Source-Optimal Training is Transfer-Suboptimal

Anonymous authors
Paper under double-blind review

Abstract

We prove that training a source model optimally for its own task is generically suboptimal when the objective is downstream transfer. We study the source-side optimization problem in L2-SP ridge regression and show a fundamental mismatch between the source-optimal and transfer-optimal source regularization: outside of a measure-zero set, $\tau_0^* \neq \tau_S^*$. We characterize the transfer-optimal source penalty τ_0^* as a function of task alignment and identify an alignment-dependent reversal: with imperfect alignment ($0 < \rho < 1$), transfer benefits from stronger source regularization, while in super-aligned regimes ($\rho > 1$), transfer benefits from weaker regularization. Additionally, in isotropic settings, the decision of whether transfer helps is independent of the target sample size and noise, depending only on task alignment and source characteristics. We verify the linear predictions in a synthetic ridge regression experiment, and we present experiments on MNIST, CIFAR-10, and 20 Newsgroups as evidence that the source-optimal versus transfer-optimal mismatch persists in standard nonlinear transfer learning pipelines.

1 Introduction

1.1 Background and Motivation

A fundamental question in transfer learning is source-side optimization: how should we train the source model if the objective is downstream transfer? We prove a source-optimal versus transfer-optimal mismatch: the source regularization that is optimal for the source task is generically suboptimal for transfer, and outside of a measure-zero set we have $\tau_0^* \neq \tau_S^*$. This mismatch is alignment-dependent: in the standard imperfect-alignment regime ($0 < \rho < 1$), transfer benefits from stronger source regularization than source-optimal training, while in a super-aligned regime ($\rho > 1$) transfer benefits from weaker regularization.

Intuitively, this divergence arises from a geometric mismatch between the source signal and the useful signal for transfer. The source estimator $\hat{\beta}_0$ naturally targets the source vector w_0. However, the optimal initialization for the target task is the projection of the target signal w_1 onto the source, which corresponds to ρw_0. In the imperfect-alignment regime ($0 < \rho < 1$), the source estimator is systematically "too large" relative to the target projection; stronger regularization is required to shrink the estimator towards the optimal scale ρ. Conversely, in the super-aligned regime ($\rho > 1$), the source estimator is "too small" (under-scaled); weaker regularization is preferred to preserve more signal magnitude, even at the cost of higher noise-induced variance. This scaling requirement exists independently of observation noise, though noise further modulates the optimal set-point. Our claims are exact for linear ridge regression; our contribution is isolating a source-side mechanism and making it analytically sharp.

Prior theoretical analyses of transfer learning typically fix the source model and optimize the target-side procedure (or characterize when transfer outperforms training from scratch). In contrast, we study the source-side optimization problem: how should the source model itself be trained to maximize downstream transfer performance? To our knowledge, no prior work derives transfer-optimal source regularization or characterizes its divergence from source-task-optimal training in the general (non-orthonormal) two-task setting we consider.

We begin our exploration through the lens of evaluating when we would expect positive or negative transfer from some source task to a target task. Previous work Dar & Baraniuk (2022); Lampinen & Ganguli (2018);

Yang et al. (2025) has shown that the phenomenon of negative transfer is a concern, where the model generated by transfer learning performs worse than if one were to train the model from scratch. Intuition and previous theoretical work Lampinen & Ganguli (2018) suggest that if the tasks and input datasets are similar enough, transfer learning techniques should help, but if the target task is sufficiently different, or in the context of continual learning the regime has shifted enough, transfer learning may harm the final model, introducing bias and noise from the initialization.

There has been notable work examining this phase boundary between when one should expect positive or negative transfer. In Dar & Baraniuk (2022); Dar et al. (2024), Dar and Baraniuk provided an explicit quantitative phase boundary for freezing/parameter-sharing style transfer in linear models, giving if-and-only-if conditions for when shared representations outperform training from scratch. In Dhifallah & Lu (2021), Dhifallah and Lu explored the phenomenon in the context of feature freezing. Additionally, feature-overlap driven transitions in linear transfer were explored by Tahir, Ganguli, and Rotskoff in Tahir et al. (2024). Related phase transitions have been characterized for other linear transfer mechanisms such as hard parameter sharing Yang et al. (2025). Complementary asymptotic analyses for fine-tuning from pretrained anchors via gradient descent appear in Ghane, Akhtiamov, and Hassibi Ghane et al. (2024), who provide universal asymptotics comparing pretrained versus fine-tuned models under distributional shifts.

The L2-SP (L2-distance to Starting Point) approach was introduced by Li, Grandvalet, and Davoine Xuhong et al. (2018), who showed empirical improvements by penalizing deviation from pre-trained parameters. This approach matches many practical fine-tuning protocols and corresponds to the isotropic elastic weight consolidation (EWC) penalty in Kirkpatrick et al. (2017). Dar, LeJeune, and Baraniuk Dar et al. (2024) recently analyzed when L2-SP transfer learning outperforms standard ridge regression, assuming task parameters are related by orthonormal transformations and focusing on optimal tuning of the target-task regularization given a fixed source model.

Our work fundamentally differs from these prior analyses by solving the source-side optimization problem. We ask: how should the source model itself be trained to maximize downstream transfer performance? This shift in perspective reveals a new phenomenon: the transfer-optimal source regularization τ_0^* generically differs from the source-task-optimal choice. We additionally emperically verify that this phenomenon persists across a range of non-linear networks and tasks, providing evidence that this is not an artifact of the linear setting, but instead may be a general phenomenon related to transfer learning pipelines.

1.2 Setup and Overview of Results

In this paper we formalize transfer learning from the lens of L2-SP ridge regression. In particular, we have two tasks, Task 0 (the source task) and Task 1 (the target task), as well as corresponding training datasets (X_0, y_0) and (X_1, y_1). We then seek to evaluate the expected out-of-sample risk on Task 1, comparing the ridge/ridgeless solution trained solely on (X_1, y_1) to the L2-SP ridge solution found by first training a ridge/ridgeless model on (X_0, y_0), and then using those model parameters as the prior for ridge/ridgeless training on (X_1, y_1).

Our analysis builds on the foundational work characterizing ridge and ridgeless regression risk in high-dimensional settings Dobriban & Wager (2018); Hastie et al. (2022), which established the precise asymptotic behavior of these estimators in overparameterized regimes and revealed phenomena such as double descent Belkin et al. (2019) and benign overfitting Bartlett et al. (2020). We employ random matrix theory techniques, specifically the deterministic equivalent framework Bai & Silverstein (2010); Couillet & Debbah (2011); Dobriban & Wager (2018); Hastie et al. (2022), to derive precise asymptotic characterizations of the estimators' risk and identify sharp phase boundaries for transfer benefit.

Our contributions can be outlined as follows. First, Theorem 3.3 provides an if-and-only-if inequality describing when transfer learning will outperform from-scratch training in the finite data, non-isotropic ridge regime. As a corollary, we show that in the finite sample, isotropic, ridgeless case the inequality takes a simple form and is interestingly independent of n_1 and σ_1. This independence is consistent with prior isotropic analyses of freezing-style transfer Dar & Baraniuk (2022); Lampinen & Ganguli (2018), and our finite-sample isotropic/ridgeless corollary recovers these patterns in the L2-SP setting.

We then examine deterministic equivalents in the asymptotic limit and Theorem 3.6 provides a DE characterization and asymptotic boundary for L2-SP ridge transfer with general (non-orthonormal) task vectors. Corollary 3.7 examines the isotropic case, where the decision criterion is determined entirely by whether or not the alignment of the two tasks surpasses a bias and noise term dependent only on Task 0.

From this we arrive at Theorem 3.8, which identifies the unique τ_0^* (source model ridge penalty) that maximizes transfer benefit, given a particular task alignment. Outside of a measure-zero set, τ_0^* does not coincide with the optimal Task 0 ridge parameter, and notably τ_0^* is independent of target sample size but depends on task alignment. Together, these results give the first explicit analytical boundary for when Euclidean L2-SP transfer helps with general (non-orthonormal) task vectors in overparameterized ridge models, including general covariance and DE limits, and provide surprising insights into optimally training source models for the purpose of transfer learning.

Finally, we establish in Corollary 3.9 an alignment-dependent phase transition for the optimal source penalty. Under standard imperfect task alignment conditions, τ_0^* is always strictly greater than the source-optimal ridge penalty. However, in super-aligned regimes this relationship reverses. This implies that maximizing transfer benefit requires adjusting regularization based on the geometric relationship between tasks, rather than optimizing for source performance alone.

Though derived in linear models, these results isolate a core mechanism that may persist in overparameterized nonlinear networks. Additionally although standard SGD fine-tuning protocols do not necessarily use an L2-SP penalty, this mechanism models the implicit regularization of SGD fine-tuning where the optimization trajectory remains anchored in the basin of the initialization. In Section 4 we validate the theory-derived phase transition in a synthetic ridge setting and then use standard transfer learning experiments on MNIST, CIFAR-10, and 20 Newsgroups to probe how source-optimal versus transfer-optimal regularization behaves in nonlinear networks, finding a consistent preference for over-regularization for transfer across all tested domains.

These results challenge the conventional practice of optimizing source models solely for their own performance. For practitioners training foundation models intended for transfer, our analysis suggests that regularization strategies should explicitly account for the downstream transfer objective and source data quality.

The remainder of the paper is outlined as follows. Section 2 sets up the model and assumptions. Section 3 presents the core mathematical results. Section 4 provides empirical validation of our theoretical predictions on synthetic data, MNIST, CIFAR-10, and 20 Newsgroups. We conclude with Section 5 and discuss implications and limitations; proofs are deferred to the Appendix.

2 Preliminaries

We let Z_0, Z_1 be $n_0, n_1 \times p$ random matrices with entries taken iid with mean 0, variance 1, and finite $2+\epsilon$ moments and Lindeberg condition (see Bai & Silverstein (2010) for more information). These assumptions are sufficient for the deterministic equivalents we will examine later, but it is safe also to simplify these assumptions taking entries in Z_i to be iid $N(0,1)$. We additionally assume we are in the overparameterized regime, with $n_0, n_1 < p-1$. We let Σ_0, Σ_1 be covariance matrices and $X_i = Z_i \Sigma_i^{1/2}$. Next we have true signal vectors w_0, w_1 and let $y_i = X_i w_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_i^2 I)$. We assume the signal norms $||w_0||, ||w_1||$ remain bounded (i.e., $O(1)$) as $p \to \infty$, ensuring that the signal and noise contributions to the risk remain comparable in the asymptotic limit.

We will frequently quantify task relatedness by the normalized alignment

$$\rho := \frac{\langle w_0, w_1 \rangle}{||w_0||^2}.$$

We refer to $0 < \rho < 1$ as the *imperfect alignment* regime and $\rho > 1$ as the *super-aligned* regime.

We adopt the L2-SP approach found in Xuhong et al. (2018) and examine the following estimators:

$$\begin{aligned}\hat{\beta}_0(\lambda_0) &:= \arg\min_{\beta}\ \|y_0 - X_0\beta\|^2 + \lambda_0\|\beta\|^2 \\ &= (X_0^\top X_0 + \lambda_0 I)^{-1} X_0^\top y_0, \\ \hat{\beta}_1^{\mathrm{S}}(\lambda_1) &:= \arg\min_{\beta}\ \|y_1 - X_1\beta\|^2 + \lambda_1\|\beta\|^2 \\ &= (X_1^\top X_1 + \lambda_1 I)^{-1} X_1^\top y_1, \\ \hat{\beta}_1^{\mathrm{TL}}(\lambda_1 \mid \hat{\beta}_0) &:= \arg\min_{\beta}\ \|y_1 - X_1\beta\|^2 + \lambda_1\|\beta - \hat{\beta}_0(\lambda_0)\|^2 \\ &= (X_1^\top X_1 + \lambda_1 I)^{-1}\big(X_1^\top y_1 + \lambda_1\,\hat{\beta}_0(\lambda_0)\big).\end{aligned}$$

We also will take the notation that $\hat{\beta}_0(0) = \lim_{\lambda_0 \searrow 0} \hat{\beta}_0(\lambda_0)$ represents the ridgeless estimator (and similarly for $\hat{\beta}_1^S$ and $\hat{\beta}_1^{TL}$). In this setting of course $\hat{\beta}_0$ represents our ridge/ridgeless estimator for Task 0, $\hat{\beta}_1^S$ is our standard ridge/ridgeless estimator for Task 1, and $\hat{\beta}_1^{TL}$ is our transfer learning estimator for Task 1 that takes the solution of Task 0 as its prior.

We note here that $\hat{\beta}_1^{\mathrm{TL}}(\lambda_1 \mid \hat{\beta}_0)$ corresponds exactly to the MAP estimator with Gaussian prior $\beta \sim N(\hat{\beta}_0, \lambda_1^{-1} I)$ and matches practical L2-SP fine-tuning Xuhong et al. (2018) and the isotropic EWC penalty Kirkpatrick et al. (2017). While our theoretical analysis assumes an explicit L2-SP penalty, standard fine-tuning protocols rely on the implicit regularization of SGD initialized at θ_0. For limited training horizons, the optimization trajectory remains anchored in the basin of θ_0, functionally approximating the L2-SP constraint. Thus, we expect the regularization-variance trade-offs identified in our ridge analysis to persist in standard fine-tuning. We also note that we do not reweight our penalty by a task metric H and stick with standard Euclidean distancing for our ridge penalty. This differs from the whitened/metric-based formulations and also more accurately reflects typical implementations of L2-SP and fine-tuning where the penalty is applied in Euclidean parameter space without whitening.

To evaluate out-of-sample performance of these estimators we will use their expected prediction risk:

$$\mathrm{Risk}_1(\beta) := \mathbb{E}_{x_1 \sim N(0,\Sigma_1)}[(x_1^\top \beta - x_1^\top w_1)^2] = \|\beta - w_1\|_{\Sigma_1}^2,$$

where $||v||_\Sigma^2 = v^\top \Sigma v$.

Our first goal is to understand when L2-SP transfer improves expected Task 1 risk, namely when

$$\mathrm{Risk}_1(\hat{\beta}_1^{\mathrm{S}}(\lambda_1)) > \mathrm{Risk}_1(\hat{\beta}_1^{\mathrm{TL}}(\lambda_1 \mid \hat{\beta}_0(\lambda_0))).$$

We will additionally make use of the Frobenius norm in Σ geometry, which we will denote as follows:

$$||A||_{\Sigma,F}^2 = \mathrm{Tr}\left(A^\top \Sigma A\right).$$

Finally, some common matrices we will be using deserve their own notation, and we define that here: let

$$M_{\lambda_1}^{(i)} = \left(X_i^\top X_i + \lambda_1 I\right)^{-1}$$

and note that

$$M_{\lambda_1}^{(1)} X_1^\top X_1 - I = -\lambda_1 M_{\lambda_1}^{(1)}.$$

Additionally let

$$P_i = X_i^+ X_i,$$

where X_i^+ is the Moore-Penrose pseudoinverse, so that P_i is the orthogonal projector in $\mathbb{R}^p$ onto $\mathrm{row}(X_i)$ and $I - P_i$ projects onto $\ker(X_i)$.

3 Mathematical Results

3.1 Finite Sample Risk Formulas

First we recall the expected out of sample risk of ridge regression (see Hastie et al. (2022) for a modern reference)

Observation 3.1. *The expected risk of ridge regression is*

$$R^S(\lambda_1) = \lambda_1^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} w_1||_{\Sigma_1}^2\right] + \sigma_1^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} X_1^\top||_{\Sigma_1,F}^2\right].$$

Next we compute the finite sample risk for L2-SP ridge regression:

Lemma 3.2. *The expected risk of the transfer estimator decomposes into pure bias, variance induced by the β_0 prior, and variance induced by estimation error:*

$$R^{TL}(\lambda_1) = B^{TL}(\lambda_1) + \sigma_0^2 V_0^{TL}(\lambda_1) + \sigma_1^2 V_1^{TL}(\lambda_1)$$

with:

$$B^{TL}(\lambda_1) = \lambda_1^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0 - M_{\lambda_1}^{(1)} w_1||_{\Sigma_1}^2\right]$$

$$V_0^{TL}(\lambda_1) = \lambda_1^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top||_{\Sigma_1,F}^2\right]$$

$$V_1^{TL}(\lambda_1) = \mathbb{E}\left[||M_{\lambda_1}^{(1)} X_1^\top||_{\Sigma_1,F}^2\right].$$

The following theorem characterizes when transfer helps by comparing the bias introduced by starting from zero (standard ridge) versus starting from the source parameters (L2-SP). Transfer succeeds when the alignment between the tasks, after appropriate filtering through the ridge resolvent and covariance geometry, exceeds a threshold determined by source bias and noise. Note that since the Task 1 variance portion of risk for both of these estimators is the same, it is independent from the decision of whether or not transfer will benefit.

Theorem 3.3. *In the finite sample case with $\lambda_1 > 0$, we gain benefit from transfer learning ($R^{TL}(\lambda_1) < R^S(\lambda_1)$) if and only if:*

$$\begin{aligned} &2\mathbb{E}\left[\langle M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0, M_{\lambda_1}^{(1)} w_1\rangle_{\Sigma_1}\right] \\ &\quad > \mathbb{E}\left[||M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0||_{\Sigma_1}^2\right] + \sigma_0^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top||_{\Sigma_1,F}^2\right]. \end{aligned}$$

As a corollary, in the ridgeless limit we consider when X_i is taken to be isotropic and Gaussian so that the Wishart formulas apply exactly:

Corollary 3.4. *In the finite case, if $\Sigma_i = I$ and $\lambda_1 = \lambda_0 = 0$ and X_i is taken to be Gaussian, then $R^{TL}(0) < R^S(0)$ if and only if:*

$$2\langle w_0, w_1\rangle > ||w_0||^2 + \sigma_0^2 \frac{p}{p - n_0 - 1}.$$

A noteable feature of the isotropic ridgeless boundary is its complete independence from n_1 and σ_1: the transfer decision depends only on task alignment $\langle w_0, w_1\rangle$ and source characteristics $(||w_0||, \sigma_0, n_0, p)$. If transfer outperforms training from scratch with 10 target samples, it also outperforms training from scratch with 10,000 target samples (and the same statement holds across target noise levels). Collecting more target data does not change whether transfer helps in this isotropic regime; only task alignment and source-side quality matter.

We also observe here that the transfer region monotonically shrinks as σ_0 grows. Thus, in order to maximize the potential transfer benefit it is essential to ensure source task noise is as small as possible.

3.2 Deterministic Equivalents and Asymptotics

We will now examine asymptotics for the ridge phase transition identified above and use deterministic equivalents to understand the limiting phase transition. To establish the core deterministic equivalents, we will scale our λ_i with n_i and let $\tau_i = \lambda_i/n_i$. We define the following ridge resolvent: Let $S_i := n_i^{-1} X_i^\top X_i$ and $\gamma_i := \lim_{p\to\infty} p/n_i$. Under standard Bai–Silverstein assumptions (e.g. bounded spectral norm of Σ_i and weak convergence of its empirical spectral distribution), the resolvent $(S_i + \tau_i I)^{-1}$ admits a deterministic equivalent of the form

$$Q_i(\tau_i) = \left(\tau_i I + \tilde{\delta}_i(\tau_i)\Sigma_i\right)^{-1},$$

where the scalar pair $(\delta_i(\tau_i), \tilde{\delta}_i(\tau_i))$ is the unique positive solution to

$$\delta_i(\tau_i) = \frac{1}{n_i}\operatorname{Tr}\left(\Sigma_i Q_i(\tau_i)\right), \qquad \tilde{\delta}_i(\tau_i) = \frac{1}{1+\delta_i(\tau_i)}.$$

We will use the notation $A_n \asymp B_n$ to denote deterministic equivalent convergence, meaning that for any sequence of deterministic matrices D_n with uniformly bounded spectral norm, we have $\operatorname{Tr}(D_n(A_n - B_n)) \to 0$ almost surely as $n \to \infty$. This is the standard notion of weak deterministic equivalence used in random matrix theory Bai & Silverstein (2010).

First, the following deterministic equivalents will be useful:

Observation 3.5. *Under standard Bai-Silverstein conditions (Bai & Silverstein (2010)), as $p, n_i \to \infty$ with $p/n_i \to \gamma_i$:*

$$n_1 M^{(1)}_{\lambda_1} = (S_1 + \tau_1 I)^{-1} \asymp Q_1(\tau_1),$$

$$\lambda_1 M^{(1)}_{\lambda_1} \asymp \tau_1 Q_1(\tau_1),$$

and

$$n_0 M^{(0)}_{\lambda_0} X_0^\top X_0 M^{(0)}_{\lambda_0} \asymp Q_0(\tau_0) - \tau_0 Q_0(\tau_0)^2.$$

Finally, define $t(\tau_0, \tau_1)$ by:

$$\lim_{p\to\infty} p^{-1} \operatorname{Tr}\left(Q_1(\tau_1)\Sigma_1 Q_1(\tau_1)\left(Q_0(\tau_0) - \tau_0 Q_0(\tau_0)^2\right)\right),$$

Under the Bai-Silverstein assumptions, the eigenvalue distributions of Σ_0 and Σ_1 converge weakly and the resolvents $Q_i(\tau_i)$ are uniformly bounded, ensuring that this limit exists and equals the limit of the corresponding trace per dimension. By independence of X_0 and X_1, we may substitute these deterministic equivalents into the results from Theorem 3.3 to arrive at the following asymptotic decision criterion:

Theorem 3.6. *In the asymptotic limit, we gain benefit from transfer learning ($R^{TL}(\lambda_1) < R^S(\lambda_1)$) if and only if:*

$$\begin{aligned} &2\left\langle Q_1(\tau_1)(I - \tau_0 Q_0(\tau_0))w_0, Q_1(\tau_1)w_1\right\rangle_{\Sigma_1} \\ &\quad > ||Q_1(\tau_1)(I - \tau_0 Q_0(\tau_0))w_0||^2_{\Sigma_1} + \sigma_0^2\gamma_0 t(\tau_0, \tau_1). \end{aligned}$$

And in the isotropic case:

Corollary 3.7. *In the isotropic case where $\Sigma_0 = \Sigma_1 = I$, we (asymptotically) gain benefit from transfer learning ($R^{TL}(\lambda_1) < R^S(\lambda_1)$) if and only if:*

$$2\left\langle w_0, w_1\right\rangle > (1 - \tau_0 a_0)||w_0||^2 + \sigma_0^2\gamma_0 a_0,$$

where a_0 is the unique positive solution to

$$\tau_0\gamma_0 a_0^2 + (\tau_0 + 1 - \gamma_0)a_0 - 1 = 0.$$

As in the finite sample isotropic ridgeless case, in the isotropic DE (ridge) setting this decision boundary is independent of τ_1, γ_1, and σ_1. In particular, whether transfer helps is a source-side question: it does not depend on target sample size, target noise, or target regularization.

As in the finite sample isotropic ridgeless case, alignment of the two tasks must pass a threshold that is independent of other target task characeristics. This transfer benefit region is again monotonically shrinking in σ_0, and may be maximized by minimizing the quantity on the right. By fixing γ_0 we arrive at the following result:

Theorem 3.8. *In the asymptotic setting with isotropic data and fixed source model overparameterization γ_0, for any fixed normalized alignment $\rho = \langle w_0, w_1 \rangle / ||w_0||^2$, there exists a unique source ridge penalty τ_0^* that maximizes the transfer benefit $\Delta R = R^S - R^{TL}$. Additionally, whenever $\langle w_0, w_1 \rangle \neq ||w_0||^2$, the optimal τ_0^* for transfer learning differs from the optimal ridge penalty for Task 0 performance.*

As a corollary, we identify that there is a consistent relationship between transfer-optimal and source-optimal regularization.

Corollary 3.9. *Let τ_0^* be the transfer-optimal regularization penalty and τ_S^* be the source-optimal regularization penalty.*

- *If the tasks are imperfectly aligned ($0 < \rho < 1$), then $\tau_0^* > \tau_S^*$ (transfer requires stronger regularization).*
- *If the tasks are super-aligned ($\rho > 1$), then $\tau_0^* < \tau_S^*$ (transfer requires weaker regularization).*

This phase transition depends solely on task alignment and holds for all noise levels $\sigma_0 > 0$.

In particular, this shows that when the tasks are imperfectly aligned ($0 < \rho < 1$), the transfer-optimal solution always uses more regularization than one would typically use for the source task alone.

4 Empirical Validation

In this section we provide a controlled synthetic experiment to validate the precise phase transition predictions. Additionally, while our theoretical analysis focuses on linear ridge regression, we study nonlinear networks on MNIST, CIFAR-10, and 20 Newsgroups to test whether the source-optimal versus transfer-optimal misalignment persists beyond linear models.

4.1 Synthetic Validation of Phase Transition

4.1.1 Experimental Setup

To verify the alignment-dependent phase transition predicted by Corollary 3.9, we conduct a controlled synthetic experiment using the generative model defined in Section 2. We set $p = 500$, $n_{source} = 250$ ($\gamma = 2.0$), and $n_{target} = 50$. We fix the source noise $\sigma_0^2 = 1.0$ and target noise $\sigma_1^2 = 0.1$.

We sweep the task alignment $\rho = \langle w_0, w_1 \rangle / ||w_0||^2$ from 0.5 to 1.5, covering both the imperfect alignment regime ($\rho < 1$) and the super-aligned regime ($\rho > 1$). For each alignment level, we: 1. Generate source and target data according to the specified alignment. 2. Train source ridge models over a logarithmic grid of regularization strengths λ. 3. Select the source-optimal λ_S^* that minimizes risk on the source task. 4. For each trained source model, fine-tune on the target task (using L2-SP with fixed target regularization $\lambda_1 = 0.1$) and identify the transfer-optimal source regularization λ_{TL}^*. 5. Compute the ratio $\lambda_{TL}^*/\lambda_S^*$.

We repeat this procedure across 10 random seeds to estimate confidence intervals.

4.1.2 Results

Figure 1 displays the ratio of transfer-optimal to source-optimal regularization as a function of task alignment. The results match our theoretical predictions:

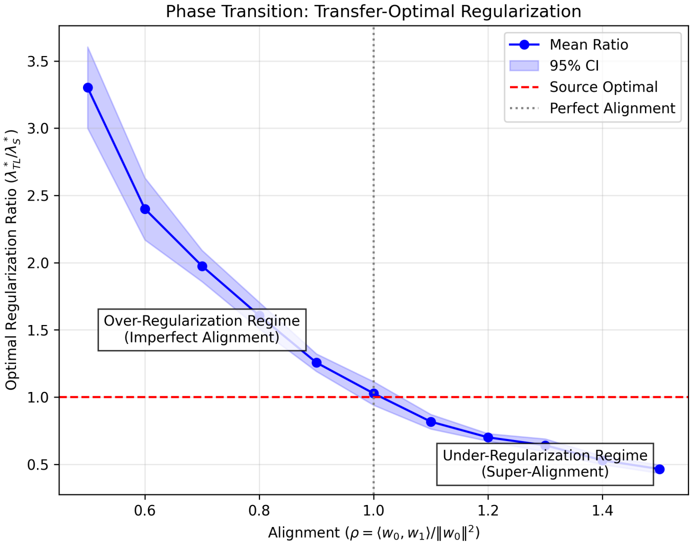

Figure 1: Synthetic validation of the alignment-dependent phase transition. The y-axis shows the ratio of transfer-optimal to source-optimal regularization ($\lambda^*_{TL}/\lambda^*_S$). A ratio > 1 indicates over-regularization is beneficial (Standard Regime), while < 1 indicates under-regularization is optimal (Super-Aligned Regime). The shaded region represents the 95% confidence interval over 10 seeds.

- **Over-Regularization Regime ($\rho < 1$):** When alignment is imperfect (standard transfer), the ratio is consistently > 1, indicating that $\lambda^*_{TL} > \lambda^*_S$. For example, at $\rho = 0.8$, the transfer-optimal penalty is approximately 1.6× larger than the source-optimal penalty.

- **Under-Regularization Regime ($\rho > 1$):** When the tasks are super-aligned, the ratio drops below 1, confirming that $\lambda^*_{TL} < \lambda^*_S$. At $\rho = 1.2$, the optimal source penalty for transfer is roughly 0.7× the source-optimal value.

- **Phase Transition:** The crossover occurs precisely at $\rho = 1.0$, matching the theoretical boundary where the target signal magnitude equals the source signal magnitude.

4.2 Nonlinear Transfer Learning Experiments

To test whether the source-optimal versus transfer-optimal mismatch persists beyond linear models, we conduct standard transfer learning experiments across three domains: vision (MNIST, CIFAR-10) and text (20 Newsgroups). Each experiment follows the same protocol: (1) train source models with varying weight decay, (2) transfer to target task with fixed fine-tuning protocol, (3) compare source-optimal versus transfer-optimal source regularization.

4.2.1 Experimental Setup

MNIST. We use a 2-layer MLP (784→128→64→5) and split digits into source (round digits: 0,3,6,8,9) and target (angular digits: 1,2,4,5,7). The target training set is subsampled to 10% to create a realistic limited-data transfer scenario.

CIFAR-10. We use a small CNN (two conv layers, one FC layer, ∼530K parameters) and split classes into source (animals: bird, cat, deer, dog, frog) and target (vehicles+: airplane, automobile, horse, ship, truck). The target training set is subsampled to 5%.

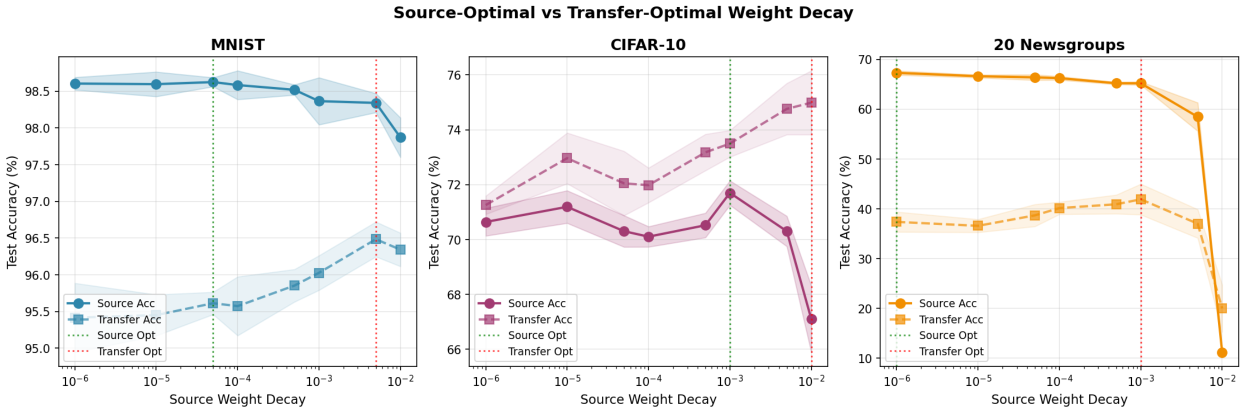

Figure 2: Source-optimal versus transfer-optimal weight decay across three standard transfer learning benchmarks. Each panel shows source accuracy (solid) and transfer accuracy (dashed) as a function of source weight decay, with vertical lines marking optimal values. Across all domains, transfer-optimal performance occurs at stronger source regularization than source-optimal performance, consistent with the over-regularization regime for imperfectly aligned tasks.

20 Newsgroups. We use TF-IDF features (5000 dimensions) with a 2-layer MLP (5000→256→128→num_classes) and split categories into source (tech: comp.*, sci.*) and target (non-tech: rec.*, talk.*, misc.*, alt.atheism). The target training set is subsampled to 10%.

For all experiments, we sweep source weight decay over $[0, 10^{-5}, 5{\times}10^{-5}, 10^{-4}, 5{\times}10^{-4}, 10^{-3}, 5{\times}10^{-3}, 10^{-2}]$ and use a fixed transfer protocol (learning rate 10^{-3}, weight decay 10^{-4}). Results are averaged over 5 random seeds.

4.2.2 Results

Figure 2 shows that across all three domains, the source-optimal and transfer-optimal regularization differ substantially. In each case, transfer-optimal performance (red dashed line) occurs at stronger source regularization than source-optimal performance (green dashed line), consistent with the over-regularization regime predicted by our theory for imperfectly aligned tasks.

MNIST: Source accuracy peaks at low weight decay ($\sim 10^{-5}$), while transfer accuracy peaks at higher weight decay ($\sim 10^{-2}$), a difference of approximately three orders of magnitude.

CIFAR-10: Source accuracy peaks around 10^{-3}, while transfer accuracy peaks at 10^{-2}, demonstrating the mismatch persists in convolutional architectures.

20 Newsgroups: Source accuracy peaks at low weight decay ($\sim 10^{-5}$), while transfer accuracy peaks around 10^{-3}, showing the phenomenon extends to text classification with TF-IDF features.

These results consistently validate that source-optimal training is suboptimal for transfer learning across vision and NLP domains, with transfer uniformly preferring stronger source regularization.

5 Conclusion

We have proven a fundamental misalignment in transfer learning: training a source model to minimize its own risk is generically suboptimal for maximizing transfer benefit. This source-optimal versus transfer-optimal divergence implies that source regularization should be chosen with the downstream objective in mind and depends on task alignment. We provide explicit transfer-versus-scratch boundaries for L2-SP ridge that hold at finite p, n and extend to general covariance via deterministic equivalents. In isotropic limits, whether transfer helps is independent of target sample size and noise: if transfer helps with 10 target samples, it helps with 10,000. Optimizing the source model target task performance yields a unique source ridge τ_0^* that differs generically from the source-task-optimal choice.

These results give an analytical foundation for the idea that pretraining optimized for transfer differs fundamentally from pretraining aimed at standalone accuracy. They suggest re-evaluating regularization and objective design in pretraining, foundation-model, and continual-learning pipelines. In particular, practitioners training source models with the goal of downstream transfer should carefully consider the alignment between source and target tasks. In standard imperfect alignment scenarios, stronger regularization is required to shrink the source estimator towards the optimal target projection.

Our experiments on MNIST, CIFAR-10, and 20 Newsgroups indicate that the misalignment between source-optimal and transfer-optimal regularization persists beyond the linear setting across vision and text domains. In all nonlinear regimes we tested, the transfer-optimal choice consistently involves stronger source regularization than source-optimal training. We did not observe a super-aligned regime where under-regularizing the source improves transfer, suggesting that exhibiting the under-regularization phase may be substantially more difficult in standard deep-learning pipelines where feature learning becomes an important aspect of the model training process.

5.1 Practical Implications

Our results provide several concrete insights for practitioners working with transfer learning systems. First, minimizing source noise is critical for maximizing the region of positive transfer, as the transfer benefit shrinks monotonically with σ_0^2. This suggests that investing in data cleaning, noise reduction, or improved labeling for the source task can pay significant dividends for downstream transfer, beyond simply improving source task performance. Second, the independence of the transfer decision from target sample size and noise in isotropic settings suggests that if transfer is beneficial for one target dataset size, it remains beneficial across different scales of target data, provided the task structure remains similar. This can simplify model selection in scenarios where target data availability is uncertain or variable.

The alignment-dependent phase transition provides actionable guidance for model training in linear ridge: when tasks exhibit standard imperfect alignment ($\langle w_0, w_1 \rangle < ||w_0||^2$), one should increase regularization beyond what would be optimal for source task performance alone, effectively trading off some source performance for better transferability. In the nonlinear regimes we tested across MNIST, CIFAR-10, and 20 Newsgroups, we consistently observed this over-regularization behavior for transfer.

Since most practical purposes likely lie in the standard imperfect alignment regime, it is likely the case that over-regularization during source training may lead to better outcomes for transfer learning. Additionally, as noted in Figure 1, this divergence can be quite non-trivial, where even with a relatively strong alignment of 0.8, the transfer-optimal regularization penalty was $\sim 60\%$ higher than source-optimal.

We note here that while our analysis focuses on L2-SP ridge regression, our experiments across MNIST, CIFAR-10, and 20 Newsgroups indicate these insights extend qualitatively to more complex fine-tuning protocols in deep learning, suggesting that foundation model training pipelines may benefit from regularization strategies that explicitly account for downstream transfer objectives rather than solely optimizing upstream performance.

5.2 Scope and Future Directions

Our results establish the source-optimal versus transfer-optimal phenomenon in the tractable linear setting and provide a foundation for understanding transfer in complex models. The precise functional forms and boundaries we derive are specific to linear ridge regression. Our experiments on MNIST, CIFAR-10, and 20 Newsgroups suggest the source/transfer regularization misalignment persists in nonlinear networks across vision and text domains, while the super-aligned under-regularization regime appears difficult to realize in standard deep-learning setups. We assume Gaussian noise and employ either exact isotropy or Bai-Silverstein conditions for our deterministic equivalent results. Extensions to more general noise models and covariance structures remain important open problems.

We focus on the L2-SP protocol, which penalizes Euclidean distance from the source parameters. This differs from more sophisticated fine-tuning methods that might incorporate task-specific metrics, feature-space regularization, or adaptive penalties based on parameter importance. Additionally, L2-SP represents a closed-form ridge solution, while practical deep learning relies on iterative optimization via gradient descent. However, the implicit regularization of early-stopped SGD typically keeps parameters close to initialization, qualitatively matching the L2-SP behavior. The recent work of Ghane, Akhtiamov, and Hassibi Ghane et al. (2024) on gradient-descent-based fine-tuning provides complementary insights for the iterative case, and connecting these perspectives remains an interesting direction.

Finally, our results focus on two-task transfer, whereas many practical scenarios involve continual learning across multiple sequential tasks or multi-task learning with many simultaneous objectives. The interplay between multiple tasks and optimal source regularization represents a rich area for future investigation.

Connecting our ridge-based findings to gradient-descent dynamics in deep learning represents an important bridge to practice. Understanding how implicit regularization in over-parameterized neural networks interacts with the mechanisms we identify could inform training procedures for foundation models. In particular, investigating whether early stopping, learning rate schedules, or architectural choices can implicitly achieve transfer-optimal regularization even when explicit penalties are not used would provide actionable insights for practitioners. Analyzing multi-task and continual learning scenarios where multiple target tasks must be considered simultaneously could reveal how to optimize source training when facing diverse downstream objectives with potentially conflicting requirements.

Though derived in the tractable setting of linear ridge regression, our results isolate a fundamental tension between optimizing for source performance versus transfer capability that likely manifests across a broad range of learning systems. By making this tension analytically precise, we hope to inform the design of pretraining objectives that explicitly account for downstream transfer goals.

References

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

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

Romain Couillet and Merouane Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

Yehuda Dar and Richard G Baraniuk. Double double descent: On generalization errors in transfer learning between linear regression tasks. *SIAM Journal on Mathematics of Data Science*, 4(4):1447–1472, 2022.

Yehuda Dar, Daniel LeJeune, and Richard G Baraniuk. The common intuition to transfer learning can win or lose: Case studies for linear regression. *SIAM Journal on Mathematics of Data Science*, 6(2):454–480, 2024.

Oussama Dhifallah and Yue M Lu. Phase transitions in transfer learning for high-dimensional perceptrons. *Entropy*, 23(4):400, 2021.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Reza Ghane, Danil Akhtiamov, and Babak Hassibi. Universality in transfer learning for linear models. *Advances in Neural Information Processing Systems*, 37:125729–125779, 2024.

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Andrew K Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. *arXiv preprint arXiv:1809.10374*, 2018.

Javan Tahir, Surya Ganguli, and Grant M Rotskoff. Features are fate: a theory of transfer learning in high-dimensional regression. *arXiv preprint arXiv:2410.08194*, 2024.

LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International conference on machine learning*, pp. 2825–2834. PMLR, 2018.

Fan Yang, Hongyang R Zhang, Sen Wu, Christopher Re, and Weijie J Su. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *Journal of Machine Learning Research*, 26(113):1–88, 2025.

A Detailed Proofs

Here we will re-state and prove the relevant Lemmas and Theorems mentioned above.

A.1 Finite Sample Risk Formulas

We begin by re-writing our L2-SP transfer learning estimator for the reader's convenience:

$$\hat{\beta}_1^{\mathrm{TL}}(\lambda_1 \mid \hat{\beta}_0) = (X_1^\top X_1 + \lambda_1 I)^{-1}\Big(X_1^\top y_1 + \lambda_1\, \hat{\beta}_0(\lambda_0)\Big).$$

We additionally note that expected risk can be computed by:

$$R(\beta) = E\left[||\beta - w_1||_{\Sigma_1}^2\right].$$

We can now prove the first important Lemma:

Lemma 3.2 The expected risk of the transfer estimator decomposes into pure bias, variance induced by the β_0 prior, and variance induced by estimation error:

$$R^{TL}(\lambda_1) = B^{TL}(\lambda_1) + \sigma_0^2 V_0^{TL}(\lambda_1) + \sigma_1^2 V_1^{TL}(\lambda_1)$$

with:

$$B^{TL}(\lambda_1) = \lambda_1^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0 - M_{\lambda_1}^{(1)} w_1||_{\Sigma_1}^2\right]$$

$$V_0^{TL}(\lambda_1) = \lambda_1^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top ||_{\Sigma_1,F}^2\right]$$

$$V_1^{TL}(\lambda_1) = \mathbb{E}\left[||M_{\lambda_1}^{(1)} X_1^\top ||_{\Sigma_1,F}^2\right].$$

Proof. We verify the decomposition as follows. First note:

$$\begin{aligned}
\hat{\beta}_1^{\mathrm{TL}}(\lambda \mid \hat{\beta}_0) - w_1 &= \lambda M_\lambda^{(1)} \hat{\beta}_0^S(\lambda_0) + \hat{\beta}_1^S(\lambda) - w_1 \\
&= \lambda M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top y_0 + M_\lambda^{(1)} X_1^\top y_1 - w_1 \\
&= \lambda M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0 + \lambda M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top \epsilon_0 + M_\lambda^{(1)} X_1^\top X_1 w_1 + M_\lambda^{(1)} X_1^\top \epsilon_1 - w_1 \\
&= \Big(\lambda M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 + \Big(M_\lambda^{(1)} X_1^\top X_1 - I\Big)\Big) w_1 + \Big(\lambda M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top\Big)\epsilon_0 + \Big(M_\lambda^{(1)} X_1^\top\Big)\epsilon_1.
\end{aligned}$$

We then take expectation of the squared Σ_1 norm of the above expression and since ϵ_0, ϵ_1 are taken independently with mean 0, the cross terms cancel and we are left with the desired result. □

We now remind the reader that the expected risk of the standard Ridge estimator is

$$R^S(\lambda_1) = \lambda_1^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} w_1||_{\Sigma_1}^2\right] + \sigma_1^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} X_1^\top ||_{\Sigma_1,F}^2\right].$$

We can therefore compare these quantities to identify when we expect Transfer to outperform training from scratch.

Theorem 3.3. In the finite sample case with $\lambda_1 > 0$, we gain benefit from transfer learning ($R^{TL}(\lambda_1) < R^S(\lambda_1)$) if and only if:

$$2\mathbb{E}\left[\langle M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0, M_{\lambda_1}^{(1)} w_1 \rangle_{\Sigma_1}\right] > \mathbb{E}\left[||M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0||_{\Sigma_1}^2\right] + \sigma_0^2 \mathbb{E}\left[||M_{\lambda_1}^{(1)} M_{\lambda_0}^{(0)} X_0^\top ||_{\Sigma_1,F}^2\right].$$

Proof. First we notice the σ_1 terms of $R^S(\lambda_1)$ and $R^{TL}(\lambda_1|\beta_0)$ exactly cancel. We are therefore left with

$$R^S(\lambda) - R^{TL}(\lambda) = \mathbb{E}\left[||(M_\lambda^{(1)} X_1^\top X_1 - I)w_1||_{\Sigma_1}^2\right] - \mathbb{E}\left[||\lambda M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0 + \left(M_\lambda^{(1)} X_1^\top X_1 - I\right) w_1||_{\Sigma_1}^2\right]$$

$$-\sigma_0^2 \lambda^2 \mathbb{E}\left[||M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top||_{\Sigma_1,F}^2\right]$$

$$= -\mathbb{E}\left[||\lambda M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0||_{\Sigma_1}^2\right] - 2\lambda \mathbb{E}\left[\langle (M_\lambda^{(1)} X_1^\top X_1 - I)w_1, M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0 \rangle_{\Sigma_1}\right]$$

$$-\sigma_0^2 \lambda^2 \mathbb{E}\left[||M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top||_{\Sigma_1,F}^2\right].$$

We note here that $(M_\lambda^{(1)} X_1^\top X_1 - I) = -\lambda M_\lambda^{(1)}$ and thus we have

$$\mathbb{E}\left[\langle (M_\lambda^{(1)} X_1^\top X_1 - I)w_1, M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0 \rangle_{\Sigma_1}\right] = -\lambda \mathbb{E}\left[\langle M_\lambda^{(1)} w_1, M_\lambda^{(1)} M_{\lambda_0}^{(0)} X_0^\top X_0 w_0 \rangle_{\Sigma_1}\right].$$

After some algebra and canceling the common λ^2 terms, we arrive at our desired inequality. □

Corollary 3.4 In the finite case, if $\Sigma_i = I$ and $\lambda_1 = \lambda_0 = 0$, and X_i is taken to be Gaussian, then $R^{TL}(0) < R^S(0)$ if and only if:

$$2\langle w_0, w_1 \rangle > ||w_0||^2 + \sigma_0^2 \frac{p}{p - n_0 - 1}.$$

Proof. We first note

$$\lim_{\lambda \searrow 0} M_{\lambda_1}^{(1)} = \lim_{\lambda \searrow 0} (X_1^\top X_1 + \lambda I)^{-1} = (X_1^\top X_1)^+ = I - X_1^+ X_1.$$

By conditioning on X_0 and taking expectation with respect to X_1, the projection $I - X_1^+ X_1$ introduces a common factor of $\frac{p-n_1}{p}$ to all terms involving w_1 or the X_0-dependent prior (due to the isotropy of X_1). Specifically, for any fixed vector v, $\mathbb{E}_{X_1}[||(I - X_1^+ X_1)v||^2] = \frac{p-n_1}{p}||v||^2$. We can thus eliminate the target projection effects from the expression. We are then left with:

$$2\mathbb{E}\left[\langle M_{\lambda_0}^{(0)} X_0^\top X_0 w_0, w_1 \rangle\right] > \mathbb{E}\left[||M_{\lambda_0}^{(0)} X_0^\top X_0 w_0||_{\Sigma_1}^2\right] + \sigma_0^2 \mathbb{E}\left[||M_{\lambda_0}^{(0)} X_0^\top||_{\Sigma_1,F}^2\right].$$

We now note that

$$\lim_{\lambda_0 \searrow 0} M_{\lambda_0}^{(0)} X_0^\top = X_0^+,$$

and thus we have

$$2\mathbb{E}\left[\langle X_0^+ X_0 w_0, w_1 \rangle\right] > \mathbb{E}\left[||X_0^+ X_0 w_0||_{\Sigma_1}^2\right] + \sigma_0^2 \mathbb{E}\left[||X_0^+||_{\Sigma_1,F}^2\right].$$

We note here that since $\Sigma_0 = I$ we have $E\left[X_0^+ X_0\right] = \frac{n_0}{p} I$. Additionally, $E\left[||X_0^+||_F^2\right] = \frac{n_0}{p-n_0-1}$. We can now take expectation to see

$$2\frac{n_0}{p}\langle w_0, w_1 \rangle > \frac{n_0}{p}||w_0||^2 + \sigma_0^2 \frac{n_0}{p - n_0 - 1}.$$

By multiplying by $\frac{p}{n_0}$ we arrive at the desired result. □

A.2 Deterministic Equivalents and Asymptotics

We will now examine asymptotics for the ridge phase transition identified above and use deterministic equivalents to understand the limiting phase transition. To establish the core deterministic equivalents, we must first let $\tau_i = \lambda_i / n_i$, and we define the following ridge resolvent: Let $S_i := n_i^{-1} X_i^\top X_i$ and $\gamma_i := \lim_{p \to \infty} p/n_i$. Under standard Bai–Silverstein assumptions, the resolvent $(S_i + \tau_i I)^{-1}$ admits a deterministic equivalent

$$Q_i(\tau_i) = \left(\tau_i I + \tilde{\delta}_i(\tau_i)\Sigma_i\right)^{-1},$$

where the scalar pair $(\delta_i(\tau_i), \tilde{\delta}_i(\tau_i))$ is the unique positive solution to

$$\delta_i(\tau_i) = \frac{1}{n_i} \operatorname{Tr}\left(\Sigma_i Q_i(\tau_i)\right), \qquad \tilde{\delta}_i(\tau_i) = \frac{1}{1+\delta_i(\tau_i)}.$$

Observation 3.5 Under standard Bai-Silverstein conditions (Bai & Silverstein (2010)), as $p, n_i \to \infty$ with $p/n_i \to \gamma_i$:

$$n_1 M_{\lambda_1}^{(1)} = (S_1 + \tau_1 I)^{-1} \asymp Q_1(\tau_1),$$

$$\lambda_1 M_{\lambda_1}^{(1)} \asymp \tau_1 Q_1(\tau_1),$$

and

$$n_0 M_{\lambda_0}^{(0)} X_0^\top X_0 M_{\lambda_0}^{(0)} \asymp Q_0(\tau_0) - \tau_0 Q_0(\tau_0)^2.$$

Proof. These equivalences follow from standard results in Bai & Silverstein (2010). Specifically, under the Bai-Silverstein conditions, the empirical resolvent $(X_i^\top X_i/n_i + \lambda_i I)^{-1}$ admits the deterministic equivalent $n_i^{-1} Q_i(\tau_i) = n_i^{-1}(\tau_i I + \delta_i(\tau_i)\Sigma_i)^{-1}$, where $\delta_i(\tau_i)$ is the unique positive solution to the Silverstein fixed-point equation. By dividing by λ_1 we arrive at the first deterministic equivalence.

On the sample side we know

$$M_{\lambda_0}^{(0)} X_0^\top = X_0^\top M_{\lambda_0}^{(0)} = n_0^{-1} X_0^\top \left(X_0^\top X_0/n_0 + \lambda_0 I\right)^{-1},$$

and the second deterministic equivalent follows.

Finally we use the following fact:

$$M_{\lambda_0}^{(0)} X_0^\top X_0 M_{\lambda_0}^{(0)} = M_{\lambda_0}^{(0)} \left((X_0^\top X_0 + \lambda_0 I) - \lambda_0 I\right) M_{\lambda_0}^{(0)} = M_{\lambda_0}^{(0)} - \lambda_0 (M_{\lambda_0}^{(0)})^2.$$

From this it is easy to see the third deterministic equation holds. □

Finally, define $t(\tau_0, \tau_1)$ by:

$$\lim_{p\to\infty} p^{-1} \operatorname{Tr}\left(Q_1(\tau_1)\Sigma_1 Q_1(\tau_1)\left(Q_0(\tau_0) - \tau_0 Q_0(\tau_0)^2\right)\right),$$

which we know exists under the Bai-Silverstein assumptions. By independence of X_0 and X_1, we may substitute these in to the results from Theorem 3.3 to arrive at the following asymptotic decision criterion:

Theorem 3.6 In the asymptotic limit, we gain benefit from transfer learning ($R^{TL}(\lambda_1) < R^S(\lambda_1)$) if and only if:

$$2\left\langle Q_1(\tau_1)(I - \tau_0 Q_0(\tau_0))w_0, Q_1(\tau_1)w_1\right\rangle_{\Sigma_1} > ||Q_1(\tau_1)(I - \tau_0 Q_0(\tau_0))w_0||_{\Sigma_1}^2 + \sigma_0^2 \gamma_0 t(\tau_0, \tau_1).$$

Proof. This result follows directly from substituting the deterministic equivalents from Observation 3.5 into the results from Theorem 3.3 combined with the independence of X_1 and X_0. □

Corollary 3.7 In the isotropic case where $\Sigma_0 = \Sigma_1 = I$, we (asymptotically) gain benefit from transfer learning ($R^{TL}(\lambda_1) < R^S(\lambda_1)$) if and only if:

$$2\langle w_0, w_1\rangle > (1 - \tau_0 a_0)||w_0||^2 + \sigma_0^2 \gamma_0 a_0,$$

where a_0 is the unique positive solution to

$$\tau_0 \gamma_0 a_0^2 + (\tau_0 + 1 - \gamma_0)a_0 - 1 = 0.$$

Proof. We note that when $\Sigma_i = I$, the deterministic equivalent is scalar: $Q_i(\tau_i) = a_i I$ where a_i is the unique positive solution to

$$\tau_i \gamma_i a_i^2 + (\tau_i + 1 - \gamma_i) a_i - 1 = 0.$$

We can substitute this in to the relevant quantities from Theorem 3.6 to see:

$$2 \langle Q_1(\tau_1)(I - \tau_0 Q_0(\tau_0))w_0, Q_1(\tau_1)w_1 \rangle_{\Sigma_1} = 2a_1^2(1 - \tau_0 a_0) \langle w_0, w_1 \rangle \text{ and}$$

$$||Q_1(\tau_1)(I - \tau_0 Q_0(\tau_0))w_0||_{\Sigma_1}^2 = a_1^2 (1 - \tau_0 a_0)^2 ||w_0||^2.$$

We now examine

$$\begin{aligned} t(\tau_0, \tau_1) &= \lim_{p \to \infty} p^{-1} Tr \left(a_1^2 I (a_0 I - \tau_0 a_0^2 I) \right) \\ &= a_1^2 Tr \left(a_0 I - \tau_0 a_0^2 I \right) / p \\ &= a_1^2 (a_0 - \tau_0 a_0^2) \\ &= a_1^2 a_0 (1 - \tau_0 a_0). \end{aligned}$$

Substituting into Theorem 3.6 and canceling the common factor a_1^2, we obtain

$$2(1 - \tau_0 a_0) \langle w_0, w_1 \rangle > (1 - \tau_0 a_0)^2 ||w_0||^2 + \sigma_0^2 \gamma_0 a_0 (1 - \tau_0 a_0).$$

Dividing by $(1 - \tau_0 a_0) > 0$ yields the stated condition. □

Theorem 3.8 In the asymptotic setting with isotropic data and fixed source model overparameterization γ_0, for any fixed normalized alignment $\rho = \langle w_0, w_1 \rangle / ||w_0||^2$, there exists a unique source ridge penalty τ_0^* that maximizes the transfer benefit $\Delta R = R^S - R^{TL}$. Additionally, outside of a measure-zero set of parameters, the optimal τ_0^* for transfer learning differs from the optimal ridge penalty for Task 0 performance.

Proof. We seek to maximize the asymptotic risk benefit $\Delta R(\tau_0) = R^S - R^{TL}(\tau_0)$. Since R^S is independent of τ_0, this is equivalent to minimizing $R^{TL}(\tau_0)$. Using the isotropic deterministic equivalents derived in Corollary 3.7, the transfer risk is proportional to:

$$J(\tau_0) = (1 - \tau_0 a_0)^2 ||w_0||^2 + \sigma_0^2 \gamma_0 a_0 (1 - \tau_0 a_0) - 2(1 - \tau_0 a_0) \langle w_0, w_1 \rangle,$$

where a_0 is implicitly defined by τ_0. Let $x(\tau_0) = 1 - \tau_0 a_0$. From the fixed-point equation $\tau_0 \gamma_0 a_0^2 + (\tau_0 + 1 - \gamma_0) a_0 - 1 = 0$, we can derive the bijection $x = \frac{a_0}{1 + \gamma_0 a_0}$. Note that a_0 is strictly decreasing in τ_0 (from ∞ to 0), and x is strictly increasing in a_0 (mapping $(0, \infty)$ to $(0, \gamma_0^{-1})$). Thus, optimizing with respect to τ_0 is equivalent to optimizing with respect to the shrinkage factor x. Substituting $a_0 = \frac{x}{1 - \gamma_0 x}$ into the objective:

$$J(x) = x^2 ||w_0||^2 + \frac{\sigma_0^2 \gamma_0 x^2}{1 - \gamma_0 x} - 2x \langle w_0, w_1 \rangle.$$

The derivative with respect to x is:

$$J'(x) = 2x||w_0||^2 + \sigma_0^2 \gamma_0 \frac{2x(1 - \gamma_0 x) + \gamma_0 x^2}{(1 - \gamma_0 x)^2} - 2 \langle w_0, w_1 \rangle = 2x||w_0||^2 + \sigma_0^2 \gamma_0 \frac{2x - \gamma_0 x^2}{(1 - \gamma_0 x)^2} - 2 \langle w_0, w_1 \rangle.$$

Let $G(x) = x||w_0||^2 + \frac{\sigma_0^2 \gamma_0}{2} \frac{x(2 - \gamma_0 x)}{(1 - \gamma_0 x)^2}$. For $x \in (0, \gamma_0^{-1})$, $G(x)$ is strictly increasing. The optimal shrinkage x^* is the unique solution to $G(x^*) = \langle w_0, w_1 \rangle$. Define the source-optimal shrinkage x_S^* as the minimizer of the source risk (which corresponds to setting $w_1 = w_0$ in the transfer risk). Thus x_S^* satisfies $G(x_S^*) = ||w_0||^2$. Comparing the two conditions:

$$G(x^*) = \langle w_0, w_1 \rangle \quad \text{and} \quad G(x_S^*) = ||w_0||^2.$$

Since G is strictly increasing/injective, $x^* = x_S^*$ if and only if $\langle w_0, w_1 \rangle = ||w_0||^2$. Thus, outside the measure-zero set where alignment exactly equals source signal power, $x^* \neq x_S^*$, implying $\tau_0^* \neq \tau_S^*$. □

Corollary 3.9 Let τ_0^* be the transfer-optimal regularization penalty and τ_S^* be the source-optimal regularization penalty.

- If the tasks are imperfectly aligned ($0 < \rho < 1$), then $\tau_0^* > \tau_S^*$ (transfer requires stronger regularization).
- If the tasks are super-aligned ($\rho > 1$), then $\tau_0^* < \tau_S^*$ (transfer requires weaker regularization).

This phase transition depends solely on task alignment and holds for all noise levels $\sigma_0 > 0$.

Proof. Recall from the proof of Theorem 3.8 that the optimal shrinkage factors x^* (transfer) and x_S^* (source) satisfy:

$$G(x^*) = \langle w_0, w_1 \rangle \quad \text{and} \quad G(x_S^*) = ||w_0||^2,$$

where $G(x)$ is a strictly increasing function. Recall also that $x(\tau_0) = 1 - \tau_0 a_0$ is a strictly decreasing function of the regularization τ_0.

- Case 1: Imperfect alignment ($\langle w_0, w_1 \rangle < ||w_0||^2$). Then $G(x^*) < G(x_S^*)$, which implies $x^* < x_S^*$. Since x decreases with τ_0, this implies $\tau_0^* > \tau_S^*$.
- Case 2: Super-alignment ($\langle w_0, w_1 \rangle > ||w_0||^2$). Then $G(x^*) > G(x_S^*)$, which implies $x^* > x_S^*$. Since x decreases with τ_0, this implies $\tau_0^* < \tau_S^*$.

□