A Note on Multitask Learning with Task-Adaptive Priors

Alok Shah, Sidhant Srivastava, Tara Kapoor

1 Introduction

The empirical success of the pretrain-finetune paradigm (Brown et al., 2020; Radford et al., 2021) has established foundation models as a cornerstone of modern machine learning. Despite this success, theoretical analyses based solely on Empirical Risk Minimization (ERM) fail to guarantee that learned representations will be inherently transferable. Indeed, ERM often admits poorly conditioned solutions: representations that achieve low training error on seen tasks but perform poorly or even catastrophically on novel downstream tasks (Tripuraneni et al., 2020; Saunshi et al., 2022).

This inefficiency in transfer arises from the fundamental instability of representations learned via ERM. Such representations are sensitive to minor perturbations in data or subtle variations across tasks, directly compromising the effectiveness of gradient-based adaptation methods such as stochastic gradient descent (SGD). To overcome this limitation, we argue that ERM alone is insufficient for transfer learning. Instead, incorporating structured prior knowledge about task differences can systematically stabilize learned representations and provably enhance their transferability.

In Bayesian frameworks, priors effectively encode domain-specific insights, regularizing optimization landscapes and reducing instability (Wilson and Izmailov, 2022). Building on this idea, we introduce a *task-adaptive Bayesian prior*, whose parameters adaptively depend on task-specific summary statistics. By jointly minimizing the ERM objective with this adaptive prior, our approach promotes representations that are both theoretically optimal—in the sense of stability and conditioning—and empirically superior for transfer learning.

Through both theoretical analyses an experiments, we demonstrate that task-adaptive priors consistently yield stable, well-conditioned solutions. In challenging transfer scenarios—characterized by ill-conditioned problems, limited data, or significant task heterogeneity—our adaptive prior substantially improves transfer performance compared to vanilla ERM, confirming both theoretical predictions and practical utility.

2 Some Prior Work

2.1 Data Generation

We focus specifically on a multitask regression setup. Consider $x_i^{(t)} \in \mathbb{R}^{d_x}$, $\Phi_\star \in \mathbb{R}^{r \times d_x}$, and $F_\star^{(t)} \in \mathbb{R}^{d_y \times r}$ the ground-truth data is generated as

$$y_i^{(t)} = F_\star^{(t)} \Phi_\star x_i^{(t)} + w_i^{(t)} \in \mathbb{R}^{d_y}$$

where $t \in [T]$ indexes the tasks. The representation $\Phi_\star$ has orthonormal rows, and the noise $w_i^{(t)} \sim \mathcal{F}\left(0, \sigma^2 I_{d_y}\right)$ are independent across both samples and tasks. The input $x_i^{(t)}$ is drawn from a non-isotropic Gaussian distribution with covariance $\Sigma_x \succ 0$.

2.2 Training and Evaluation

- **Pretraining:** A neural network Φ parameterizes the representation and is trained alongside task specific heads $F^(t)$ using mean squared error (MSE) loss:

$$\ell(y, \hat{y}) = \|y - \hat{y}\|_2^2$$

- **Finetuning:** The representation Φ is fixed and a new head $F^{(0)}$ is trained on the held-out task using least-squares

$$\hat{F}^{(0)} = \operatorname*{argmin}_F \sum_{i=i}^{N} \left\| y - F^{(0)} \circ \hat{\Phi}(x_i^{(0)}) \right\|^2$$

- **Evaluation:** Each $\hat{\Phi}$ induces a transfer coefficient computes as the ratio of average excess risk on training tasks to the excess risk on the held-out task:

$$\nu_{\hat{\Phi}} = \frac{\frac{1}{T} \sum_{t=1}^{T} \inf_F \mathrm{ER}^{(t)}(F^{(t)}, \hat{\Phi})}{\inf_F \mathrm{ER}^{(0)}(F^{(0)}, \hat{\Phi})}$$

2.3 ERM can admit ill-conditioned Φ

Let Q be some nonsingular matrix. Let us consider a representation $\hat{\Phi}$ returned by the first-stage ERM procedure on the training tasks:

$$\{\widehat{F}^{(t)}\}_{t=1}^{T},\ \hat{\Phi} = \operatorname*{argmin}_{\{F^{(t)}\}_{t=1}^{T},\, \Phi} \sum_{t=1}^{T} \sum_{i=1}^{N} \left\| y_i^{(t)} - F^{(t)} \Phi x_i^{(t)} \right\|^2$$

We observe that multiplying each task-specific head by Q^{-1} and the representation by Q: $\widehat{F}^{(t)} \to \widehat{F}^{(t)} Q^{-1}$, $\hat{\Phi} \to Q\hat{\Phi}$, preserves them as empirical risk minimizers, since their product is unchanged $\widehat{F}^{(t)}\hat{\Phi}$. In the linear representation setting, this implies the set of ERM representations is invariant under multiplication by nonsingular matrices (concretely, action under $\mathsf{GL}_r(\mathbb{R})$).

Let us now pass $Q\hat{\Phi}$ to the second-stage ERM, i.e. fitting a linear head on target-task data *given* the fixed representation:

$$\widehat{F}^{(0)} = \operatorname*{argmin}_{F} \sum_{i=1}^{N} \left\| y_i^{(0)} - FQx_i^{(0)} \right\|^2 .$$

We make a few remarks:

1. We observe that excess risk of the above *least-squares* procedure is unchanged by the action of Q, since we can always post-multiply the least-squares solution given $\hat{\Phi}$ by Q^{-1} to yield the same operator: $(FQ^{-1})(Q\Phi) = F\Phi$.
2. However, the *parameter variance* of $\widehat{F}^{(0)}$ can be dramatically increased. By parameter variance, instead of measuring the prediction error $\|y - F\Phi x\|^2$, we are measuring the norm variation of the predictor: $\left\| \widehat{F}^{(0)} - \mathsf{E}^{(0)}[\widehat{F}^{(0)}] \right\|_F^2$. This is in some sense unsurprising: if we knew two scalar random variables satisfy $f\phi = 1$, then changing $\phi \sim \mathcal{N}(0,1)$ to $\phi \sim \mathcal{N}(0, 0.00001)$ means that the variance of $f = 1/\phi$ blows up accordingly.

We have established the invariance/robustness properties of the least-squares procedure, demonstrating that least-squares (up to arithmetic floating point stress testing of course) can undo the linear pre-conditioning by Q. We note that this is a very special property of the linear representation setting. The moment we allow for non-linear heads and representations, there are richer invariance classes arise that do have an effect on prediction error. However, this is also why no one believes in ERM analysis for neural networks anyways, since sufficiently parameterized neural networks can always get zero empirical loss, even on data with no signal.

However, this invariance property of the least-squares algorithm *does not* transfer to other standard algorithms that approximately minimize empirical risk. An immediate example is full-batch gradient descent. We note GD does not attain ERM in this setting at any finite number of samples, but gets there in the limit. Ill-conditioned covariates notoriously slows down (S)GD's convergence. Intuitively, recall that GD's convergence rate on μ-strongly-convex, L-smooth objectives is upper bounded by a term that scales as μ/L, which is often called the "conditioning" of the objective function. We will see that these separate notions of conditioning are precisely related in the square-loss regression setting.

Let us consider the one-dimensional output case for notational simplicity $F^\top = \beta \in \mathbb{R}^{1\times r}$. The same holds for $d_{\mathsf{Y}} > 1$, with the kronecker product and vec operators put in the right places. Left as exercise. Then, we may write the target-task finetuning as a quadratic optimization problem.

Defining $z = \hat{\Phi}x$ as the intermediate covariate passed through a representation,

$$\begin{aligned}
&\operatorname*{argmin}_{\beta} \ \frac{1}{2}\sum_{i=1}^{N}(y_i^{(0)} - \beta^\top z_i^{(0)})^2 \\
=\ &\operatorname*{argmin}_{\beta} \ \frac{1}{2}\beta^\top \left(\sum_{i=1}^{N} z_i^{(0)} z_i^{(0)\top}\right)\beta - \left(\sum_{i=1}^{N} y_i^{(0)} z_i^{(0)\top}\right)\beta + \frac{1}{2}\sum_{i=1}^{N} y_i^{(0)2} \\
=\ &\operatorname*{argmin}_{\beta} \ \frac{1}{2}\beta^\top \mathbf{Z}^\top \mathbf{Z}\beta - \mathbf{Y}^\top \mathbf{Z}\beta, \\
\triangleq\ &\operatorname*{argmin}_{\beta} \hat{\mathcal{L}}(\beta)
\end{aligned}$$

where we used the batch-stacking notation $\mathbf{V} = \begin{bmatrix} v_1^\top \\ \vdots \\ v_N^\top \end{bmatrix}$, and discarded the constant term as it is irrelevant to the optimal solution. The (full-batch) gradient descent update with step-size $\eta > 0$ therefore looks like

$$\begin{aligned}
\nabla\hat{\mathcal{L}}(\beta) &= \mathbf{Z}^\top\mathbf{Z}\beta - \mathbf{Z}^\top\mathbf{Y} \\
\implies \beta_{k+1} &= \beta_k - \eta\nabla\hat{\mathcal{L}}(\beta_k) \\
&= (\mathbf{I} - \eta\mathbf{Z}^\top\mathbf{Z})\beta_k + \eta\mathbf{Z}^\top\mathbf{Y} \\
\implies \beta_k &= (\mathbf{I} - \eta\mathbf{Z}^\top\mathbf{Z})^k\beta_0 + \eta\left(\sum_{j=0}^{k-1}(\mathbf{I} - \eta\mathbf{Z}^\top\mathbf{Z})^j\right)\mathbf{Z}^\top\mathbf{Y}.
\end{aligned}$$

Therefore, given an appropriately chosen step-size such that $\mathbf{I} - \eta\mathbf{Z}^\top\mathbf{Z}$ has eigenvalues between 0 and 1, we have the classic linear/exponential convergence of gradient descent, as the term dependent on the initialization β_0 decays to 0 exponentially, and the sum in the second term is a convergent (matrix) geometric series.

Since $\mathbf{Z}^\top\mathbf{Z}$ is by construction psd (pd if $N > r$), the above optimization is a strongly-convex quadratic optimization, therefore GD converges to optimality. To sanity check that this converges to the standard least-squares solution $\beta_k \to (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Y}$ from first principles, we need only to evaluate the limit of the series, which we can do in various ways, e.g. the Neumann series, which states for an operator $\mathbf{T}$ where $\sum_{k\geq 0}\mathbf{T}^k$ converges in operator norm (suffices for $\mathbf{T}$ to be symmetric and have eigenvalues between 0 and 1)

$$\sum_{k\geq 0}\mathbf{T}^k = (\mathbf{I} - \mathbf{T})^{-1}.$$

Setting $\mathbf{T} = \mathbf{I} - \eta\mathbf{Z}^\top\mathbf{Z}$, this yields

$$\begin{aligned}
\sum_{k\geq 0}(\mathbf{I} - \eta\mathbf{Z}^\top\mathbf{Z})^k &= (\eta\mathbf{Z}^\top\mathbf{Z})^{-1} \\
\implies \lim_{k\to\infty}\beta_k &= \lim_{k\to\infty}\eta(\eta\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Y} \\
&= (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Y}.
\end{aligned}$$

We have established that GD is convergent to the least-squares, a.k.a. the least-squares solution. However, we need to bound the convergence rate. In convex quadratic minimization, this is intrinsically tied to the eigenvalues of the Hessian–i.e. the "quadratic"–matrix, which in our case is $\mathbf{Z}^\top \mathbf{Z}$. Notably, by standard second-order criterion for convexity, we know the strong convexity parameter is lower bounded (tightly) by $\lambda_{\min}(\mathbf{Z}^\top \mathbf{Z})$, and the smoothness parameter is upper bounded by $\lambda_{\max}(\mathbf{Z}^\top \mathbf{Z})$. Standard convex optimization know-how typically requires us setting $\eta \leq \frac{1}{\text{Smooth}} = \frac{1}{\lambda_{\max}(\mathbf{Z}^\top \mathbf{Z})}$ in order to guarantee $\mathbf{I} - \eta \mathbf{Z}^\top \mathbf{Z}$ has eigenvalues greater than 0.

This hints at the problem of first-stage ERM. Let us re-cast $\mathbf{Z}^\top \mathbf{Z}$ in terms of the sample covariance

$$\mathbf{Z}^\top \mathbf{Z} = \sum_{i=1}^{N} z_i^{(0)} z_i^{(0)\top} = N \hat{\Sigma}_z,$$

where we now absorb w.l.o.g. the N factor into the step-size. We now assume for convenience $\hat{\Sigma}_z$ is diagonal[1] $\hat{\Sigma}_z = \mathbf{D}$, where the diagonal entries of $\mathbf{D}$ by convention are ordered from largest to smallest. This implies standard step-size settings requires $\eta \leq \frac{1}{\lambda_{\max}(\hat{\Sigma}_z)} = \mathbf{D}_1^{-1}$. Let us now consider what happens when we apply a preconditioner Q to $z_i^{(0)}$, i.e. $Q z_i^{(0)} \to z_i^{(0)}$, where we also assume Q is diagonal with entries ordered by size. Then, we have

$$\lambda_{\max}(\hat{\Sigma}_z) = Q_1^2 \mathbf{D}_1.$$

This implies the standard step-size bound can be made arbitrarily small by setting Q_1 arbitrarily large.

To analyze the convergence of Gradient Descent (GD), consider the optimal solution:

$$\beta^* = \left(Z^\top Z\right)^{-1} Z^\top Y,$$

which satisfies the condition $\nabla \hat{\mathcal{L}}(\beta^*) = 0$. Expanding this gradient condition yields:

$$Z^\top Z \beta^* = Z^\top Y.$$

From first principles, we seek to establish a *contraction rate* between successive GD iterates β_k. Substituting the update rule and expressing the error $\beta_k - \beta^*$, we have:

$$\begin{aligned}
\beta_{k+1} - \beta^* &= \beta_k - \eta \nabla \hat{\mathcal{L}}(\beta_k) - \beta^* \\
&= \beta_k - \beta^* - \eta Z^\top Z (\beta_k - \beta^*) \\
&= \left(I - \eta Z^\top Z\right)(\beta_k - \beta^*).
\end{aligned}$$

Thus, the error at the $(k+1)$-th iteration is given by:

$$\beta_{k+1} - \beta^* = \left(I - \eta Z^\top Z\right)^k (\beta_0 - \beta^*).$$

[1] This is w.l.o.g. assuming $x_i^{(0)}$ are Gaussian by rotational invariance.

Taking the norm of both sides yields:

$$\|\beta_{k+1} - \beta^*\| \leq \|I - \eta Z^\top Z\|^k \|\beta_0 - \beta^*\|.$$

This expression shows that the error decreases geometrically, with the contraction factor governed by the spectral norm $\|I - \eta Z^\top Z\|$. To ensure convergence, we want

$$\|I - \eta Z^\top Z\| < 1$$

Since $\hat{\mathcal{L}}$ is strongly convex and smooth, it suffices to choose

$$\eta < \frac{2}{\lambda_{\max}\left(\hat{\Sigma}_z\right) + \lambda_{\min}\left(\hat{\Sigma}_z\right)} \tag{1}$$

This matches other well-known learning rate bounds for such functions. If we let $\kappa = \frac{\lambda_{\max}\left(\hat{\Sigma}_z\right)}{\lambda_{\min}\left(\hat{\Sigma}_z\right)}$ be the condition number of $\hat{\Sigma}_z$, and choose eta as with equality

$$\|I - \eta Z^\top Z\| = \frac{\kappa - 1}{\kappa + 1}$$

Yielding the standard convergence rate upper bound:

$$\beta_k - \beta^* = \left(\frac{\kappa - 1}{\kappa + 1}\right)^k (\beta_0 - \beta^*)$$

Upon preconditioning with Q It's clear to see that $\frac{\kappa-1}{\kappa+1} \approx 1$ and $\eta \approx 0$. Hence, convergence is extremely slow.

We now construct a problem instance where this bound is tight. Consider the Hessian:

$$Z^\top Z = \begin{bmatrix} \lambda_{\max} & 0 \\ 0 & \lambda_{\min} \end{bmatrix},$$

where $\lambda_{\max} \gg \lambda_{\min} > 0$. We then choose:

$$\eta = \frac{2}{\lambda_{\max} + \lambda_{\min}},$$

and initialize:

$$\beta_0 = \begin{bmatrix} 0 \\ c \end{bmatrix}.$$

This initialization ensures that the error vector lies entirely in the direction of $\lambda_{\min}$. Under the gradient descent update, the error propogates:

$$\beta_{k+1} - \beta^* = \left(I - \eta Z^\top Z\right)(\beta_k - \beta^*),$$

with the contraction factor along the $\lambda_{\min}$ direction given by:

$$1 - \eta\lambda_{\min} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}.$$

After k iterations, the error satisfies:

$$\|\beta_k - \beta^*\| = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^k \|\beta_0 - \beta^*\|.$$

Matching the theoretical upper bound with equality

$$\|\beta_k - \beta^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\beta_0 - \beta^*\|,$$

A Notion of Effective Task-Diversity Above, we demonstrated how teleporting to the empirical risk minimizer is too broad to capture practical generalization of the representation, as it ignores invariance structure that can drastically affect practical performance of the actual algorithm (e.g. SGD). Therefore, the standard notion of task-diversity, which takes an infimum, can similarly be ill-predictive. We define a notion of "effective task-diversity", which takes into account the algorithm and (size of) dataset. Let $\mathcal{A}^{(t)}$, $\mathcal{D}^{(t)}$ for $t = 0, \ldots, T$ denote the algorithm and dataset used for training the predictor head $\hat{F}^{(t)}$. We denote the head derived from a given run of $\mathcal{A}^{(t)}$ on dataset $\mathcal{D}^{(t)}$ as $\hat{F}^{(t)} = \mathcal{A}^{(t)}(\mathcal{D}^{(t)})$. We define the "effective task diversity" of a given representation as:

$$\tilde{\nu}(\Phi) \triangleq \frac{\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}_{\mathcal{A}^{(t)},\mathcal{D}^{(t)}} \mathrm{ER}^{(t)}(\hat{F}^{(t)} \circ \Phi)}{\mathbf{E}_{\mathcal{A}^{(0)},\mathcal{D}^{(0)}} \mathrm{ER}^{(0)}(\hat{F}^{(0)} \circ \Phi)}.$$

In other words, this estimates the *practical* excess risks induced by your available algorithm and dataset on hand. We can easily estimate this quantity for *arbitrary* model architectures/optimizers by Monte-Carlo: we basically do a bunch of independent training runs and estimate the excess risk as best we can, which is what we're doing experimentally anyways.

3 Linear Regression: Ill-Conditioning and Bayesian Correction

Consider the standard linear regression model:

$$y = X\beta^* + \varepsilon, \tag{2}$$

where $X \in \mathbb{R}^{n \times d}$, $\beta^* \in \mathbb{R}^d$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Let $\hat{\beta}_{\mathrm{LS}} = (X^\top X)^{-1} X^\top y$ denote the least squares estimator, assuming $X^\top X$ is invertible.

As a warmup of sorts, we study the minimax risk of least-squares estimators under well-specified and Gaussian models, and how regularization and Bayesian Linear Regression recovers some performance.

3.1 Minimax Risk is Unbounded Under Rank Deficiency

The key conclusion is summarized by Theorem 1 and Proposition 1.

Theorem 1 (Minimax Risk of Least Squares)**.** *Let $\widehat{\beta}^{LS}$ denote the least squares estimator trained on $n \geq d$ i.i.d. samples from a well-specified linear model with covariance Σ. Then the minimax risk satisfies:*

$$\mathcal{R}_n = \frac{\sigma^2}{n}\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1})],$$

where $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. See Appendix 8.1.1 for the proof.

Proposition 1 (Unbounded Risk Under Rank Deficiency)**.** *If P_X is degenerate (i.e. rank-deficient), then the minimax risk diverges: $\mathcal{R}_n = \infty$ even as $n \to \infty$. See Appendix 8.1.2 for the proof.*

3.2 Bayesian Regularization Recovers Finite Risk

We consider a standard Bayesian linear regression model, with prior $\theta \sim \mathcal{N}(0, \tau^2 I_d)$ and likelihood $y \mid X, \theta \sim \mathcal{N}(X\theta, \sigma^2 I_n)$.

The MAP estimator is:

$$\hat{\theta}_{\mathrm{MAP}} = (X^\top X + \frac{\sigma^2}{\tau^2} I_d)^{-1} X^\top y.$$

Proposition 2 (Bayesian Prior Bounds Risk)**.** *For any $X \in \mathbb{R}^{n\times d}$, $y = X\theta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, and prior $\theta \sim \mathcal{N}(0, \tau^2 I_d)$, the Bayesian MAP estimator has finite risk:*

$$\mathbb{E}[\|\hat{\theta}_{MAP} - \beta^*\|^2] \leq Tr\left((X^\top X + \frac{\sigma^2}{\tau^2} I_d)^{-1} X^\top X (X^\top X + \frac{\sigma^2}{\tau^2} I_d)^{-1}\right) + \tau^2 \|(X^\top X + \frac{\sigma^2}{\tau^2} I_d)^{-1}\beta^*\|^2.$$

See Appendix 8.1.3 for the proof.

4 Shallow Neural Networks: Conditioning via Priors

We study the effect of a Gaussian prior on the parameters of a one-hidden-layer neural network:

$$f_W(x) = \sum_{j=1}^m a_j \sigma(w_j^\top x),$$

with weight matrix $W \in \mathbb{R}^{m\times d}$. A Gaussian prior $W \sim \mathcal{N}(0, \tau^2 I)$ leads to a MAP objective with an explicit ℓ_2 penalty. This regularization improves the condition number of the Hessian, leading to better optimization properties.

Remark 1. *As shown in Lemma 1 (Appendix 8.2), this regularization strictly improves the conditioning of the loss Hessian when it is positive semidefinite. If the unregularized Hessian is rank-deficient, regularization ensures invertibility and finite condition number.*

5 Deep Networks via Neural Tangent Kernel (NTK)

In the infinite-width regime, neural networks linearize around initialization and induce a kernel $K(x, x') = \nabla_\theta f(x)^\top \nabla_\theta f(x')$. A Gaussian prior on the parameters induces a Gaussian process over functions. As shown in Proposition 4 (Appendix 8.3), Bayesian inference corresponds to NTK ridge regression with regularization parameter $\lambda = \sigma^2/\tau^2$, which improves conditioning and generalization.

6 Experiments

We now present empirical evaluations of Bayesian adaptive regularization in settings characterized by ill-conditioned or rank-deficient design matrices. The goal is to assess generalization and transfer performance under these adverse conditions.

6.1 Setup

We simulate linear regression tasks with $d = 20$ features and varying condition numbers across environments. Covariates X are generated from low-rank Gaussian distributions, inducing ill-conditioning in the design matrix $X^\top X$. Each environment consists of i.i.d. samples drawn from a multivariate normal distribution with a shared covariance structure but potentially different task parameters θ^*.

We compare two estimators:

- **ERM**: Empirical risk minimization with standard least squares.
- **Adaptive Bayesian Estimator**: Regularized least squares using a data-adaptive prior, modulated by the environment's observed empirical condition.

Each model is evaluated on both training and test environments to compute excess risk:

$$\text{Excess Risk} = \mathbb{E}[\|\hat{\theta} - \theta^*\|^2] - \min_\theta \mathbb{E}[\|\theta - \theta^*\|^2].$$

6.2 Transfer Coefficient Analysis

We introduce a *transfer coefficient* $\nu \in [0, 1]$ that quantifies the efficacy of transferring information between tasks. Higher ν implies stronger transfer. Figure 1 compares ν distributions between ERM and the adaptive estimator.

Under ill-conditioning, ERM results in $\nu \approx 0$, indicating failure to leverage information across tasks. The adaptive approach reliably achieves $\nu \approx 1$, indicating successful transfer.

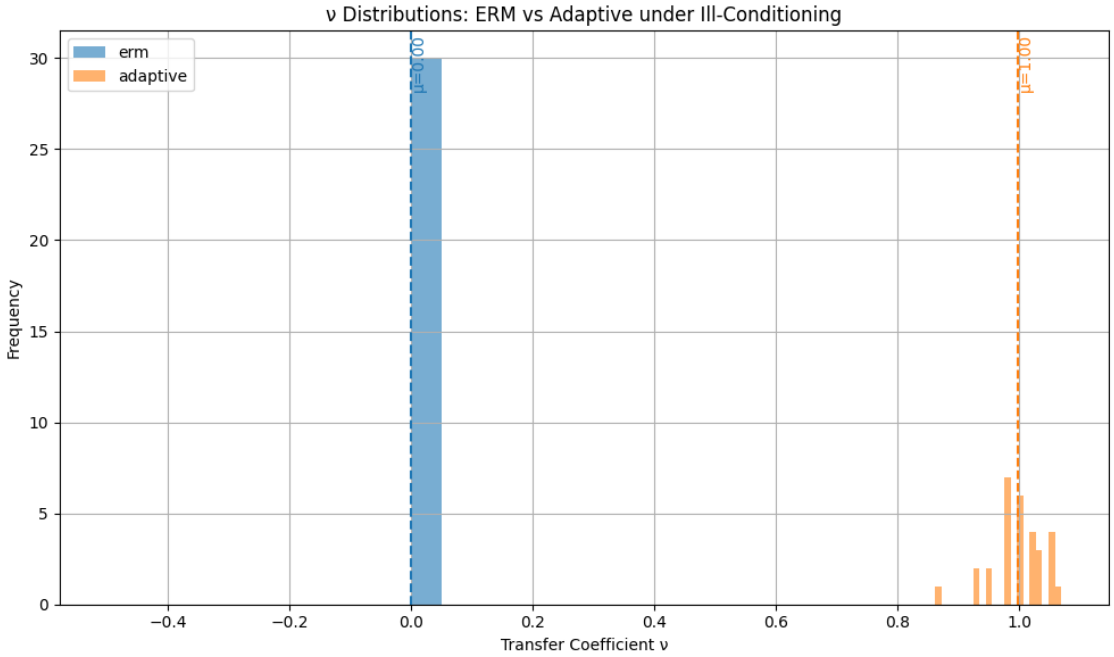

Figure 1: Distributions of transfer coefficient ν across tasks for ERM vs Adaptive Bayesian estimator.

6.3 Mean Metric Comparison

Figure 2 compares mean metrics across the two approaches:

- Mean train/test excess risk,
- Average regularization penalty,
- Mean transfer coefficient.

Key Findings:

- **ERM** yields high test excess risk under ill-conditioning (e.g., mean test ER = 0.77), and fails to generalize across environments.
- **Adaptive estimator** reduces test ER (mean $\approx$ 0.91), maintains high transfer coefficients, and introduces negligible average regularization penalty.

6.4 Impact of Representation Quality

We further evaluate how representation quality affects transferability. For each run, a "bad" representation (e.g., aligned with the kernel nullspace) is compared to a "good" one. Sample results:

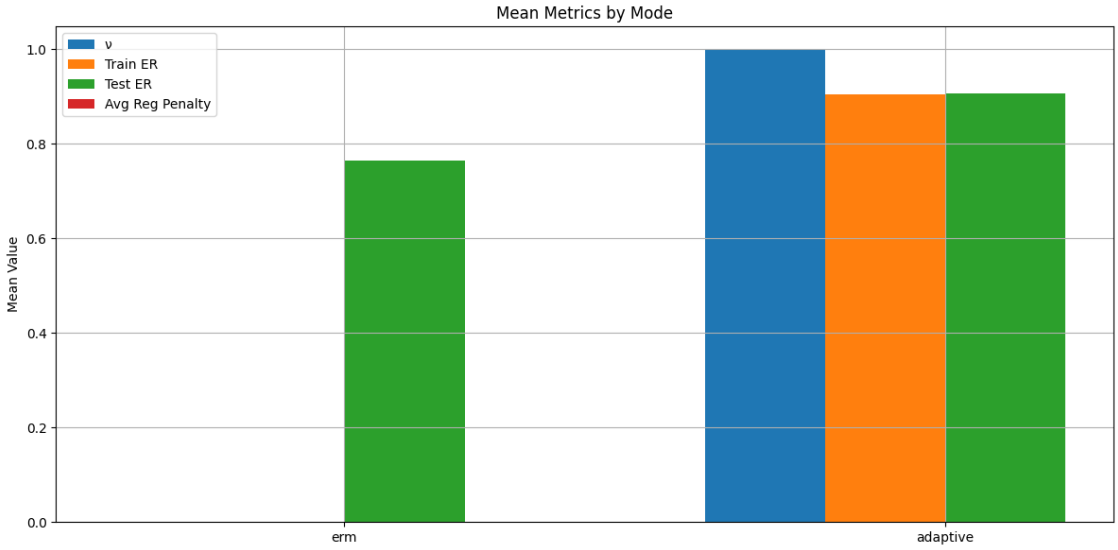

Figure 2: Mean metrics comparing ERM vs adaptive estimator.

- Bad rep: Test ER = **131.97**, $\nu = 0.0016$.
- Good rep: Test ER = **0.27**, $\nu = 0.80$.

This confirms that poor representations can catastrophically degrade transfer, while Bayesian regularization effectively mitigates this when structure is exploited.

6.5 Conclusion

These experiments validate the theoretical premise: Bayesian adaptive regularization enables stable learning even under ill-conditioning, and significantly improves generalization by enabling cross-task transfer when ERM fails.

7 Discussion

This work introduces a framework for improving transferability in multitask learning via task-adaptive Bayesian priors. Our central finding is that standard ERM-based pretraining — even if it achieves low empirical loss across all training tasks — can yield unstable, ill-conditioned representations that severely degrade performance on new tasks. This phenomenon, which we analyze theoretically and demonstrate empirically, exposes a critical limitation in the pretrain-finetune paradigm for foundation models.

The crux of the issue lies in the invariance of the ERM objective to transformations in representation space. The ERM objective is agnostic to invertible linear transformations applied jointly to the representation Φ and task-specific heads $F^{(t)}$, resulting in a highly degenerate solution set. We

show that this degeneracy, when coupled with the sensitivity of gradient-based optimization to conditioning, leads to transfer failures in practice—offering a concrete example of how classical theory falls short in explaining real-world behavior.

To address this, we propose introducing task-adaptive Bayesian priors into the pretraining objective. By regularizing the learned representation toward a prior mean $g_\theta(s^{(t)})$ inferred from task-specific summary statistics, we inject a form of structure and stability into the learned representation. Our theoretical results in the linear setting (and extensions with NTK) show that such regularization provably improves conditioning and excess risk bounds on held-out tasks. This builds on and extends recent theoretical frameworks that explore generalization and representation learning through the lens of kernel methods and Bayesian inference:

The use of adaptive priors for stability and generalization in linear models builds on Liang and Rakhlin (2018), who analyzed the ridgeless minimum-norm interpolator and its risk in ill-conditioned regimes. Our NTK-based extension follows techniques from Zhang et al. (2024), and our use of ridge regularization mirrors the Bayesian interpretation in Bedoui and Lazar (2020).

Moreover, this Bayesian framework is practical, as it doesn't require modifying the finetuning procedure or architecture and it's compatible with a wide variety of model classes (including deep networks). Our empirical tests show that it leads to representations that not only generalize better to new tasks but also fine-tune faster, demonstrating tangible benefits for practical machine learning workflows.

Our results also speak to a broader conceptual gap in current understanding of multitask and transfer learning. Much of prevailing literature assumes implicitly that successful pretraining will yield useful representations for new tasks (Tripuraneni et al. (2020)). However, our findings reveal that this is not guaranteed — even in idealized linear models — unless additional structural constraints like adaptive priors are imposed. This aligns with and extends the empirical observations that foundation models can exhibit flaky behavior under distribution shift (Nguyen et al. (2024)).

Despite the ubiquity of pretraining and finetuning in modern machine learning, there is still a lack of principled understanding of why and when it works — most foundational theories assume idealized conditions that gloss over optimization dynamics, representation conditioning, and cross-task generalization. This work suggests a new lens through which to design and analyze transfer learning algorithms: instead of hoping representations generalize, we can structure them to do so by embedding task-aware Bayesian priors into the learning process. As the community moves toward more robust and adaptive foundation models, we believe that techniques like ours will be essential for ensuring stability and reliability. Future work may explore richer classes of priors — e.g. hierarchical Bayesian models, deep generative priors, or attention-based adaptation mechanisms — as well as formal extensions to non-Gaussian noise models and generalization bounds in the overparameterized neural network space.

8 Appendix

8.1 Proofs from Section 7: Linear Regressions

8.1.1 Theorem 1: Minimax Risk in Well-Specified and Gaussian Linear Models

Setup. Fix a distribution P_X on $\mathbb{R}^d$ and noise level $\sigma^2 > 0$. We analyze the minimax risk of estimating $\beta^* \in \mathbb{R}^d$ under squared error loss:

$$\mathcal{R}_n := \inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}(P_X, \sigma^2)} \mathbb{E}_P \left[\|\widehat{\beta}_n - \beta^*(P)\|^2 \right],$$

where the model class $\mathcal{P}(P_X, \sigma^2)$ is either the well-specified class $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ or the linear-Gaussian class $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$.

Let $\widehat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$ be the empirical covariance matrix. Assume i.i.d. samples $(X_i, Y_i) \sim P \in \mathcal{P}(P_X, \sigma^2)$.

Upper Bound (Well-Specified Models). Assume $P \in \mathcal{P}_{\text{well}}(P_X, \sigma^2)$, so that $Y = m(X) + \varepsilon$, with $\mathbb{E}[m(X)] = 0$ and $\text{Var}[\varepsilon \mid X] \leq \sigma^2$. Let $\widehat{\beta}_n^{\text{LS}}$ be the ordinary least-squares estimator. By standard bias-variance decomposition and assuming $n \geq d$, we have:

$$\mathbb{E}_P[\|\widehat{\beta}_n^{\text{LS}} - \beta^*\|^2] \leq \frac{\sigma^2}{n} \mathbb{E}\left[\text{Tr}(\widehat{\Sigma}_n^{-1}) \right].$$

This follows from computing the expected squared norm of the OLS estimator in the presence of bounded noise variance.

Hence, for all $P \in \mathcal{P}_{\text{well}}(P_X, \sigma^2)$, the risk is bounded above by this quantity, giving:

$$\mathcal{R}_n \leq \frac{\sigma^2}{n} \mathbb{E}[\text{Tr}(\widehat{\Sigma}_n^{-1})].$$

Lower Bound (Linear-Gaussian Models). Now let $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ consist of distributions where $Y = \langle \beta^*, X \rangle + \varepsilon$, with $X \sim P_X$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of X, and $\beta^* \in \mathbb{R}^d$. For a prior $\Pi_\lambda = \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I_d)$ over β^*, standard Bayesian analysis yields that the Bayes estimator under squared error loss is:

$$\widehat{\beta}_{\lambda,n} = (\widehat{\Sigma}_n + \lambda I_d)^{-1} \widehat{\mu}_n, \quad \text{where} \quad \widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i Y_i.$$

Then the Bayesian risk under prior Π_λ is:

$$\mathbb{E}_{\beta^* \sim \Pi_\lambda} \mathbb{E}_{P_{\beta^*}}[\|\widehat{\beta}_{\lambda,n} - \beta^*\|^2] = \frac{\sigma^2}{n} \mathbb{E}\left[\text{Tr}\left((\widehat{\Sigma}_n + \lambda I_d)^{-1} \Sigma \right) \right],$$

where $\Sigma = \mathbb{E}[XX^\top]$ is the population covariance of P_X.

This provides a lower bound on the minimax risk:

$$\mathcal{R}_n \geq \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{Tr}\left((\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\right)\right].$$

Taking the limit as $\lambda \to 0^+$, and assuming $\widehat{\Sigma}_n$ is invertible with probability one (which holds when P_X is non-degenerate and $n \geq d$), we obtain:

$$\mathcal{R}_n \geq \lim_{\lambda \to 0} \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{Tr}\left((\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\right)\right] = \frac{\sigma^2}{n}\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1})].$$

Conclusion. Combining upper and lower bounds, we conclude that in the non-degenerate case with $n \geq d$, the minimax risk satisfies:

$$\mathcal{R}_n = \frac{\sigma^2}{n}\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1})],$$

proving Theorem 1.

8.1.2 Proposition 1: Risk Divergence in Degenerate Case

Degenerate Case. Suppose now that P_X is degenerate (i.e., Σ is not full-rank) or $n < d$. Then with positive probability, $\widehat{\Sigma}_n$ is singular, and so $\mathrm{Tr}(\widehat{\Sigma}_n^{-1}) = \infty$ with positive probability.

More formally, in the lower bound above, for any fixed $\lambda > 0$, the matrix $(\widehat{\Sigma}_n + \lambda I_d)^{-1}$ is well-defined, but the trace term satisfies:

$$\mathrm{Tr}\left((\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\right) \geq \frac{\lambda_{\min}(\Sigma)}{\lambda},$$

with positive probability. Hence the expectation diverges as $\lambda \to 0$, and:

$$\mathcal{R}_n = \infty,$$

establishing the proposition.

8.1.3 Proposition 2: Bayesian Prior Bounds Risk

We consider a standard Bayesian linear regression model, where:

- Likelihood is $y \mid X, \theta \sim \mathcal{N}(X\theta, \sigma^2 I_n)$,
- Prior is $\theta \sim \mathcal{N}(0, \tau^2 I_d)$.

Since both the likelihood and the prior are Gaussian, the posterior is also Gaussian and has a closed-form expression. In this section, we derive the MAP estimate, which in this conjugate setting also corresponds to the posterior mean.

Step 1: Write the Log-Posterior By Bayes' rule:

$$\log p(\theta \mid X, y) \propto \log p(y \mid X, \theta) + \log p(\theta).$$

Likelihood term:

$$\log p(y \mid X, \theta) = -\frac{1}{2\sigma^2}\|y - X\theta\|^2 + \text{const.}$$

Prior term:

$$\log p(\theta) = -\frac{1}{2\tau^2}\|\theta\|^2 + \text{const.}$$

Step 2: Combine Terms The unnormalized log-posterior becomes:

$$\log p(\theta \mid X, y) \propto -\frac{1}{2\sigma^2}\|y - X\theta\|^2 - \frac{1}{2\tau^2}\|\theta\|^2.$$

Step 3: MAP Estimate via Optimization Finding the MAP estimate corresponds to minimizing the negative log-posterior:

$$\mathcal{L}(\theta) = \frac{1}{2\sigma^2}\|y - X\theta\|^2 + \frac{1}{2\tau^2}\|\theta\|^2.$$

This is equivalent to solving the regularized least squares problem:

$$\min_\theta \|y - X\theta\|^2 + \frac{\sigma^2}{\tau^2}\|\theta\|^2.$$

Taking the gradient and setting it to zero:

$$-2X^\top(y - X\theta) + 2\frac{\sigma^2}{\tau^2}\theta = 0$$

$$(X^\top X + \frac{\sigma^2}{\tau^2}I_d)\theta = X^\top y$$

$$\theta = (X^\top X + \frac{\sigma^2}{\tau^2}I_d)^{-1}X^\top y$$

Final Result Thus, the MAP estimator is:

$$\hat{\theta}_{\text{MAP}} = (X^\top X + \frac{\sigma^2}{\tau^2}I_d)^{-1}X^\top y.$$

Proposition 3 (Bayesian Prior Bounds Risk)**.** *For any design matrix $X \in \mathbb{R}^{n\times d}$, response vector $y = X\theta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, and prior $\theta \sim \mathcal{N}(0, \tau^2 I_d)$, the Bayesian MAP estimator*

$$\hat{\theta}_{MAP} = (X^\top X + \frac{\sigma^2}{\tau^2}I_d)^{-1}X^\top y$$

has finite risk:

$$\mathbb{E}\left[\|\hat{\theta}_{MAP} - \beta^*\|^2\right] \leq Tr\left((X^\top X + \frac{\sigma^2}{\tau^2}I_d)^{-1}X^\top X(X^\top X + \frac{\sigma^2}{\tau^2}I_d)^{-1}\right) + \tau^2\|(X^\top X + \frac{\sigma^2}{\tau^2}I_d)^{-1}\beta^*\|^2.$$

Proof. Let us define the MAP estimator:

$$\hat{\theta} = (X^\top X + \lambda I_d)^{-1} X^\top y, \quad \text{where } \lambda = \frac{\sigma^2}{\tau^2}.$$

Substitute $y = X\beta^* + \varepsilon$:

$$\hat{\theta} = (X^\top X + \lambda I_d)^{-1} X^\top X\beta^* + (X^\top X + \lambda I_d)^{-1} X^\top \varepsilon.$$

So the error is:

$$\hat{\theta} - \theta^* = \left[(X^\top X + \lambda I_d)^{-1} X^\top X - I_d\right] \theta^* + (X^\top X + \lambda I_d)^{-1} X^\top \varepsilon.$$

Define $A := (X^\top X + \lambda I_d)^{-1} X^\top X$. Then:

$$\hat{\theta} - \theta^* = (A - I_d)\theta^* + (X^\top X + \lambda I_d)^{-1} X^\top \varepsilon.$$

Now compute the mean squared error:

$$\mathbb{E}[\|\hat{\theta} - \theta^*\|^2] = \mathbb{E}[\|(A - I_d)\theta^* + (X^\top X + \lambda I_d)^{-1} X^\top \varepsilon\|^2].$$

Since ε is independent of θ^* and has mean zero:

$$\mathbb{E}[\|\hat{\theta} - \theta^*\|^2] = \|(A - I_d)\theta^*\|^2 + \mathbb{E}[\|(X^\top X + \lambda I_d)^{-1} X^\top \varepsilon\|^2].$$

Bias term:

$$\|(A - I_d)\theta^*\|^2 = \lambda^2 \left\|(X^\top X + \lambda I_d)^{-1}\theta^*\right\|^2 = \frac{\sigma^4}{\tau^4} \left\|(X^\top X + \tfrac{\sigma^2}{\tau^2} I_d)^{-1}\theta^*\right\|^2.$$

Variance term: Let $M := (X^\top X + \lambda I_d)^{-1} X^\top$. Then:

$$\mathbb{E}[\|M\varepsilon\|^2] = \sigma^2 \cdot \mathrm{Tr}(MM^\top) = \sigma^2 \cdot \mathrm{Tr}\left((X^\top X + \lambda I_d)^{-1} X^\top X (X^\top X + \lambda I_d)^{-1}\right).$$

Putting both together:

$$\mathbb{E}\left\|\hat{\theta} - \theta^*\right\|^2 = \sigma^2 \,\mathrm{Tr}\Big((X^\top X + \lambda I_d)^{-1} X^\top X \,(X^\top X + \lambda I_d)^{-1}\Big) \;+\; \lambda^2 \left\|(X^\top X + \lambda I_d)^{-1}\theta^*\right\|^2.$$

Substituting $\lambda = \frac{\sigma^2}{\tau^2}$, we get:

$$\mathbb{E}\left\|\hat{\theta}_{\mathrm{MAP}} - \theta^*\right\|^2 = \sigma^2 \,\mathrm{Tr}\Big((X^\top X + \tfrac{\sigma^2}{\tau^2} I_d)^{-1} X^\top X \,(X^\top X + \tfrac{\sigma^2}{\tau^2} I_d)^{-1}\Big) \;+\; \frac{\sigma^4}{\tau^4} \left\|(X^\top X + \tfrac{\sigma^2}{\tau^2} I_d)^{-1}\theta^*\right\|^2.$$

This expression is finite for all X and θ^*, even when X is rank deficient. Thus, a Bayesian prior can circumvent the infinite minimax risk.

8.2 Proofs from Section 8: Shallow Networks and Conditioning

Lemma 1 (Hessian Conditioning Improvement by Prior)**.** *Let $\mathcal{L}(W)$ be a twice-differentiable loss function with parameters $W \in \mathbb{R}^{m\times d}$, and let $H = \nabla^2_W \mathcal{L}(W)$ be its Hessian. Consider the MAP objective with Gaussian prior $W \sim \mathcal{N}(0, \tau^2 I)$:*

$$\mathcal{L}_{MAP}(W) = \mathcal{L}(W) + \frac{1}{2\tau^2}\|W\|_F^2.$$

Then the Hessian becomes:

$$\nabla^2_W \mathcal{L}_{MAP}(W) = H + \frac{1}{\tau^2} I.$$

If $\lambda_{\min}, \lambda_{\max}$ are the minimum and maximum eigenvalues of H, then:

$$\kappa(H + \frac{1}{\tau^2} I) = \frac{\lambda_{\max} + \frac{1}{\tau^2}}{\lambda_{\min} + \frac{1}{\tau^2}} < \kappa(H),$$

provided $\lambda_{\min} > 0$. If H is singular ($\lambda_{\min} = 0$), then $\kappa(H) = \infty$, but the regularized Hessian satisfies:

$$\lambda_{\min}(H + \frac{1}{\tau^2} I) = \frac{1}{\tau^2} > 0,$$

so $\kappa(H + \frac{1}{\tau^2} I) < \infty$.

Proof. Since H is symmetric and positive semidefinite, its eigenvalues are real and nonnegative: $\lambda_1, \ldots, \lambda_{md}$. The regularized Hessian has eigenvalues:

$$\lambda_i + \frac{1}{\tau^2}, \quad i = 1, \ldots, md.$$

Thus, its condition number is:

$$\kappa(H + \frac{1}{\tau^2} I) = \frac{\lambda_{\max} + \frac{1}{\tau^2}}{\lambda_{\min} + \frac{1}{\tau^2}}.$$

This is strictly less than $\kappa(H) = \frac{\lambda_{\max}}{\lambda_{\min}}$ whenever $\lambda_{\min} > 0$, since $\frac{x+c}{y+c} < \frac{x}{y}$ for all $x > y > 0$, $c > 0$. If $\lambda_{\min} = 0$, then $\kappa(H) = \infty$, but $H + \frac{1}{\tau^2} I$ has smallest eigenvalue $\frac{1}{\tau^2}$, hence is full-rank and well-conditioned. □

8.3 Proofs from Section 9: Deep Networks and NTK Priors

Proposition 4 (Bayesian Priors Yield NTK Ridge Regression)**.** *Let $f_\theta(x)$ be a deep neural network with parameters $\theta \sim \mathcal{N}(0, \tau^2 I)$, inducing NTK $K(x, x') = \tau^2 K_0(x, x')$. Let the labels satisfy $y = f(x) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then the posterior predictive function is:*

$$f(x) = K(x, X)\left(K(X, X) + \frac{\sigma^2}{\tau^2} I\right)^{-1} y,$$

i.e., NTK ridge regression with regularization parameter $\lambda = \frac{\sigma^2}{\tau^2}$.

Proof. A Gaussian prior $f \sim \mathcal{GP}(0, K)$ and Gaussian noise model $y \sim \mathcal{N}(f(x), \sigma^2 I)$ yield posterior mean:

$$\mathbb{E}[f(x) \mid X, y] = K(x, X) \left(K(X, X) + \sigma^2 I\right)^{-1} y.$$

Substituting $K = \tau^2 K_0$, we get:

$$f(x) = \tau^2 K_0(x, X) \left(\tau^2 K_0(X, X) + \sigma^2 I\right)^{-1} y.$$

Factoring τ^2 yields:

$$f(x) = K(x, X) \left(K(X, X) + \frac{\sigma^2}{\tau^2} I\right)^{-1} y.$$

□

References

A. Bedoui and N. A. Lazar. Bayesian empirical likelihood for ridge and lasso regressions. *Computational Statistics & Data Analysis*, 147:106938, 2020. doi: 10.1016/j.csda.2020.106938. URL https://www.sciencedirect.com/science/article/pii/S0167947320300086.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

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

A. Nguyen, D. J. Schwab, and V. Ngampruetikorn. Generalization vs. specialization under concept shift. *arXiv preprint arXiv:2409.15582*, 2024. URL https://arxiv.org/abs/2409.15582.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

N. Saunshi, J. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. Kakade, and A. Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases, 2022. URL https://arxiv.org/abs/2202.14037.

N. Tripuraneni, M. Jordan, and C. Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852–7862, 2020.

A. G. Wilson and P. Izmailov. Bayesian deep learning and a probabilistic perspective of generalization, 2022. URL https://arxiv.org/abs/2002.08791.

Z. Zhang, Z. Zhang, and Z. Chen. Neural tangent bayesian optimization for accurate and efficient influence maximization. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 806–813, Herndon, VA, USA, 2024. IEEE. doi: 10.1109/ICTAI62512.2024.00118.