\kappa - 3\kappa^2 - \kappa = \kappa(5-3\kappa).$$

For $\kappa \geq 2$ we have $5 - 3\kappa < 0$, so $N'_{\overline{B}_1}(\kappa,\kappa) \leq 0$. Since $N_{\overline{B}_1}$ is concave on $[\kappa,\infty)$, its derivative $N'_{\overline{B}_1}(s,\kappa)$ is non-increasing in s on this interval. Hence, for all $s \geq \kappa$, $N'_{\overline{B}_1}(s,\kappa) \leq N'_{\overline{B}_1}(\kappa,\kappa) \leq 0$, so $N_{\overline{B}_1}(\cdot,\kappa)$ is non-increasing on $[\kappa,\infty)$. Together with $N_{\overline{B}_1}(\kappa,\kappa) < 0$ this implies

$$N_{\overline{B}_1}(s,\kappa) \leq N_{\overline{B}_1}(\kappa,\kappa) \leq 0 \qquad \text{for all } s \geq \kappa \geq 2,$$

which proves the lemma. □

Lemma F.6. *Let $\overline{B}_2(\sigma,\kappa)$ be given by (77). Then, for every $\kappa > 1$ and all $\sigma \geq 0$, $\overline{B}_2(\kappa,\sigma) \geq 0$.*

Proof. Recall the definition

$$\overline{B}_2(\sigma,\kappa) = \frac{2\tau_0^2}{(1+\tau_0)((1+\tau_0)^2-\kappa)}.$$

Both the numerator and the denominator are positive for $\kappa > 1$ and $\sigma > 0$ (as shown in the previous lemmas), therefore the claim holds. □

Lemma F.7. *Let $\overline{C}_2(\sigma,\kappa)$ be given by (77). Then, for every $\kappa \geq 2$ and all $\sigma \geq 0$, $\overline{C}_2(\kappa,\sigma) \geq 0$.*

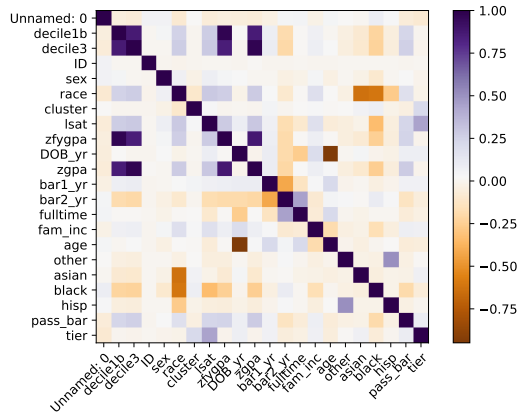
Proof. Recall the definition

$$\overline{C}_2(\sigma,\kappa) = \frac{2\tau_0^3(\tau_0^2-1)}{(1+\tau_0)^4((1+\tau_0)^2-\kappa)}.$$

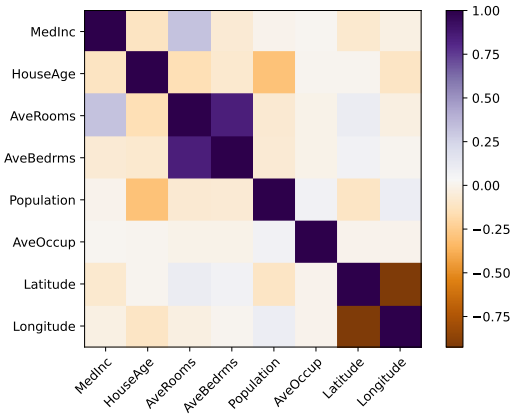
Let $s(\sigma) = 1 + \tau_0$. The denominator is strictly positive for $\kappa > 1$. The term $2\tau_0^3$ is also strictly positive. Thus, the sign of $\overline{C}_2$ is determined by the sign of $(\tau_0^2 - 1)$. We rewrite this term as:

$$\tau_0^2 - 1 = (s(\sigma)-1)^2 - 1 = s(\sigma)^2 - 2s(\sigma) = s(\sigma)(s(\sigma)-2).$$

Since $s(\sigma) \geq \kappa > 1$, $s(\sigma)$ is positive. The sign is therefore determined by $(s(\sigma) - 2)$. We are given $\kappa \geq 2$, which gets $s(\sigma) \geq \kappa \geq 2$. Therefore, $s(\sigma) - 2 \geq 0$ for all $\sigma \geq 0$ and the claim holds. □



(a) Empirical covariance of LSAC dataset.



(b) Empirical covariance of Housing dataset.

G. Details for the experimental setup

For both datasets, the curves are obtained by running 100 equally spaced values of λ with the same splits, so that the observations focus on the performative effect. Data is split uniformly at random across the different steps.

Housing. We keep all features of the dataset and normalize them. We center the target feature since we use a linear regression without intercept. Following Cyffers et al. (2024), we fix the features affected by the performative effect to be `MedInc`, `AveBedrms`, and `AveOccup`, with all values of b set equal.

LSAC. We keep only one feature in cases of redundant encoding, drop features that are too strongly correlated with the target `GPA` ($\rho > 0.6$), and randomly select roughly half of the features to be affected by the performative effect. The names of these features are reported in Table 1. We normalize all features and center the target. All coefficients of b are equal.

Empirical Covariance. In Figures 9a and 9b, we observe that the features do not follow the assumptions made on the data matrix X in the theoretical part, despite exhibiting similar behavior in the experiments. This illustrates that our findings on how to scale regularization remain useful for more general datasets.

Table 1. Features of the LSAC dataset

Category	Feature name
Redundant	`male` (same as `sex`), `parttime` (same as `fulltime`), `decile1` (same as `decile1b`)
With $\rho > 0.6$	`ugpa`, `index6040`, `dnn bar pass prediction`
With $b_{\text{feat}} = \bar{b}$	`Unnamed0`, `decile1b`, `decile3`, `other`, `asian`, `black`, `hisp`, `pass bar`, `tier`