LEARNED SHARED SPACE FOR KERNEL-BASED META-LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Kernel methods are attractive for few-shot regression because squared-loss predictors can be computed in closed form from a small set of observations. Their performance depends on how well the kernel captures inherent structure across tasks, motivating the learning of a kernel representation through meta-learning. However, learning the kernel alone does not provide a learned initial function as in standard gradient-based meta-learning. We develop a kernel meta-learning method, based on inducing points, to learn both a common function space and initial function within it. The learned initial function is then transferred to each new task and adapted by functional gradient descent. We characterize the resulting finite adaptation trajectory to better understand the role of gradient steps during adaptation. We further derive excess transfer risk bounds that quantify the effects of approximation, initialization, and adaptation steps. Under the stated assumptions including boundedness, regularity and spectral decay condition of the kernels, we provide an early-stopping rule and the resulting risk bounds, and describe the trade-off between the approximation error of the learned common space along with the cost of learning the inducing points.

1 INTRODUCTION

Meta-learning uses experience across tasks to learn from a small amount of data and generalize well to unseen data. Among various meta-learning algorithms, gradient-based methods (Finn et al., 2017; Raghu et al., 2020) learn an initialization for quick task adaptation. Kernel methods can obtain squared-loss predictors through kernel ridge regression (KRR) or Gaussian process regression (GPR), instead of gradient-based updates. These methods exploit the kernel's advantage in the few-shot learning settings (for example, quick adaptation through closed-form solvers along with non-parametric nature of prediction). However, these methods depend heavily on the expressivity of the chosen kernel function. Kernel meta-learning addresses this by learning such kernel parameters and corresponding representation (Bertinetto et al., 2019; Patacchiola et al., 2020; Zhou et al., 2021; Falk et al., 2022). We study how to learn a function space and a shared initial function within it, and how that starting point affects a finite sequence of functional gradient updates.

To learn a good shared initial function across tasks, its kernel head must exhibit a common basis across tasks. In general, changing the support dataset changes the basis functions; transferring the head coefficients generally does not preserve the original meaning of the function. Moreover, the number of coefficients increase with the number of support dataset size (and the number of tasks), increasing storage requirements. To address this problem, we instead use inducing points over a latent space to provide a common kernel basis (Quiñonero-Candela & Rasmussen, 2005; Snelson & Ghahramani, 2005; Titsias, 2009). This induced space makes the shared initial function well defined and transferable across tasks.

We initialize the inducing points, feature extractor, and the kernel parameters with a sparse-variational objective (Titsias, 2009). Given the resulting induced space, we minimize a quadratic objective to obtain a kernel head, producing a shared initial function over the induced space. We then refine the kernel representation and inducing points using query losses, through the standard gradient based updates. At test time, the representation is fixed and the function is adapted through functional gradient descent.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our contributions are summarized as follows.

1. We develop a kernel meta-learning method that learns inducing points, which define an induced space, and a shared initial function in this space for task adaptation. Inducing points define a kernel basis, making the shared initial function transferable across tasks, while query losses refine the representation used by both initialization and adaptation.
2. We characterize the adaptation trajectory and establish support-conditional and uniform excess transfer risk bounds. These bounds distinguish approximation and initialization errors from the effects of residuals and label noise. Under the stated assumptions, we derive early-stopping rules and balanced risk rates that quantify the trade-off between the approximation accuracy of the induced space and the cost of learning it.

2 Related work

Gradient-based meta-learning. MAML learns an initialization for quick task adaptation, while ANIL concentrates adaptation in the task head (Finn et al., 2017; Raghu et al., 2020). Learned metrics, latent subspaces, and context variables further restrict the adaptable coordinates (Lee & Choi, 2018; Zintgraf et al., 2019; Rusu et al., 2019); representation-recovery theory studies when shared features are learned (Collins et al., 2022).

Closed-form and kernel meta-learning. R2-D2 learns representations by differentiating through a ridge-regression solution (Bertinetto et al., 2019). Deep Kernel Transfer learns deep kernels for Gaussian-process prediction across tasks (Patacchiola et al., 2020), while MetaVRF and iKML learn kernels through random-feature representations (Zhen et al., 2020; Falk et al., 2022). Dense Gaussian processes also learn inducing-variable representations for few-shot prediction (Wang et al., 2021). We learn a shared initial function within an induced space and analyze the finite-step adaptation via functional gradient updates.

Sparse Gaussian processes and inducing points. Variational sparse Gaussian processes treat inducing points as variational parameters and optimize them together with kernel hyperparameters through a marginal-likelihood lower bound (Titsias, 2009). We use this objective to pre-train the feature extractor, kernel parameters, and inducing points. In our formulation, the inducing points define a common finite-dimensional function space (which we call induced space) across tasks. We learn a shared initial function in this space and refine the representation using query losses.

Functional updates and early stopping. MetaFun learns iterative updates of a functional representation for task prediction (Xu et al., 2020). Early stopping acts as spectral regularization in least-squares and kernel regression (Yao et al., 2007; Raskutti et al., 2014; Ali et al., 2019). We study functional gradient descent from a meta-learned initial function, characterize its adaptation trajectory, and derive early-stopping rules from excess transfer risk bounds.

3 Method

The method learns a common set of kernel basis functions and its corresponding spanned space called the induced space, and an initial combination of them, called the shared initial function. To this end, we first define the inducing points and its induced space, and then describe how the representation, i.e., the kernel basis functions, and the head of the shared initial function are trained.

3.1 RKHS preliminaries and shared bases via inducing points

Let $S \sim \mathcal{P}_{\text{task}}$ index a regression task with regression function $f_S^*(x) = \mathbb{E}[Y \mid X = x, S]$. We write $P_{s|S}$ and $P_{q|S}$ for its support and query input distributions. Each meta-training task has a support dataset $D_m^s = \{(X_{m,i}^s, Y_{m,i}^s)\}$ and a query dataset $D_m^q = \{(X_{m,i}^q, Y_{m,i}^q)\}$ of size n each, respectively. For the analysis, we use a common sample size n for the support and query sets; the experimental sample sizes are specified in Appendix F. Conditional on the task, observations within each set are independent and identically distributed, and the two sets are independent. Labels satisfy $Y = f_S^*(X) + \epsilon$ with $\mathbb{E}[\epsilon \mid X, S] = 0$. There are M_0 pre-training tasks and M meta-training tasks.

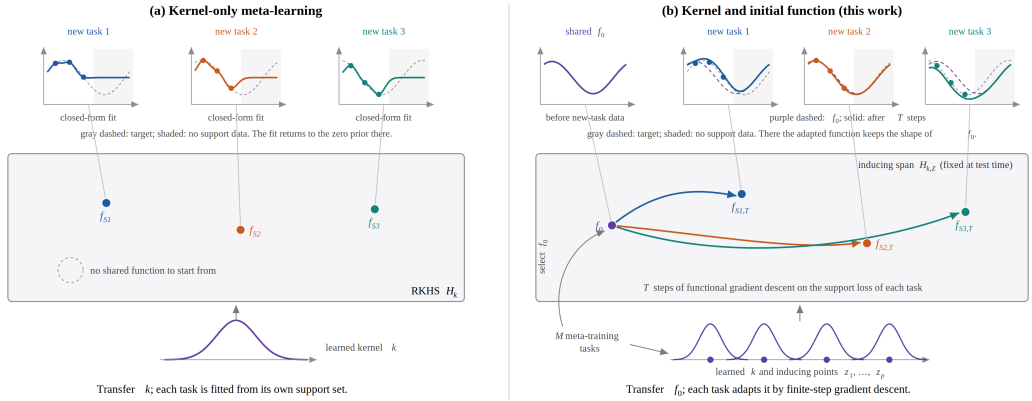

Figure 1: Kernel meta-learning vs Our method.

Let $\mathcal{D}_{\mathrm{VI}}$ and $\mathcal{D}_{\mathrm{meta}}$ denote the pre-training and meta-training datasets, respectively. A test task and its observations are independent of these training data.

Let $h_\psi : \mathcal{X} \to \mathcal{U}$ be a feature extractor mapping inputs to the latent space $\mathcal{U}$ and $k = k_\theta$ a positive-definite kernel on $\mathcal{U}$. For inducing points $Z = (z_1, \ldots, z_p) \in \mathcal{U}^p$, assume $K_{ZZ} = [k(z_i, z_j)]_{i,j=1}^p \succ 0$. Define

$$\mathcal{H}_{k,Z} = \mathrm{span}\{k(z_j, \cdot) : 1 \le j \le p\}, \qquad \vartheta = (\psi, Z, \theta). \tag{1}$$

The whitened features and predictor are

$$\phi_\vartheta(x) = K_{ZZ}^{-1/2} k(Z, h_\psi(x)), \qquad f_{\vartheta,\beta}(x) = \phi_\vartheta(x)^\top \beta, \quad \beta \in \mathbb{R}^p. \tag{2}$$

The corresponding latent-space function is $k(\cdot, Z)K_{ZZ}^{-1/2}\beta$, whose RKHS norm is $\|\beta\|_2$. Thus Euclidean gradient descent in β is the RKHS gradient update restricted to $\mathcal{H}_{k,Z}$. When the representation is understood, kernel expressions involving an input x include the composition with h_ψ.

For $a \in \{s, q\}$, let Φ_m^a have rows $\phi_\vartheta(X_{m,i}^a)^\top$ and let Y_m^a be the label vector. The empirical loss is

$$\widehat{L}_{D_m^a}(f_{\vartheta,\beta}) = \frac{1}{2n}\|\Phi_m^a \beta - Y_m^a\|_2^2. \tag{3}$$

We use the support and query operators

$$\begin{aligned} H_m &= \frac{(\Phi_m^s)^\top \Phi_m^s}{n}, \quad b_m = \frac{(\Phi_m^s)^\top Y_m^s}{n}, \\ G_m &= \frac{(\Phi_m^q)^\top \Phi_m^q}{n}, \quad g_m = \frac{(\Phi_m^q)^\top Y_m^q}{n}. \end{aligned} \tag{4}$$

At fixed ϑ, using $\nabla_\beta \widehat{L}_{D_m^a}$ adaptation from β takes the form

$$\beta_{m,0} = \beta, \qquad \beta_{m,t+1} = (I - \tau H_m)\beta_{m,t} + \tau b_m. \tag{5}$$

We use E steps during meta-training and T steps at test time. The step size is denoted by τ in both cases. After E steps, the adaptation map is given by

$$\begin{aligned} &\mathcal{A}_{m,E}^\vartheta(\beta) = A_{m,E}\beta + d_{m,E}, \\ &A_{m,E} = (I - \tau H_m)^E, \qquad d_{m,E} = \tau \sum_{t=0}^{E-1} (I - \tau H_m)^t b_m. \end{aligned} \tag{6}$$

3.2 Variational pre-training and shared initialization

Variational pre-training learns a space for representing the information across tasks. It provides an initial representation using task data and the kernel approximation error. Then following the shared initialization criterion, we choose the shared initial function from which adaptation begins.

For pre-training task m, define the Nyström matrix $Q_m = K_{X_m Z} K_{ZZ}^{-1} K_{Z X_m}$. The sparse variational objective is the average negative evidence lower bound

$$-\frac{1}{M_0}\sum_{m=1}^{M_0}\mathcal{L}_{\mathrm{VI},m}(\vartheta) = \frac{1}{M_0}\sum_{m=1}^{M_0}\left[-\log\mathcal{N}(Y_m; 0, Q_m + \sigma_0^2 I) + \frac{\mathrm{tr}(K_{X_m X_m} - Q_m)}{2\sigma_0^2}\right]. \tag{7}$$

Here σ_0^2 is the pre-training likelihood variance. This objective initializes ψ, Z, and θ. Its trace term measures the kernel approximation error on the observed inputs. Appendix B relates population projection residuals to this trace error and states the assumptions underlying the residual bound. Detailed description appears in Section 4.1.

For a fixed representation, we choose a kernel head using the quadratic criterion derived in Appendix C:

$$\begin{aligned}\widehat{\mathcal{E}}_{\mathrm{RKHS}}(\beta) = \frac{1}{2M}\sum_{m=1}^{M}\Big[&\|(I - A_{m,E})\beta - d_{m,E}\|_2^2 \\ &- \eta(G_m\beta - g_m)^\top K_{ZZ}(G_m\beta - g_m)\Big].\end{aligned} \tag{8}$$

The first term measures the squared RKHS displacement after E step of support updates. Its average is the training adaptation energy in Lemma 18, which enters the optimization bound of Section 4.3. The second term is a gradient term from query losses motivated by a first-order loss expansion. Together they define a shared initial function used for the starting point of the adaptation. Appendix C provides the motivation and computational details. The quadratic and linear coefficients of this objective are

$$\begin{aligned}P_{M,E} &= \frac{1}{M}\sum_{m=1}^{M}\left[(I - A_{m,E})^\top(I - A_{m,E}) - \eta G_m K_{ZZ} G_m\right], \\ q_{M,E} &= \frac{1}{M}\sum_{m=1}^{M}\left[(I - A_{m,E})d_{m,E} - \eta G_m K_{ZZ} g_m\right].\end{aligned} \tag{9}$$

In practice, we add the damping term $\gamma\beta^\top K_{ZZ}^{-1}\beta/2$ and choose $\gamma \geq 0$ so that $P_{M,E} + \gamma K_{ZZ}^{-1} \succ 0$. The resulting minimizer is

$$\beta^\star(\vartheta) = \left(P_{M,E} + \gamma K_{ZZ}^{-1}\right)^{-1} q_{M,E}. \tag{10}$$

The undamped case is obtained with $\gamma = 0$ when $P_{M,E}$ is positive definite. This notation keeps $P_{M,E}$ equal to the undamped matrix in Appendix C. The minimizer depends differentiably on ϑ wherever the coefficient matrix is nonsingular and both the matrix and $q_{M,E}$ are differentiable in ϑ.

3.3 Outer training and test-time adaptation

The outer objective evaluates query losses after support adaptation from $\beta^\star(\vartheta)$:

$$\mathcal{F}_{M,E}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M}\widehat{L}_{D_m^q}\left(f_{\vartheta,\mathcal{A}_{m,E}^{\vartheta}(\beta^\star(\vartheta))}\right). \tag{11}$$

We update ϑ by differentiating the outer objective via gradient descent. In practice, the averages in (9) and (11) are taken over each mini-batch. After outer training, we recompute the shared initialization using the meta-training tasks and set $\widehat{\beta}_0 = \beta^\star(\widehat{\vartheta})$.

On a fresh task, $\widehat{\vartheta}$ and $\widehat{\beta}_0$ are fixed. We apply (5) to its support dataset for T steps and use $f_{\widehat{\vartheta},\beta_{S,T}}$ for query prediction. The training and test-time operations are summarized in Algorithm 1.

4 Theory

In this section we provide the theoretical results of our algorithm. We first bound the approximation error of the induced space. We then study excess transfer risks for a fixed new task (support-conditional bound) and for an average over meta-training data and new tasks (uniform bound). These results describe the contributions of the representation, the shared initial function, and the number of adaptation steps. Full proofs are given in the Appendix.

Algorithm 1: Meta-learning a shared initial function and meta-testing.
Input: $\mathcal{D}_{\mathrm{VI}}$, $\mathcal{D}_{\mathrm{meta}}$; number of inducing points p; depths E, T; step size τ; query-gradient weight η; damping parameter γ chosen as in (10). **Output:** $\widehat{\vartheta}$, $\widehat{\beta}_0$, and a fresh-task predictor.

1. **Pre-training.** Initialize $\vartheta = (\psi, Z, \theta)$ and minimize (7) on $\mathcal{D}_{\mathrm{VI}}$. Construct the induced space.
2. **Meta-training.** Repeat the following steps over meta-batches.
   (a) **Shared initialization.** Compute the operators (4), the E-step maps (6), and the coefficients (9), using averages over the current meta-batch. Compute $\beta^\star(\vartheta)$ from (10).
   (b) **Representation update.** Adapt each task head to $\mathcal{A}^{\vartheta}_{m,E}(\beta^\star(\vartheta))$ and update ϑ using the gradient of (11).
3. **Final initialization.** Fix $\widehat{\vartheta}$; recompute (9) over the meta-training tasks and store $\widehat{\beta}_0 = \beta^\star(\widehat{\vartheta})$ from (10).
4. **Meta-test.** On a fresh task, initialize with $f_{\widehat{\vartheta},\widehat{\beta}_0}(x) = \phi_{\widehat{\vartheta}}(x)^\top \widehat{\beta}_0$. Apply (5) to its support dataset for T steps and return $f_{\widehat{\vartheta},\beta_{S,T}}$ for query prediction.

4.1 Approximation by the induced space

The inducing points determine the basis functions shared across tasks. Both the shared initial function and the adapted predictors are expressed in this basis. The first result quantifies its approximation error for a fixed feature map and kernel.

For inducing points $Z = (z_1, \cdots, z_p)$, write

$$g_Z(x) = k(x,x) - k(x,Z)K_{ZZ}^{-1}k(Z,x),$$
$$R(Z) = \mathbb{E}_S \mathbb{E}_{X\sim P_{s|S}}[g_Z(X)],$$
$$\widehat{R}_{M_0,n}(Z) = \frac{1}{M_0 n}\sum_{m=1}^{M_0} \mathrm{tr}\big(K_{X_m X_m} - Q_m\big). \tag{12}$$

Here X_m contains the n observations from pre-training task m. Let $\bar{f}_{S,Z}$ be a best approximation to f_S^* in $\mathcal{H}_{k,Z}$ under $L^2(P_{s|S})$, and set $r_{S,Z} = f_S^* - \bar{f}_{S,Z}$.

Proposition 1 (Approximation by the induced space)**.** *Fix h_ψ and k_θ independently of the pre-training samples. Assume the sampling and kernel regularity conditions of Theorem 8 in Appendix B, and suppose $\|f_S^*\|_{\mathcal{H}_k} \le B$ for almost every task. Assume also the spectral-decay of covariance operator Assumption 1 in Appendix B, $\lambda_j \le C_\alpha j^{-(1+\alpha)}$, and that $\mathcal{Z}_p$ contains a population minimizer of $R(Z)$. If*

$$\widehat{Z}_{\mathrm{tr}} \in \operatorname*{arg\,min}_{Z\in\mathcal{Z}_p} \widehat{R}_{M_0,n}(Z),$$

then, with probability at least $1-\delta$ over pre-training,

$$\mathbb{E}_S\left[\|r_{S,\widehat{Z}_{\mathrm{tr}}}\|^2_{L^2(P_{s|S})} \,\middle|\, \mathcal{F}_{\mathrm{VI}}\right] \le B^2\left[\frac{2^{\alpha+1}C_\alpha}{\alpha p^\alpha} + 2\Delta_{M_0,n}(\epsilon,\delta)\right]. \tag{13}$$

Here $\Delta_{M_0,n}(\epsilon,\delta)$ is the trace-estimation error with the rate

$$\Delta_{M_0,n}(\epsilon,\delta) = \tilde{O}\left(\sqrt{\frac{p\dim(\mathcal{U}) + \log(1/\delta)}{M_0}}\left(1 + \frac{1}{\sqrt{n}}\right) + \frac{1}{n}\right).$$

The rate follows from the covering bound and the choice of ϵ in Appendix B.

The residual $\|r_{S,\widehat{Z}_{\mathrm{tr}}}\|^2_{L^2(P_{s|S})}$ measures the kernel variation at the observed inputs that is misspecified by the induced space. Under the RKHS-norm bound, its population average controls an upper bound on the average task approximation error. Proposition 1 separates the approximation available with p inducing points from the error when the points are selected from finitely many tasks. It follows directly from the projection bound and minimization result in Appendix B.

The same trace residual $\hat{R}_{M_0,n}(Z)$ appears in the variational objective (7). With a theoretical assumption, we can apply its rate to the final learned representation (Section 4.4).

4.2 SUPPORT-CONDITIONAL EXCESS TRANSFER RISK

We next fix the learned representation and shared initial function and consider adaptation to a new task. Let $\mathcal{F}_{\text{train}}$ contain all training data and define a sigma field $\mathcal{G}_S = \sigma(S, X_S, \mathcal{F}_{\text{train}})$. Conditioning on $\mathcal{G}_S$ fixes the task, the learned quantities, and the new support inputs. The remaining randomness is in the support labels. We assume

$$\mathbb{E}[\varepsilon_S \mid \mathcal{G}_S] = 0, \qquad \mathbb{E}[\varepsilon_S \varepsilon_S^\top \mid \mathcal{G}_S] \preceq \sigma^2 I_n,$$

and independent query dataset given the task.

Write $H_S = \Phi_S^\top \Phi_S / n = U_S \operatorname{diag}(\lambda_{S,i}) U_S^\top$, let $I_S = \{i : \lambda_{S,i} > 0\}$, and let $\Pi_{S,0}$ be the orthogonal projector onto $\ker H_S$. For the query loss $L_S(f) = \frac{1}{2}\mathbb{E}_{(X,Y)\sim P_{q|S}}[(f(X) - Y)^2]$, choose the minimum-norm population least-squares solution

$$\bar{\beta}_S \in \arg\min_{\beta \in \mathbb{R}^p} \|f_{\widehat{\vartheta},\beta} - f_S^*\|^2_{L^2(P_{q|S})}, \quad r_S(x) := f_S^*(x) - f_{\widehat{\vartheta},\bar{\beta}_S}(x). \tag{14}$$

Define $e_{S,0} = U_S^\top(\widehat{\beta}_0 - \bar{\beta}_S)$, and write $r_S(X_S) \in \mathbb{R}^n$ for the residual evaluated at the support inputs. This projection uses the query distribution, which may differ from the support distribution used in Proposition 1.

Theorem 2 (Support-conditional excess transfer risk)**.** *Under the preceding assumptions, suppose* $k(x,x) \le 1$ *and* $0 \le \tau\lambda_{S,i} \le 1$ *for every* i*. For every integer* $T \ge 0$*, define*

$$\mathcal{R}_S^{sc}(T) = \mathbb{E}\Big[L_S(f_{\widehat{\vartheta},\beta_{S,T}}) - L_S(f_S^*) \,\Big|\, \mathcal{G}_S\Big]$$

where $\beta_{S,T}$ *denotes the task head obtained after* T *steps of functional gradient descent starting from* $\widehat{\beta}_0$*. Then*

$$\begin{aligned}\mathcal{R}_S^{sc}(T) \le \frac{1}{2}\|r_S\|^2_{L^2(P_{q|S})} + \frac{1}{2}\|\Pi_{S,0}(\widehat{\beta}_0 - \bar{\beta}_S)\|_2^2 + \sum_{i\in I_S}(1-\tau\lambda_{S,i})^{2T} e_{S,0}(i)^2 \\ + \frac{\|r_S(X_S)\|_2^2 + \sigma^2}{n}\sum_{i\in I_S}\frac{[1-(1-\tau\lambda_{S,i})^T]^2}{\lambda_{S,i}}.\end{aligned} \tag{15}$$

The first term measures what the induced space cannot represent under the query distribution. The second measures the mismatch between the shared initial function and the projected task function in directions that cannot be updated using support dataset. Note that both are independent of adaptation depth. The remaining initial error contracts along directions observed through the support datasets, while the final term increases as adaptation fits the task residual and label noise σ. The bound thus explains how the induced space and shared initial function affect the subsequent adaptation on a fresh task S. Theorem 14 in Appendix D provides the proof.

4.3 UNIFORM EXCESS TRANSFER RISK AND STOPPING

We now average over meta-training and a fresh task, conditional on $\mathcal{F}_{\text{VI}}$:

$$\begin{aligned}\mathcal{R}^u(T) &= \mathbb{E}\Big[L_S(f_{\widehat{\vartheta},\beta_{S,T}}) - L_S(f_S^*) \,\Big|\, \mathcal{F}_{\text{VI}}\Big], \\ \rho_p &= \mathbb{E}\Big[\|r_S\|^2_{L^2(P_{s|S})} \,\Big|\, \mathcal{F}_{\text{VI}}\Big].\end{aligned} \tag{16}$$

The risk also averages the fresh support sample. We assume matched support and query distributions and the boundedness, sampling, regularity condition over Hessian matrices and spectral decay conditions of Appendix E.

Theorem 3 (Uniform excess transfer risk)**.** *Under these conditions, for every integer* $T \ge 1$*,*

$$\mathcal{R}^u(T) \le \frac{A'}{\tau T} + 4(1+B_y^2)\sqrt{L_n}\left(\frac{\tau T \rho_p}{\sqrt{n}} + \frac{\tau^2 T^2 \sigma^2}{n^{3/2}}\right) + C_0. \tag{17}$$

where A'*,* L_n*, and* C_0 *are defined in Appendix E,* (21)*.*

Here A' depends on the training adaptation energy and an across-task term; C_0 collects the terms independent of T, including representation complexity. Thus incomplete adaptation decreases with T, while residual and noise contributions increase linearly and quadratically, respectively.

Corollary 4 (Bound-derived stopping rule)**.** *Under Theorem 3, suppose $A' > 0$ and set*

$$T_0 = \min\left\{ \frac{n^{1/4}}{\tau}\left[\frac{A'}{4(1+B_y^2)\sqrt{L_n}\rho_p}\right]^{1/2}, \frac{\sqrt{n}}{\tau}\left[\frac{A'}{4(1+B_y^2)\sqrt{L_n}\sigma^2}\right]^{1/3}\right\}, \qquad T^* = \max\{1, \lceil T_0 \rceil\}. \tag{18}$$

This balances the decreasing term against each increasing term. The resulting bound is given in (22) *in Appendix E.*

This rule involves population quantities, including ρ_p; the experiments instead use plug-in rules based on the conditional bound (15).

4.4 Dependence on inducing points and task numbers

Let $d = \dim(\mathcal{U})$, and let D_ψ, D_{hyp} denote the feature-extractor and kernel-parameter dimensions. We also assume spectral-decay condition with exponent $\alpha > 0$.

Corollary 5 (Risk bound at the selected stopping rule)**.** *Under the assumptions of Theorem 3 and Proposition 1, suppose the covering and growth conditions of Lemma 11 and the conditions of Appendix E hold. If the final representation satisfies the residual condition* (23) *in Appendix E, then for fixed $\delta \in (0,1)$, with probability at least $1-\delta$ over pre-training data,*

$$\mathcal{R}^u(T^*) \le \mathcal{A}^*_{\text{sub}} + \widetilde{O}\left(p^{-\alpha} + \sqrt{\frac{pd}{M_0}} + \sqrt{\frac{pd + D_\psi + D_{\text{hyp}}}{M}} + \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{M}}\right). \tag{19}$$

Here $\mathcal{A}^*_{\text{sub}}$ is the T-independent term defined in (20)*.*

The first term inside $\widetilde{O}$ describes approximation; the next two account for learning the inducing points and representation from M_0 and M tasks. The $p^{-\alpha}$ rate concerns the empirical trace minimizer and requires the stated residual assumptions to apply to the final learned representation. The notation $\widetilde{O}$ absorbs L_n, $\log p$, and covering logarithms, with fixed norm and noise bounds absorbed into constants. Appendix E gives the proof of the above rate. It also provides a statement over the choice of the size of p; with an ideal choice of p, the above excess transfer risk achieves the rate of $\mathcal{A}^*_{\text{sub}} + \tilde{O}\left(\frac{1}{\min\{M_0, M\}^{\frac{\alpha}{2\alpha+1}}} + \frac{1}{\sqrt{M}} + \frac{1}{\sqrt{n}}\right)$.

5 Experiments

We compare predictive performance with MAML, R2-D2, DKT, and iKML on five synthetic and three real world datasets, and examine pre-training, initialization, and adaptation depth. Appendix F provides the settings and additional results.

5.1 Setup

Synthetic datasets use inputs $x \sim U[-5, 5]$ and Gaussian label noise with standard deviation 0.1. *Sinusoidal* draws $A\sin(x+\phi)$ with random amplitude A and phase ϕ. *Omega* additionally adds frequency from $\omega \sim U[0.7, 1.3]$. *Misfit* adds a high-frequency component $0.5\sin(7x+\phi_2)$. *Oracle* draws from the span of 32 fixed RBF sections; this does not imply exact representation in the learned inducing space. *Structured Oracle* adds a shared nonzero mean and structured coefficient variation on Oracle dataset. The real world datasets are Beijing air quality (Zhang et al., 2017), METR-LA, and PEMS-BAY (Li et al., 2018), each split into disjoint task sets for pre-training, meta-training, validation and meta-test. We use $n = 10$ for synthetic tasks and $n = 5$ for real world tasks. Detailed settings are listed in Appendix F.

Table 1: Test query MSE averaged over three training runs. For our method, val(T^*) and bound(T^*) report query MSE at depths selected by validation MSE and the plug-in conditional bound (15), respectively. The two rules use separate draws of validation tasks. MAML uses $T = 10$. On the synthetic datasets, all methods are evaluated on the same 300 test tasks per run. On the real world datasets, each method is evaluated on 200 test tasks per run, and the test tasks of Ours are sampled independently of those of the baselines. Bold denotes the lowest reported mean. †iKML returned near-zero predictors in some runs; Table 4 reports every run. Detailed settings are in Table 5.

Dataset	Ours (val(T^*))	Ours (bound(T^*))	MAML	R2-D2	iKML	DKT
Synthetic ($n = 10$)						
Sinusoidal	**0.0355**	0.0439	0.3776	0.0464	2.8762†	0.1937
Omega	**0.1780**	0.1792	0.6115	0.2025	1.4132†	0.1892
Misfit	0.2236	0.3432	0.5088	**0.2158**	2.9532†	1.2080
Oracle	0.4297	0.4467	1.2006	0.4714	0.2983	**0.2959**
Structured	**0.3592**	0.3615	0.5140	0.6209	1.3312†	0.5055
Real ($n = 5$)						
Air quality	0.0503	0.0527	**0.0479**	0.0665	0.0896	0.1755
METR-LA	**0.3463**	0.3857	0.4078	0.4289	0.5516	0.6491
PEMS-BAY	0.0514	**0.0513**	0.0555	0.0534	0.0834	0.1656

5.2 Predictive evaluation

Table 1 reports test query MSE at global depths selected by validation MSE (val) or by the estimated conditional bound averaged over the validation tasks (bound) which is explained in detail in Appendix F.4. Ours and MAML use training depth $E = 10$; the MAML column uses test depth $T = 10$. R2-D2 and iKML use the implementations and settings specified in Appendix F.3. Tables 8 and 9 in Appendix F.4 report the bound-selected results and additional depth comparisons.

Under these settings, Ours has the lowest reported mean on METR-LA and PEMS-BAY, while MAML has the lowest mean on Air quality. Several iKML synthetic runs return near-zero predictors, which substantially affect its averages. The remaining runs in Table 4 achieve lower synthetic errors than the corresponding means of Ours. The comparison therefore reflects both prediction accuracy and optimization outcomes under the reported configurations.

5.3 Task-specific stopping using Conditional Excess-Risk Bound

We evaluate the plug-in stopping rule from (15) by estimating the quantities in the bound using 200 additional labeled observations per test task, and evaluate on a separate set of 200 query observations. Here the depth is selected separately for each test task, whereas Table 1 uses a global depth. Figure 2 reports the median per-task regret, defined as the query MSE at the selected depth divided by the minimum query MSE over the reference depths specified in Appendix F.7. Median ratios are 1.07, 1.13, 1.12, and 1.05 on Sinusoidal, Omega, Oracle, and Structured-Oracle, respectively, and 1.30 on Misfit. Appendix F.7 gives the full evaluation details and regret distributions.

5.4 Effects of pre-training and initialization

Figure 3 examines the effects of pre-training and shared head initialization at test-time. The left panel compares final models with pre-training (stage I) and random inducing points followed by the subsequent training stages, using validation-selected depths. The model with stage I gives lower query MSE in this comparison. The right panel fixes each learned kernel and inducing points and changes only the head (stage II) before test-time adaptation. The proposed shared head gives lower query MSE on the real world datasets. Appendices F.5 and F.6 give the full results with synthetic datasets as well as the real world datasets. These results identify the benefits of the initial function within the learned space that depend on the task distribution.

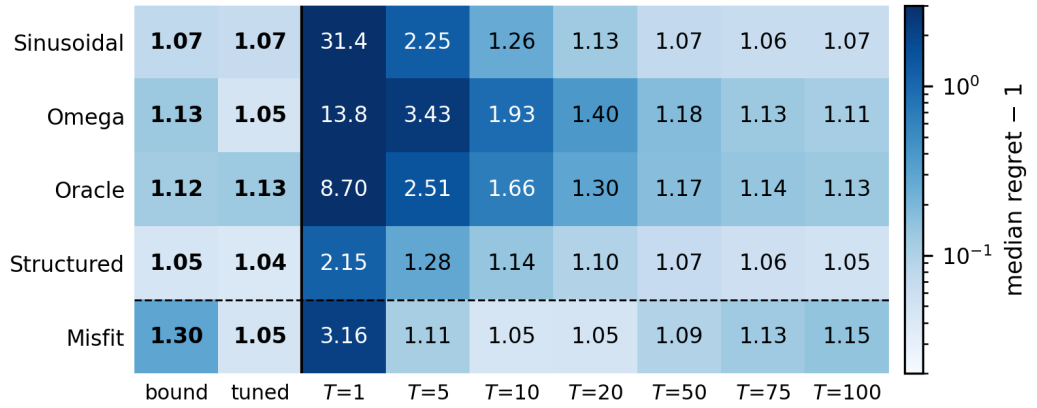

Figure 2: Median per-task regret of the bound-selected depth (bound), a single depth tuned on validation tasks (tuned) and fixed depths, on independent query sets.

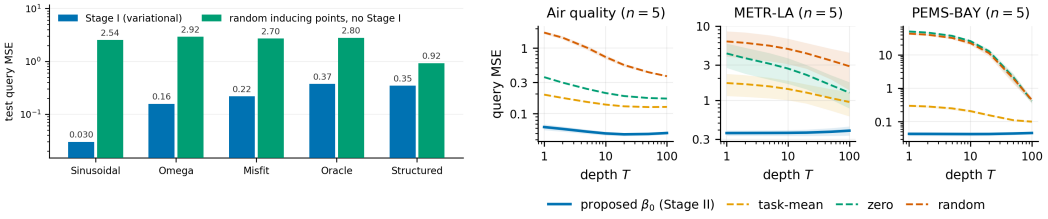

Figure 3: Ablations of pre-training and shared head initialization. Left: final-model query MSE with adaptation depth selected on validation tasks. Right: query MSE across adaptation depths for different shared head initializations, with the kernel and inducing points fixed.

6 Limitations and Conclusion

Limitations. Our analysis focuses on squared-loss adaptation within a finite induced space. The stopping experiments require additional task-specific labels to estimate quantities in the conditional bound, and the resulting plug-in rule is less effective in some settings. Although the proposed model is a hybrid approach of both kernel and gradient descent optimization, some tasks favor closed-form kernel regression. For example, successful iKML training can achieve lower errors in synthetic datasets.

Conclusion. We presented a kernel meta-learning method that learns an induced space and a shared initial function for functional gradient adaptation. The induced space allows the shared initial function to be defined well over the tasks and provides an explicit expression for gradient-based adaptation. Our analysis characterizes the adaptation trajectory and derives conditional and uniform excess transfer risk bounds with early-stopping rules. Experiments on synthetic and real world regression tasks show comparable predictive performance, and the ablations demonstrate benefits from pre-training and the shared initial function in several settings.

Reproducibility Statement

Appendix A collects the notation, sampling conventions, and auxiliary lemmas. Appendices B and C give the induced space analysis and derive the shared initial function. Appendices D and E respectively provide the conditional and uniform excess transfer risk bounds and early stopping depths with corresponding balanced risks. Code and exact configurations are provided in the supplementary material.

REFERENCES

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Kernel interpolation with continuous volume sampling. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 725–735. PMLR, 2020. URL `https://proceedings.mlr.press/v119/belhadji20a.html`.

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019.

Liam Collins, Aryan Mokhtari, Sewoong Oh, and Sanjay Shakkottai. MAML and ANIL provably learn representations. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

John I. T. Falk, Carlo Ciliberto, and Massimiliano Pontil. Implicit kernel meta-learning using kernel integral forms. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Massimiliano Patacchiola, Jack Turner, Elliot J. Crowley, Michael O'Boyle, and Amos Storkey. Bayesian meta-learning for the few-shot setting via deep kernels. In *Advances in Neural Information Processing Systems*, 2020.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959, 2005.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *International Conference on Learning Representations*, 2020.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014.

Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

Michalis K. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

Ze Wang, Zichen Miao, Xiantong Zhen, and Qiang Qiu. Learning to learn dense gaussian processes for few-shot learning. In *Advances in Neural Information Processing Systems*, 2021.

Jin Xu, Jean-François Ton, Hyunjik Kim, Adam R. Kosiorek, and Yee Whye Teh. MetaFun: Meta-learning with iterative functional updates. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315, 2007.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A*, 473(2205):20170457, 2017. doi: 10.1098/rspa.2017.0457.

Xiantong Zhen, Haoliang Sun, Yingjun Du, Jun Xu, Yilong Yin, Ling Shao, and Cees G. M. Snoek. Learning to learn kernels with variational random features. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 11409–11419, 2020.

Yufan Zhou, Zhenyi Wang, Jiayi Xian, Changyou Chen, and Jinhui Xu. Meta-learning with neural tangent kernels. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=Ti87Pv5Oc8`.

Luisa M. Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

A NOTATION AND PRELIMINARIES

We first collect the notation, sampling conventions, and auxiliary results used in the proofs. Appendix B studies the inducing-space residual, Appendix C derives the shared initialization criterion, and Appendices D and E establish the two excess transfer risk bounds.

A.1 NOTATION

Table 2: Notation used in the analysis. Definitions involving the learned representation are conditional on the training data as specified in Appendix A.2.

Symbol	Meaning
Tasks and observations	
$S,\ \mathcal{P}_{\mathrm{task}}$	A task and the distribution from which tasks are sampled.
$P_{s\|S},\ P_{q\|S}$	Task-conditional support and query distributions; the same notation denotes the input marginal when the integrand depends only on X.
$D_m^s,\ D_m^q$	Support and query samples for meta-training task m. In empirical support losses, $D_m = D_m^s$.
$D_S,\ X_S,\ Y_S$	The fresh support sample and its input and label vectors. We also write $X_S^s = X_S$.
$M_0,\ M,\ n$	Numbers of pre-training tasks and meta-training tasks, and observations in each task sample. Each support and query sample has n observations; pre-training also uses n observations per task.
$\mathcal{F}_{\mathrm{VI}},\ \mathcal{F}_M$	Sigma-fields generated by pre-training and meta-training, respectively, including the corresponding algorithmic randomness.
$\mathcal{F}_{\mathrm{train}},\ \mathcal{G}_S$	$\mathcal{F}_{\mathrm{VI}} \vee \mathcal{F}_M$ and $\sigma(\mathcal{F}_{\mathrm{train}}, S, X_S)$, respectively.
$f_S^*,\ \varepsilon_S$	The task regression function and the support noise vector $Y_S - f_S^*(X_S)$.
$\sigma^2,\ \sigma_0^2$	An upper bound on the conditional support-noise variance and the positive likelihood variance in the pre-training objective.
Representation and adaptation	
$h_\psi,\ k_\theta,\ \vartheta$	Feature extractor, latent-space kernel, and full representation parameter $\vartheta = (\psi, Z, \theta)$.
$\mathcal{U},\ d,\ p,\ Z$	Latent space $\mathcal{U} \subseteq \mathbb{R}^d$, its dimension, number of inducing points, and inducing points $Z = (z_1, \dots, z_p)$.
$K_{ZZ},\ K_S,\ Q_S$	Inducing Gram matrix, support Gram matrix, and support Nyström matrix $K(X_S, Z)K_{ZZ}^{-1}K(Z, X_S)$.
$\mathcal{H}_k,\ \mathcal{H}_{k,Z}$	The RKHS of the reference kernel and its induced space $\mathrm{span}\{k(\cdot, z_j) : 1 \le j \le p\}$.
$\phi_\vartheta,\ \Phi_D^\vartheta$	Whitened feature vector in $\mathbb{R}^p$ and the matrix with rows $\phi_\vartheta(X_i)^\top$.
$\beta,\ f_{\vartheta,\beta}$	Whitened head coefficient and predictor $\phi_\vartheta(\cdot)^\top\beta$.
$H_D^\vartheta,\ b_D^\vartheta$	Support Hessian $(\Phi_D^\vartheta)^\top\Phi_D^\vartheta/n$ and vector $(\Phi_D^\vartheta)^\top Y_D/n$.
$G_m,\ g_m$	Query Hessian $(\Phi_m^q)^\top\Phi_m^q/n$ and vector $(\Phi_m^q)^\top Y_m^q/n$.
$\tau,\ E,\ T$	Head step size, number of support updates during meta-training, and number of test-time support updates.
$A_D^\vartheta,\ A_{m,E}$	One-step matrix $I - \tau H_D^\vartheta$ and E-step matrix $(I - \tau H_m)^E$.
$d_{D,T}^\vartheta,\ \mathcal{A}_T^\vartheta$	Affine update term and adaptation map, defined in Appendix A.2.
$P_{M,E},\ q_{M,E}$	Quadratic and linear coefficients of the shared initialization criterion in (9).
$\eta,\ \gamma$	Weight of the query-gradient term and damping parameter in the initialization objective.
$\beta^\star(\vartheta),\ \widehat{\beta}_0$	The minimizer in (10) and its value at the final representation.
$U_S,\ \Lambda_S,\ \lambda_{S,i}$	Orthonormal eigenbasis, diagonal eigenvalue matrix, and eigenvalues of H_S.
$I_S,\ \Pi_{S,0},\ \Pi_{S,+}$	Positive-eigenvalue index set and orthogonal projectors onto $\ker H_S$ and $\mathrm{range}\, H_S$.
Projections, residuals, and risks	
P_Z	RKHS-orthogonal projection onto $\mathcal{H}_{k,Z}$.
$\bar{f}_{S,Z},\ r_{S,Z}$	A best induced space approximation under $L^2(P_{s\|S})$ and the residual $f_S^* - \bar{f}_{S,Z}$.

Continued on the next page

Table 2 continued

Symbol	Meaning
$\Sigma_S,\ \bar{\beta}_S$	Query feature second-moment matrix and minimum-norm population least-squares coefficient at the final representation.
$\bar{f}_S,\ r_S$	Query projection in the learned space. $f_{\widehat{\vartheta},\bar{\beta}_S}$ and residual $f_S^* - \bar{f}_S$. They also give the support projection when $P_{s\|S} = P_{q\|S}$.
$\widehat{\beta}_m$	An empirical least-squares minimizer for the support sample D_m in the induced space.
$e_{S,t},\ \xi_S$	Spectral coefficient error $U_S^\top(\beta_{S,t} - \bar{\beta}_S)$ and forcing vector $U_S^\top(b_S - H_S\bar{\beta}_S)$.
$\theta_{S,t},\ \mu_S$	Spectral coordinates $U_S^\top\beta_{S,t}$ and $U_S^\top b_S$, used locally in Appendix D.
$L_S,\ \widehat{L}_D$	Population query loss and empirical squared loss, both normalized by $1/2$.
$\mathcal{R}_S^{sc}(T),\ \mathcal{R}^u(T)$	Excess risks conditional on $\mathcal{G}_S$ and $\mathcal{F}_{\mathrm{VI}}$, respectively; see (15) and (16).
$\mathcal{R}_{\mathrm{within}}(T), \mathcal{R}_{\mathrm{across}}(T), \mathcal{O}(T)$	Within-task, across-task, and optimization terms in (20).
$\mathcal{A}^*_{\mathrm{sub}}$	Test-time independent (T-independent) term defined in (20).
$g_Z(x),\ R(Z),\ \widehat{R}_{M_0,n}(Z)$	Pointwise kernel projection error, its population average, and its pooled empirical average.
$\widehat{Z}_{\mathrm{tr}}$	Reference empirical trace minimizer over $\mathcal{Z}_p$.
ρ_p	Final population squared residual averaged over meta-training randomness and a fresh task, conditional on $\mathcal{F}_{\mathrm{VI}}$.
ρ_S	Common scalar contraction factor in Corollary 15; it is distinct from ρ_p.
$F_S,\ B_S,\ V_S$	Floor, initial-error coefficient, and residual/noise coefficient of the conditional excess risk bound in Corollary 15.
Δ_S	Frobenius deviation of the augmented feature second-moment matrix in Lemma 16.
$\widehat{\mathcal{P}}_{E,M}, A', a_0, b_1, b_2, C_0, L_n$	Adaptation energy and quantities in the uniform risk bound and stopping rule; see Lemma 18 and (21).
Bounds and complexity	
$B_y,\ R_0,\ B$	Bounds on labels, the initial head norm, and the target RKHS or population projected-head norms, where each bound is assumed.
$\mu > 0$	Lower bound on positive training-Hessian eigenvalues in Lemma 18; distinct from the vector μ_S.
$\kappa^2,\ L_k,\ \Lambda$	Kernel diagonal bound, kernel Lipschitz constant, and uniform lower bound on $\lambda_{\min}(K_{ZZ})$ over $\mathcal{Z}_p$.
$\lambda_j,\ C_\alpha,\ \alpha$	Eigenvalues and polynomial-decay parameters of the kernel covariance operator in Assumption 1.
$\mathcal{Z}_p,\ R$	Admissible inducing points and the latent-space radius used in $\mathcal{Z}_p \subseteq (B_d(0,R))^p$.
$L_g,\ \Delta_{M_0,n}(\epsilon,\delta)$	Lipschitz constant of g_Z in Z and the trace-concentration error in Corollary 10.
$\mathcal{V},\ d_{\mathcal{V}},\ L_{\mathcal{V}}$	Admissible representation class, its metric, and the uniform representation Lipschitz bound for the meta-level (empirical) loss.
$D_\psi,\ D_{\mathrm{hyp}}$	Dimensions of the feature-extractor and kernel-parameter spaces; the inducing points contribute pd parameters.
$\mathcal{N}(\epsilon,\mathcal{V},d_{\mathcal{V}})$	Internal ϵ-covering number of $\mathcal{V}$.
$\epsilon,\ \delta,\ C_K$	Covering radius, failure probability, and universal constant in the matrix Khintchine bound.
ω_i	Independent Rademacher variables, also independent of the observations: $\mathbb{P}(\omega_i = 1) = \mathbb{P}(\omega_i = -1) = 1/2$.

For vectors, $\langle u, v\rangle = u^\top v$ and $\|u\|_2^2 = u^\top u$; these expressions contain no sample-size normalization. For a matrix A, $\|A\|_{\mathrm{op}}$ and $\|A\|_F$ denote its operator and Frobenius norms, $A^\dagger$ its Moore–Penrose pseudoinverse, and $A^\top$ its transpose. The exponent T denotes an integer power, so A^T and $A^\top$ have different meanings. For symmetric matrices, $A \preceq B$ means that $B - A$ is positive semidefinite. We write $[a]_+ = \max\{a, 0\}$ and use natural logarithms. For a probability measure P,

$$\|g\|^2_{L^2(P)} = \int |g(x)|^2\, dP(x).$$

The notation $\widetilde{O}$ suppresses logarithmic factors; the arguments of those factors are retained in the preceding explicit bounds.

A.2 Sampling, coordinates, and projections

Sampling and conditioning. A task S determines a regression function $f_S^*(x) = \mathbb{E}[Y \mid X = x, S]$ and support and query distributions. The regression function is assumed square-integrable under the input distributions used for the projections below. Within each task, observations in each sample are independent and identically distributed; the support and query samples are independent given the task. The fresh task and its observations are independent of training. The training sigma-fields and the support-conditioning sigma-field are

$$\mathcal{F}_{\mathrm{train}} = \mathcal{F}_{\mathrm{VI}} \vee \mathcal{F}_M, \qquad \mathcal{G}_S = \sigma(\mathcal{F}_{\mathrm{train}}, S, X_S).$$

The final representation $\widehat{\vartheta}$ and initialization $\widehat{\beta}_0$ are $\mathcal{F}_{\mathrm{train}}$-measurable. Conditioning on $\mathcal{G}_S$ therefore fixes the representation, initialization, task, and support inputs. The conditional analysis averages the support label noise and assumes

$$\mathbb{E}[\varepsilon_S \mid \mathcal{G}_S] = 0, \qquad \mathbb{E}[\varepsilon_S \varepsilon_S^\top \mid \mathcal{G}_S] \preceq \sigma^2 I_n.$$

The query noise has conditional mean zero given its input, the task, the training data, and the support sample. In the uniform excess transfer risk analysis, $\mathbb{E}[\cdot \mid \mathcal{F}_{\mathrm{VI}}]$ also averages the meta-training datasets, their associated query samples and algorithmic randomness, and the fresh task and support sample. Thus a residual defined using $\widehat{\vartheta}$ is also averaged over meta-training. Independence across tasks is imposed conditional on $\mathcal{F}_{\mathrm{VI}}$ as in Appendix E.

Whitened coordinates and empirical operators. For a fixed representation $\vartheta = (\psi, Z, \theta)$, kernel evaluations involving an observed input include the feature map:

$$k(x, Z) = k_\theta(h_\psi(x), Z), \qquad k(x, x') = k_\theta(h_\psi(x), h_\psi(x')).$$

The inducing matrix is $K_{ZZ} = k_\theta(Z, Z) \succ 0$. The associated latent-space function $k_\theta(\cdot, Z)K_{ZZ}^{-1/2}\beta$ has RKHS norm $\|\beta\|_2$. Its prediction on an observed input is

$$f_{\vartheta,\beta}(x) = \phi_\vartheta(x)^\top \beta, \qquad \phi_\vartheta(x) = K_{ZZ}^{-1/2} k_\theta(Z, h_\psi(x)).$$

Positivity of the Gram matrix on $Z \cup \{h_\psi(x)\}$ gives

$$0 \le \|\phi_\vartheta(x)\|_2^2 = k(x, Z)K_{ZZ}^{-1}k(Z, x) \le k(x, x).$$

For $D = \{(X_i, Y_i)\}_{i=1}^n$, define

$$\Phi_D^\vartheta = \begin{pmatrix} \phi_\vartheta(X_1)^\top \\ \vdots \\ \phi_\vartheta(X_n)^\top \end{pmatrix}, \qquad H_D^\vartheta = \frac{(\Phi_D^\vartheta)^\top \Phi_D^\vartheta}{n}, \qquad b_D^\vartheta = \frac{(\Phi_D^\vartheta)^\top Y_D}{n}.$$

In particular, $\mathrm{rank}(H_D^\vartheta) \le \min\{n, p\}$ and, when $k(x, x) \le 1$,

$$\|H_D^\vartheta\|_{\mathrm{op}} \le \mathrm{tr}\, H_D^\vartheta = \frac{1}{n}\sum_{i=1}^n \|\phi_\vartheta(X_i)\|_2^2 \le 1.$$

The one-step matrix, affine term, and adaptation map are

$$A_D^\vartheta = I - \tau H_D^\vartheta, \qquad d_{D,T}^\vartheta = \tau \sum_{t=0}^{T-1} (A_D^\vartheta)^t b_D^\vartheta,$$

$$\mathcal{A}_T^\vartheta(\beta; D) = (A_D^\vartheta)^T \beta + d_{D,T}^\vartheta.$$

This agrees with $\mathcal{A}_{m,E}^\vartheta$ in the main text when $D = D_m$ and $T = E$. Superscripts ϑ are omitted when the representation is fixed. The diagonalization $H_S = U_S \Lambda_S U_S^\top$ defines

$$I_S = \{i : \lambda_{S,i} > 0\}, \qquad \Pi_{S,+} = H_S^\dagger H_S, \qquad \Pi_{S,0} = I - \Pi_{S,+}.$$

Population and empirical projections. The population and empirical losses are

$$L_S(f) = \frac{1}{2}\mathbb{E}_{(X,Y)\sim P_{q|S}}[(f(X)-Y)^2], \qquad \widehat{L}_D(f) = \frac{1}{2|D|}\sum_{(X_i,Y_i)\in D}(f(X_i)-Y_i)^2.$$

At the final representation, the minimum-norm coefficient of the population query projection is

$$\Sigma_S = \mathbb{E}_{X\sim P_{q|S}}[\phi_{\widehat{\vartheta}}(X)\phi_{\widehat{\vartheta}}(X)^\top \mid S, \mathcal{F}_{\mathrm{train}}],$$
$$\bar{\beta}_S = \Sigma_S^\dagger \mathbb{E}_{X\sim P_{q|S}}[\phi_{\widehat{\vartheta}}(X) f_S^*(X) \mid S, \mathcal{F}_{\mathrm{train}}], \qquad \bar{f}_S = f_{\widehat{\vartheta},\bar{\beta}_S}, \quad r_S = f_S^* - \bar{f}_S.$$

Finite-dimensionality gives a closed subspace in $L^2(P_{q|S})$, so the projected function exists and is unique as an L^2 equivalence class. The coefficient above selects its minimum Euclidean norm representation. Orthogonality gives

$$\|f_{\widehat{\vartheta},\beta} - f_S^*\|^2_{L^2(P_{q|S})} = \|r_S\|^2_{L^2(P_{q|S})} + \|f_{\widehat{\vartheta},\beta} - \bar{f}_S\|^2_{L^2(P_{q|S})}.$$

Hence

$$\|f_S^* - \bar{f}_{S,Z}\|^2_{L^2(P_{s|S})} \le \|f_S^* - P_Z f_S^*\|^2_{L^2(P_{s|S})}.$$

The reference trace quantity is defined on inducing points with $K_{ZZ} \succ 0$:

$$g_Z(x) = k(x,x) - k(x,Z)K_{ZZ}^{-1}k(Z,x),$$
$$R(Z) = \mathbb{E}_S \mathbb{E}_{X\sim P_{s|S}} g_Z(X),$$
$$\widehat{R}_{M_0,n}(Z) = \frac{1}{M_0 n}\sum_{m=1}^{M_0}\sum_{i=1}^{n} g_Z(X_{m,i}).$$

The trace concentration result concerns a fixed feature map and kernel. Its empirical minimizer is denoted by $\widehat{Z}_{\mathrm{tr}}$. Applying the resulting residual estimate to the final representation uses the additional comparison in (23).

In Appendix E, support and query distributions coincide. The final residual is then

$$\rho_p = \mathbb{E}[\|r_S\|^2_{L^2(P_{s|S})} \mid \mathcal{F}_{\mathrm{VI}}].$$

Coverings and measurable classes. For a metric space $(\mathcal{V}, d_{\mathcal{V}})$, an internal ϵ-net is a finite subset $\{\vartheta_1,\dots,\vartheta_N\} \subseteq \mathcal{V}$ such that every $\vartheta \in \mathcal{V}$ is within distance ϵ of a net point. The least possible N is $\mathcal{N}(\epsilon, \mathcal{V}, d_{\mathcal{V}})$. For inducing points, we use the Frobenius norm, equivalently the Euclidean norm after identifying Z with a vector in $\mathbb{R}^{pd}$. All suprema, minimizers, and conditional expectations used below are assumed measurable; separable continuous parameter classes provide the usual sufficient setting.

A.3 Auxiliary lemmas

The following results are stated in the forms used below. We provide proof for each lemma if necessary. In probabilistic statements, an identical conditional version holds whenever the hypotheses hold under the specified conditional law. In particular, the Rademacher arguments in Lemma 17 condition on $\mathcal{F}_{\mathrm{VI}}$, and their empirical versions also condition on the task samples.

Lemma A.1 (Euclidean coverings and extension from a net)**.** *If $\mathcal{W} \subseteq B_d(0,R)$ and $\epsilon > 0$, then its internal covering number satisfies*

$$\mathcal{N}(\epsilon, \mathcal{W}, \|\cdot\|_2) \le (1 + 2R/\epsilon)^d.$$

If each member of a family $\{h_\vartheta : \vartheta \in \mathcal{V}\}$ satisfies $|h_\vartheta(x) - h_{\vartheta'}(x)| \le L d_{\mathcal{V}}(\vartheta,\vartheta')$, then any internal ϵ-net $\{\vartheta_k\}_{k=1}^N$ obeys

$$\sup_{\vartheta\in\mathcal{V}}\left\{\mathbb{E}h_\vartheta(X) - \frac{1}{M}\sum_{i=1}^{M} h_\vartheta(X_i)\right\} \le \max_{k\le N}\left\{\mathbb{E}h_{\vartheta_k}(X) - \frac{1}{M}\sum_{i=1}^{M} h_{\vartheta_k}(X_i)\right\} + 2L\epsilon.$$

The same statement holds with an absolute value around each deviation.

Proof. Choose a maximal ϵ-separated subset of $\mathcal{W}$. The open balls of radius $\epsilon/2$ around its points are disjoint and contained in $B_d(0, R + \epsilon/2)$. Comparing their volumes gives the cardinality bound, and maximality makes the subset an internal ϵ-net. For the second assertion, choose a net point within ϵ of ϑ. Both the population and empirical averages change by at most $L\epsilon$. □

Lemma A.2 (Empirical minimization under uniform deviation)**.** *Let R and $\widehat{R}$ be finite real-valued functions on a nonempty set $\mathcal{W}$, and let $\widehat{w} \in \arg\min_{w\in\mathcal{W}} \widehat{R}(w)$. If the right-hand side below is finite, then*

$$R(\widehat{w}) - \inf_{w\in\mathcal{W}} R(w) \le 2 \sup_{w\in\mathcal{W}} |R(w) - \widehat{R}(w)|.$$

Proof. For every $w \in \mathcal{W}$, insert $\widehat{R}(\widehat{w})$ and $\widehat{R}(w)$ into $R(\widehat{w}) - R(w)$. The difference $\widehat{R}(\widehat{w}) - \widehat{R}(w)$ is nonpositive; each remaining difference is bounded by the displayed supremum. Take the infimum over w. □

Lemma A.3 (Hoeffding's lemma and inequality (Hoeffding, 1963))**.** *If $a \le X \le b$ almost surely, then, for every $s \in \mathbb{R}$,*

$$\mathbb{E}e^{s(X-\mathbb{E}X)} \le e^{s^2(b-a)^2/8}.$$

Consequently, for independent $X_1, \dots, X_N$ with $a_i \le X_i \le b_i$ and $\sum_i (b_i - a_i)^2 > 0$,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbb{E}X_i)\right| \ge t\right) \le 2\exp\left(-\frac{2N^2t^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right), \qquad t > 0.$$

For K such averages with a common range width $b - a$, the union bound replaces the prefactor 2 by $2K$.

Lemma A.4 (Bounded differences (McDiarmid, 1989))**.** *Let $X_1, \dots, X_N$ be independent. Suppose changing only coordinate i of the input to a measurable function h changes its value by at most c_i. Then*

$$\mathbb{E}e^{s(h-\mathbb{E}h)} \le \exp\left(\frac{s^2}{8}\sum_{i=1}^{N} c_i^2\right), \qquad s \in \mathbb{R}.$$

If $\sum_i c_i^2 > 0$, it follows that

$$\mathbb{P}(h - \mathbb{E}h \ge t) \le \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right), \qquad t > 0.$$

Lemma A.5 (Second moments in a Hilbert space)**.** *For independent, mean-zero, square-integrable random elements X_i of a real Hilbert space,*

$$\mathbb{E}\left\|\sum_{i=1}^{N} X_i\right\|^2 = \sum_{i=1}^{N} \mathbb{E}\|X_i\|^2.$$

In particular, for i.i.d. square-integrable Y_i,

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mathbb{E}Y_i)\right\| \le \frac{\{\mathbb{E}\|Y_1\|^2\}^{1/2}}{\sqrt{N}}.$$

For fixed vectors v_i and independent Rademacher signs ω_i,

$$\mathbb{E}_\omega\left\|\sum_{i=1}^{N}\omega_i v_i\right\| \le \left(\mathbb{E}_\omega\left\|\sum_{i=1}^{N}\omega_i v_i\right\|^2\right)^{1/2} = \left(\sum_{i=1}^{N}\|v_i\|^2\right)^{1/2}.$$

These statements apply to matrices equipped with the Frobenius norm.

Rademacher complexity convention. For a measurable real-valued function class $\mathcal{J}$ and a fixed sample $D_{1:M}$, we use the normalization

$$\widehat{\mathfrak{R}}_{D_{1:M}}(\mathcal{J}) = \mathbb{E}_\omega \left[\sup_{j\in\mathcal{J}} \frac{1}{M} \sum_{m=1}^{M} \omega_m j(D_m) \,\middle|\, D_{1:M} \right].$$

Lemma A.6 (Expected symmetrization (Bartlett & Mendelson, 2002)). *Let $D_1, \dots, D_M$ be i.i.d., let D be an independent observation with the same law, and let $\mathcal{J}$ be a fixed measurable class with an integrable bound. Then*

$$\mathbb{E}_{D_{1:M}} \sup_{j\in\mathcal{J}} \left\{ \mathbb{E}_D j(D) - \frac{1}{M} \sum_{m=1}^{M} j(D_m) \right\} \le 2\mathbb{E}_{D_{1:M}} \widehat{\mathfrak{R}}_{D_{1:M}}(\mathcal{J}).$$

For the conditional version, $\mathcal{J}$ may depend on the conditioning sigma-field, and the observations must be i.i.d. under that conditional law.

Lemma A.7 (Finite union of empirical Rademacher classes). *Fix $D_{1:M}$ and nonempty classes $\mathcal{J}_1, \dots, \mathcal{J}_K$. If $|j(D_m)| \le b$ for all $j \in \bigcup_k \mathcal{J}_k$ and all m, then*

$$\widehat{\mathfrak{R}}_{D_{1:M}} \left(\bigcup_{k=1}^{K} \mathcal{J}_k \right) \le \max_{k\le K} \widehat{\mathfrak{R}}_{D_{1:M}}(\mathcal{J}_k) + b\sqrt{\frac{2\log K}{M}}.$$

Lemma A.8 (Matrix Khintchine bound (Tropp, 2015, Theorem 4.1.1)). *For fixed real symmetric matrices $Q_1, \dots, Q_M \in \mathbb{R}^{p\times p}$ and independent Rademacher variables $\{\omega_i\}_{i=1}^M$,*

$$\mathbb{E}_\omega \left\| \sum_{m=1}^{M} \omega_m Q_m \right\|_{\mathrm{op}} \le C_K \sqrt{\log(2p)} \left\| \left(\sum_{m=1}^{M} Q_m^2 \right)^{1/2} \right\|_{\mathrm{op}},$$

where C_K is a universal constant. Thus $\|Q_m\|_{\mathrm{op}} \le b$ implies an upper bound $C_K b\sqrt{M\log(2p)}$. The statement also holds after conditioning on any variables that determine the matrices and are independent of the Rademacher variables.

Lemma A.9 (Least-squares descent and spectral bounds). *Let $\widehat{L}_D(\beta) = \|\Phi_D\beta - Y_D\|_2^2/(2n)$, $H = \Phi_D^\top\Phi_D/n$, and $b = \Phi_D^\top Y_D/n$. Suppose $0 \preceq \tau H \preceq I$, with $\tau > 0$, and let $A = I - \tau H$. Then $b \in \operatorname{range} H$, and the update $\beta^+ = A\beta + \tau b$ satisfies*

$$\widehat{L}_D(\beta^+) - \widehat{L}_D(\beta) = -\tau(H\beta - b)^\top (I - \tau H/2)(H\beta - b) \le 0.$$

For every integer $T \ge 1$ and each positive eigenvalue λ of H,

$$\begin{aligned} 0 \le 1 - (1-\tau\lambda)^T &\le \min\{1, \tau T\lambda\}, \\ \frac{\{1-(1-\tau\lambda)^T\}^2}{\lambda} &\le \min\{\tau T, \tau^2 T^2 \lambda\}, \\ \|HA^{2T}\|_{\mathrm{op}} &\le \frac{1}{2e\tau T}. \end{aligned}$$

If $\operatorname{tr} H \le 1$, these bounds give

$$\sum_{\lambda_j > 0} \frac{\{1-(1-\tau\lambda_j)^T\}^2}{\lambda_j} \le \tau^2 T^2 \operatorname{tr} H \le \tau^2 T^2.$$

Finally, if every positive eigenvalue of H is at least $\mu > 0$, then, for any vector v and integer $E \ge 1$,

$$\|(I - A^E)v\|_2 \ge [1 - (1-\tau\mu)^E] \|\Pi_+ v\|_2, \qquad \Pi_+ = H^\dagger H.$$

Proof. The equality $\ker H = \ker \Phi_D$ implies $b \in (\ker H)^\perp = \operatorname{range} H$. Expanding the quadratic loss after the update proves the descent identity. Bernoulli's inequality gives the first scalar bound. Squaring its two upper bounds gives $\{1-(1-\tau\lambda)^T\}^2/\lambda \le \min\{1/\lambda, \tau^2T^2\lambda\} \le \tau T$, as well as the stated $\tau^2T^2\lambda$ bound. Also $\lambda(1-\tau\lambda)^{2T} \le \lambda e^{-2\tau T\lambda} \le (2e\tau T)^{-1}$. Diagonalization gives the operator and trace statements. On $\operatorname{range} H$, every eigenvalue of $I - A^E$ is at least $1 - (1-\tau\mu)^E$. It is zero on $\ker H$. □

Lemma A.10 (Tail integration for a dependent product)**.** *Let F, U be nonnegative random variables with $U \le b$ almost surely and $\mathbb{E}[U] \le m$. Suppose $\mathbb{P}(F > a + t) \le e^{-ct^2}$ for $t > 0$, where $a \ge 0$ and $c > 0$. For every $t_0 > 0$,*

$$\mathbb{E}[FU] \le am + \int_0^\infty \min\{m, be^{-ct^2}\}\, dt \le m(a + t_0) + \frac{b\, e^{-ct_0^2}}{2ct_0}.$$

Proof. Use $F \le a + (F - a)_+$ and Tonelli's theorem. For every $t > 0$, $\mathbb{E}[U\mathbf{1}_{\{F>a+t\}}] \le \min\{m, b\mathbb{P}(F > a + t)\}$. Split the integral at t_0 and use $\int_{t_0}^\infty e^{-ct^2}\, dt \le t_0^{-1} \int_{t_0}^\infty te^{-ct^2}\, dt = e^{-ct_0^2}/(2ct_0)$. □

Lemma A.3 and the covering and empirical-minimization lemmas are used for connecting the trace term and residual. The Hilbert-space moment, bounded-differences, spectral, and tail-integration lemmas are used for the within-task gap. Symmetrization, the finite-union bound, and the matrix and vector Rademacher bounds are used for the across-task gap. The last spectral inequality also links training adaptation energy to optimization error.

B Connection between variational learning of inducing points and residual reduction

Given the task $S \sim \mathcal{P}_{\text{task}}$, inducing point $Z = (z_1, \cdots, z_p)$ and the kernel k, we define an objective on the subspace $\mathcal{H}_{k,Z}$ as

$$\mathcal{L}_{S,Z}(f) := \mathbb{E}_{x\sim P_{s|S}}\left[\frac{1}{2}\left(f(x) - f_S^*(x)\right)^2\right], \quad f \in \mathcal{H}_{k,Z}.$$

Since $\mathcal{H}_{k,Z}$ is finite-dimensional, it admits a minimizer $\bar{f}_{S,Z} \in \mathcal{H}_{k,Z}$. From this point, choose a minimizer $\bar{f}_{S,Z}$ and define the residual function $r_{S,Z}$ as $r_{S,Z} := f_S^* - \bar{f}_{S,Z} \in L^2(P_{s|S})$. From this, we can quantify the left over term produced after the approximation inside the subspace $\mathcal{H}_{k,Z}$.

Lemma 6. *For any task S with $f_S^* \in \mathcal{H}_k$ and inducing points Z where $K_{ZZ} \succ 0$,*

$$\mathbb{E}_{x\sim P_{s|S}}[r_{S,Z}(x)^2] \le \|f_S^*\|_{\mathcal{H}_k}^2 \mathbb{E}_{x\sim P_{s|S}}\left[g_Z(x)\right]$$

where $g_Z(x) := k(x,x) - k(x,Z)K_{ZZ}^{-1}k(Z,x)$.

Proof. Let $P_Z : \mathcal{H}_k \to \mathcal{H}_{k,Z}$ denote the orthogonal projection operator from $\mathcal{H}_k$ to $\mathcal{H}_{k,Z}$. Then, from the well-known identity

$$P_Z f(\cdot) = k(Z,\cdot)^\top K_{ZZ}^{-1} f(Z),$$

we have that

$$\begin{aligned} g_Z(x) &= k(x,x) - k(x,Z)K_{ZZ}^{-1}k(Z,x) \\ &= \langle k(x,\cdot), k(x,\cdot)\rangle_{\mathcal{H}_k} - \left(P_Z k(x,\cdot)\right)(x) \\ &= \langle k(x,\cdot), k(x,\cdot)\rangle_{\mathcal{H}_k} - \langle P_Z k(x,\cdot), P_Z k(x,\cdot)\rangle_{\mathcal{H}_k} \\ &= \|k(x,\cdot)\|_{\mathcal{H}_k}^2 - \|(P_Z k(x,\cdot))\|_{\mathcal{H}_k}^2 \\ &= \|k(x,\cdot) - (P_Z k(x,\cdot))\|_{\mathcal{H}_k}^2 \end{aligned}$$

where the last line follows from the orthogonality of P_Z.

As $\bar{f}_{S,Z}$ is a minimizer of $f \to \mathbb{E}_{x\sim P_{s|S}}\left[((f(x) - f_S^*(x))^2\right]$ over the space $\mathcal{H}_{k,Z}$,

$$\mathbb{E}_{x\sim P_{s|S}}\left[r_{S,Z}(x)^2\right] = \mathbb{E}_{x\sim P_{s|S}}\left[(\bar{f}_{S,Z}(x) - f_S^*(x))^2\right] \le \mathbb{E}_{x\sim P_{s|S}}\left[\left((P_Z f_S^*)(x) - f_S^*(x)\right)^2\right]$$

as $P_Z f_S^* \in \mathcal{H}_{k,Z}$. From the reproducing property and self-adjointness of P_Z,

$$f_S^*(x) - (P_Z f_S^*)(x) = \langle f_S^* - P_Z f_S^*,\ k(x,\cdot)\rangle_{\mathcal{H}_k} = \langle f_S^*, k(x,\cdot) - (P_Z k(x,\cdot))\rangle_{\mathcal{H}_k}.$$

Consequently, by the Cauchy-Schwarz inequality

$$|f_S^*(x) - (P_Z f_S^*)(x)| \le \|f_S^*\|_{\mathcal{H}_k} \|k(x,\cdot) - P_Z k(x,\cdot)\|_{\mathcal{H}_k} = \|f_S^*\|_{\mathcal{H}_k} \sqrt{g_Z(x)}.$$

Using this bound, we have that

$$\mathbb{E}_{x\sim P_{s|S}}[r_{S,Z}(x)^2] \le \mathbb{E}_{x\sim P_{s|S}}\left[\left((P_Z f_S^*)(x) - f_S^*(x)\right)^2\right] \le \|f_S^*\|_{\mathcal{H}_k}^2 \mathbb{E}_{x\sim P_{s|S}}\left[g_Z(x)\right].$$

□

Let K_S denote the Gram matrix of task S from the sample $X_S = (x_{S,1}, \cdots, x_{S,n})$. That is,

$$(K_S)_{i,j} = k(x_{S,i}, x_{S,j}).$$

Define the Nyström matrix induced by Z as

$$Q_S = Q(X_S; Z) = K(X_S, Z)K(Z, Z)^{-1}K(Z, X_S).$$

Titsias (2009) proposes a variational approximation that introduces a Gaussian variational distribution $q(u)$ over inducing points $u = f(Z) \in \mathbb{R}^p$ and optimizes a variational lower bound (ELBO) on the exact GP marginal likelihood. Collapsing the optimal $q(u)$, the resulting bound can be written in the form

$$\mathcal{L}_{\text{VI}}(Z) = \log \mathcal{N}(Y_S|0, Q_S + \sigma_0^2 I) - \frac{1}{2\sigma_0^2}\text{Trace}\left(K_S - Q_S\right).$$

Importantly, the trace term is exactly the total variance left unexplained by the inducing subspace and therefore encourages selecting Z that reduces the Nyström projection error. We provide a lemma on how this can be connected to our residual quantity.

Lemma 7. *For any task S with training sample $X_S = (x_{S,1}, \cdots, x_{S,n})$ and inducing point Z,*

$$\frac{1}{n}\text{Trace}(K_S - Q_S) = \frac{1}{n}\sum_{i=1}^{n} g_Z(x_{S,i}).$$

Proof. This follows directly from the definition of g_Z. Indeed,

$$(K_S - Q_S)_{ii} = k(x_{S,i}, x_{S,i}) - k(Z, x_{S,i})^\top K_{ZZ}^{-1} k(Z, x_{S,i}) = g_Z(x_{S,i}).$$

We obtain the identity by taking the trace sum over $K_S - Q_S$. □

Note that the above identity is satisfied even under the pooled data setting. That is, consider M_0 tasks and for each task we have data sample $X_m = (X_{m,1}, \cdots, X_{m,n})$, $m = 1, \cdots, M_0$, define

$$X := (x_{1,1}, \cdots, x_{M_0,n}), \quad K_{\text{pool}} := K(X, X), \quad Q_{\text{pool}} := K(X, Z)K_{ZZ}^{-1}K(Z, X).$$

Then it follows that

$$\frac{1}{M_0 n}\text{Trace}(K_{\text{pool}} - Q_{\text{pool}}) = \frac{1}{M_0}\sum_{m=1}^{M_0}\frac{1}{n}\text{Trace}(K_m - Q_m) = \frac{1}{M_0 n}\sum_{m,i=1}^{(M_0,n)} g_Z(x_{m,i}).$$

For a fixed feature map and kernel parameters, Lemma 6 bounds the task residual by the kernel projection error. If $\|f_S^*\|_{\mathcal{H}_k} \le B$ for every task, then

$$\mathbb{E}_S \|r_{S,Z}\|_{L^2(P_{s|S})}^2 \le B^2 R(Z).$$

The pooled trace is the empirical average $\widehat{R}_{M_0,n}(Z)$ associated with $R(Z)$. The following uniform bound quantifies their difference and applies to an empirical trace minimizer.

Theorem 8. *Fix the reference feature map h_ψ and kernel k_θ independently of the samples used in this theorem. Draw M_0 independent tasks and denote their observation laws by $\mathcal{D}_m^{\text{VI}}$. Conditional on the tasks, sample $(X_{m,i}, Y_{m,i})_{i=1}^n$ independently from each law, with independence across tasks. Let*

$\mathcal{Z}_p \subseteq \mathcal{U}^p$ be compact. Assume that k_θ is L_k-Lipschitz in each argument, $\sup_{u\in\mathcal{U}} k_\theta(u,u) \le \kappa^2 \le 1$, and $\lambda_{\min}(K_{ZZ}) \ge \Lambda > 0$ for every $Z \in \mathcal{Z}_p$. Define

$$R(Z) := \mathbb{E}_{S\sim\mathcal{P}_{task}}\mathbb{E}_{x\sim P_{s|S}}\left[g_Z(x)\right].$$

Let us define a pooled estimate of $R(Z)$ as

$$\hat{R}_{M_0,n}(Z) := \frac{1}{M_0 n}\sum_{m,i} g_Z(x_{m,i}) = \frac{1}{M_0}\sum_{m=1}^{M_0}\frac{1}{n}\text{Trace}\left(K(X_m,X_m) - Q(X_m;Z)\right)$$

and denote its measurable minimizer as $\hat{Z}_{\text{tr}} \in \arg\min_{Z\in\mathcal{Z}_p}\hat{R}_{M_0,n}(Z)$. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$,

$$R(\hat{Z}_{\text{tr}}) \le \inf_{Z\in\mathcal{Z}_p} R(Z) + 2\sqrt{\frac{\log\mathcal{N}(\epsilon,\mathcal{Z}_p)+\log(4/\delta)}{2M_0}} + 2\sqrt{\frac{\log\mathcal{N}(\epsilon,\mathcal{Z}_p)+\log(4/\delta)}{2M_0 n}} + 8L_g\epsilon.$$

Here $\mathcal{N}(\epsilon,\mathcal{Z}_p)$ is the ϵ-covering number of $\mathcal{Z}_p$ and L_g is the Lipschitz constant of the function g_Z satisfying $|g_Z(x) - g_{Z'}(x)| \le L_g\|Z-Z'\|_2$ for all $x\in\mathcal{X}$, and $Z, Z' \in \mathcal{Z}_p$.

Proof. Let us define the per-task estimate of $R(Z)$ as

$$\bar{R}(Z) = \frac{1}{M_0}\sum_{m=1}^{M_0}\mathbb{E}[g_Z(x)|\mathcal{D}_m^{\text{VI}}].$$

The empirical comparison in Lemma A.2 gives the first step below. Writing $Z^* \in \arg\min_{Z\in\mathcal{Z}_p} R(Z)$, whose existence follows from compactness and continuity here, we can decompose $R(\hat{Z}_{\text{tr}}) - \inf_{Z\in\mathcal{Z}_p} R(Z)$ as

$$\begin{aligned}
R(\hat{Z}_{\text{tr}}) - \inf_{Z\in\mathcal{Z}_p} R(Z) &\le R(\hat{Z}_{\text{tr}}) - \hat{R}_{M_0,n}(\hat{Z}_{\text{tr}}) \\
&\quad + \hat{R}_{M_0,n}(\hat{Z}_{\text{tr}}) - \hat{R}_{M_0,n}(Z^*) + \hat{R}_{M_0,n}(Z^*) - R(Z^*) \\
&\le \sup_{Z\in\mathcal{Z}_p}|R(Z) - \hat{R}_{M_0,n}(Z)| + 0 + \sup_{Z\in\mathcal{Z}_p}|R(Z) - \hat{R}_{M_0,n}(Z)| \\
&= 2\sup_{Z\in\mathcal{Z}_p}|R(Z) - \hat{R}_{M_0,n}(Z)| \\
&\le 2\sup_{Z\in\mathcal{Z}_p}|R(Z) - \bar{R}(Z)| + 2\sup_{Z\in\mathcal{Z}_p}|\bar{R}(Z) - \hat{R}_{M_0,n}(Z)|.
\end{aligned}$$

First, we bound the term $\sup_{Z\in\mathcal{Z}_p}|\bar{R}(Z) - R(Z)|$. As $g_Z(x) = \|k(x,\cdot) - (P_Z k(x,\cdot))\|^2_{\mathcal{H}_k}$, we have that $0 \le g_Z(x) \le \kappa^2$ and hence

$$0 \le \mathbb{E}[g_Z(x)|\mathcal{D}_m^{\text{VI}}] \le \kappa^2 \quad \text{a.s.}$$

Now as $\mathbb{E}_m[\mathbb{E}[g_Z(x)|\mathcal{D}_m^{\text{VI}}]] = R(Z)$ and that $\mathbb{E}[g_Z(x)|\mathcal{D}_m^{\text{VI}}]$ are i.i.d. under the task distribution $\mathcal{P}_{\text{task}}$, we apply Lemma A.3 and a union bound over the ϵ-net $\mathcal{Z}_\epsilon \subset \mathcal{Z}_p$ so that for any $t > 0$,

$$\mathbb{P}\left(\sup_{Z\in\mathcal{Z}_\epsilon}|R(Z) - \bar{R}(Z)| \ge t\right) \le 2\mathcal{N}(\epsilon,\mathcal{Z}_p)\exp\left(-\frac{2M_0 t^2}{\kappa^4}\right).$$

For any $\delta\in(0,1)$, taking $t = \kappa^2\sqrt{\frac{\log\mathcal{N}(\epsilon,\mathcal{Z}_p)+\log(4/\delta)}{2M_0}}$ we have that with probability at least $1-\frac{\delta}{2}$,

$$\sup_{Z\in\mathcal{Z}_\epsilon}|R(Z) - \bar{R}(Z)| \le \kappa^2\sqrt{\frac{\log\mathcal{N}(\epsilon,\mathcal{Z}_p)+\log(4/\delta)}{2M_0}}.$$

To extend this bound over the entire $\mathcal{Z}_p$, we first verify the Lipschitzness of g_z. Indeed,

$$|g_Z(x) - g_{Z'}(x)| = |k(x,Z)K_{ZZ}^{-1}k(Z,x) - k(x,Z')K_{Z'Z'}^{-1}k(Z',x)|$$

$$= |\underbrace{k(x,Z)K_{ZZ}^{-1}k(Z,x) - k(x,Z')K_{ZZ}^{-1}k(Z',x)}_{:=T_1}$$
$$+ \underbrace{k(x,Z')K_{ZZ}^{-1}k(Z',x) - k(x,Z')K_{Z'Z'}^{-1}k(Z',x)}_{:=T_2}|$$
$$\leq |T_1| + |T_2|.$$

Now using the identity for a symmetric matrix M,

$$u^\top Mu - v^\top Mv = (u-v)^\top Mu + v^\top M(u-v),$$

consequently

$$\begin{aligned}
|T_1| &= |k(x,Z)K_{ZZ}^{-1}k(Z,x) - k(x,Z')K_{ZZ}^{-1}k(Z',x)| \\
&= |(k(x,Z) - k(x,Z'))K_{ZZ}^{-1}k(Z,x) + k(x,Z')K_{ZZ}^{-1}(k(Z,x) - k(Z',x))| \\
&\leq \|k(x,Z) - k(x,Z')\|_2 \|K(Z,Z)^{-1}\|_{\text{op}} (\|k(x,Z)\|_2 + \|k(x,Z')\|_2) \\
&\leq L_k \|Z - Z'\| \|K(Z,Z)^{-1}\|_{\text{op}} (2\kappa^2\sqrt{p})
\end{aligned}$$

For T_2, observe that

$$\begin{aligned}
|T_2| &\leq |k(x,Z')(K_{ZZ}^{-1} - K_{Z'Z'}^{-1})k(Z',x)| \\
&\leq \|k(x,Z')\|_2^2 \|K_{ZZ}^{-1} - K_{Z'Z'}^{-1}\|_{\text{op}} \\
&= \|k(x,Z')\|_2^2 \|K_{ZZ}^{-1}(K_{Z'Z'} - K_{ZZ})K_{Z'Z'}^{-1}\|_{\text{op}} \\
&\leq \kappa^4 p \|K(Z,Z)^{-1}\|_{\text{op}} \|K(Z',Z')^{-1}\|_{\text{op}} \|K_{Z'Z'} - K_{ZZ}\|_{\text{op}}
\end{aligned}$$

From simple matrix calculation,

$$\|K_{Z'Z'} - K_{ZZ}\|_{\text{op}} \leq \|K_{Z'Z'} - K_{ZZ}\|_F \leq 2L_k\sqrt{p}\|Z - Z'\|_2.$$

Hence,

$$|T_2| \leq 2L_k p\sqrt{p}\kappa^4 \|K(Z,Z)^{-1}\|_{\text{op}} \|K(Z',Z')^{-1}\|_{\text{op}} \|Z - Z'\|_2$$

Since we have assumed $\boldsymbol{\lambda}_{\min}(K_{ZZ}) \geq \Lambda > 0$ for all $Z \in \mathcal{Z}_p$, we have that $\|K(Z,Z)^{-1}\|_{\text{op}} \|K(Z',Z')^{-1}\|_{\text{op}} \leq \frac{1}{\Lambda^2}$. Thus, by setting L_g as

$$L_g = \frac{2\kappa^2 L_k\sqrt{p}}{\Lambda} + \frac{2\kappa^4 L_k p\sqrt{p}}{\Lambda^2},$$

$g_Z(x)$ is Lipschitz to $Z \in \mathcal{Z}_p$ for all $x \in \mathcal{X}$.
Consequently, $\sup_{Z\in\mathcal{Z}_p}|R(Z) - \bar{R}(Z)| \leq \sup_{Z\in\mathcal{Z}_\epsilon}|R(Z) - \bar{R}(Z)| + 2L_g\epsilon$, and hence

$$\sup_{Z\in\mathcal{Z}_p}|R(Z) - \bar{R}(Z)| \leq \kappa^2\sqrt{\frac{\log\mathcal{N}(\epsilon,\mathcal{Z}_p) + \log(4/\delta)}{2M_0}} + 2L_g\epsilon$$

with probability at least $1 - \frac{\delta}{2}$.

Now for the term $\left|\hat{R}_{M_0,n}(Z) - \frac{1}{M_0}\sum_{m=1}^{M_0}\mathbb{E}[g_Z(x)|\mathcal{D}_m^{\text{VI}}]\right| = |\hat{R}_{M_0,n}(Z) - \bar{R}(Z)|$, we first define

$$U_{m,i} = g_Z(x_{m,i}) - \mathbb{E}[g_Z(x)|\mathcal{D}_m^{\text{VI}}] \quad \text{and} \quad U_m = \frac{1}{n}\sum_{i=1}^{n} U_{m,i}.$$

Then

$$\hat{R}_{M_0,n}(Z) - \bar{R}(Z) = \frac{1}{M_0}\sum_{m=1}^{M_0}\frac{1}{n}\sum_{i=1}^{n}\left(g_Z(x_{m,i}) - \mathbb{E}[g_Z(x)|\mathcal{D}_m^{\text{VI}}]\right) := \frac{1}{M_0}\sum_{m=1}^{M_0} U_m.$$

Note that $U_{m,i}$ lies under $[-\kappa^2,\kappa^2]$ with mean 0. Thus, for all $\lambda \in \mathbb{R}$

$$\mathbb{E}[\exp(\lambda U_{m,i})|\mathcal{D}_m^{\text{VI}}] \leq \exp\left(\frac{\lambda^2\kappa^4}{8}\right)$$

and hence,

$$\mathbb{E}[\exp(\lambda U_m)|\mathcal{D}_m^{\text{VI}}] = \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\frac{\lambda}{n} U_{m,i}\right)|\mathcal{D}_m^{\text{VI}}\right] \leq \prod_{i=1}^{n} \exp\left(\frac{\lambda^2\kappa^4}{8n^2}\right) = \exp\left(\frac{\lambda^2\kappa^4}{8n}\right)$$

As for different tasks, the samples are independent, U_m are also independent,

$$\mathbb{E}_{m\sim\mathcal{P}_{\text{task}}}\mathbb{E}_{\mathcal{D}_m^{\text{VI}}}\left[\exp\left(\lambda\sum_{m=1}^{M_0} U_m\right)\right] = \mathbb{E}_{m\sim\mathcal{P}_{\text{task}}}\prod_{m=1}^{M_0}\mathbb{E}_{\mathcal{D}_m^{\text{VI}}}\left[\exp(\lambda U_m)\right] \leq \exp\left(\frac{M_0\lambda^2\kappa^4}{8n}\right).$$

Applying the moment-generating-function bound in Lemma A.3 conditionally on the tasks, followed by exponential Markov inequality, gives, for all $t > 0$,

$$\begin{aligned}\mathbb{P}\left(\sum_{m=1}^{M_0} U_m \geq M_0 t\right) &\leq \exp(-\lambda M_0 t)\,\mathbb{E}_{m\sim\mathcal{P}_{\text{task}}}\mathbb{E}_{\mathcal{D}_m^{\text{VI}}}\left[\exp\left(\lambda\sum_{m=1}^{M_0} U_m\right)\right] \\ &\leq \exp\left(-\lambda M_0 t + \frac{M_0\lambda^2\kappa^4}{8n}\right).\end{aligned}$$

Taking $\lambda^* = \frac{4nt}{\kappa^4}$ minimizes the above Chernoff bound. Plugging it in yields

$$\mathbb{P}\left(\sum_{m=1}^{M_0} U_m \geq M_0 t\right) \leq \exp\left(-\frac{2M_0 n t^2}{\kappa^4}\right)$$

Now similarly as above, we can apply this inequality over ϵ-net $\mathcal{Z}_\epsilon \subset \mathcal{Z}_p$. Hence,

$$\sup_{Z\in\mathcal{Z}_p} |\hat{R}_{M_0,n}(Z) - \bar{R}(Z)| \leq \kappa^2\sqrt{\frac{\log\mathcal{N}(\epsilon,\mathcal{Z}_p) + \log(4/\delta)}{2M_0 n}} + 2L_g\epsilon$$

Adding this under union bound and using that $\kappa \leq 1$, with probability at least $1-\delta$

$$R(\hat{Z}_{\text{tr}}) - \inf_{Z\in\mathcal{Z}_p} R(Z) \leq 2\sqrt{\frac{\log\mathcal{N}(\epsilon,\mathcal{Z}_p) + \log(4/\delta)}{2M_0}} + 2\sqrt{\frac{\log\mathcal{N}(\epsilon,\mathcal{Z}_p) + \log(4/\delta)}{2M_0 n}} + 8L_g\epsilon.$$

□

We provide a lemma that helps to directly connect the above result to the excess transfer risk bounds.

Assumption 1. *(Polynomial decay of the kernel spectrum). Suppose that $\mathcal{U}$ is a Borel space, $h_\psi : \mathcal{X} \to \mathcal{U}$ is measurable, and $k_\theta : \mathcal{U}\times\mathcal{U} \to \mathbb{R}$ (where we simply write $k(x,x') = k_\theta(h_\psi(x), h_\psi(x'))$) is jointly measurable, symmetric and positive semidefinite. Assume that $\mathbb{E}_{S\sim\mathcal{P}_{task}}\mathbb{E}_{X\sim P_{s|S}}[k(X,X)] < \infty$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ be the eigenvalues of the kernel covariance operator $\mathbb{E}_S\mathbb{E}_{x\sim P_{s|S}}[k(\cdot,x)\otimes k(\cdot,x)]$. Assume that this operator has rank at least p and that, for some constants $C_\alpha > 0$ and $\alpha > 0$,*

$$\lambda_j \leq C_\alpha j^{-(1+\alpha)} \quad \textit{for every } j \geq 1.$$

Lemma 9. *Under Assumption 1, for the fixed kernel parameters h_ψ and k_θ, we assume that Z is independently drawn as*

$$S_i \overset{\text{i.i.d}}{\sim} \mathcal{P}_{\text{task}}, \quad X_i|S_i \sim P_{s|S_i}, \quad z_i = h_\psi(X_i), \quad i = 1,\cdots,p.$$

The symbol $\mathbb{E}_Z[\cdot]$ denotes expectation over the above distribution over Z. Then,

$$\inf_{\substack{Z\in\mathcal{U}^p \\ K_{ZZ}\succ 0}} R(Z) = \inf_{\substack{Z\in\mathcal{U}^p \\ K_{ZZ}\succ 0}} \mathbb{E}_{S\sim\mathcal{P}_{task}}\mathbb{E}_{X\sim P_{s|S}}[g_Z(X)] \leq 2\sum_{j>\lceil p/2\rceil}\lambda_j \leq \frac{2^{\alpha+1}C_\alpha}{\alpha p^\alpha}.$$

Proof. In this proof, determinant-weighted quantities are defined to be zero on singular inducing points. From the continuous-volume-sampling identity (Belhadji et al., 2020), we have that

$$\mathbb{E}\left[\det(K_{ZZ})\right] = p! \sum_{1\leq j_1<\cdots<j_p} \prod_{l=1}^{p}\lambda_{j_l}$$

As the right-hand side is strictly positive (rank of the operator is at least p), the following

$$\frac{\mathbb{E}\left[\mathbf{1}_{\{K_{ZZ}\succ 0\}}\det(K_{ZZ})\mathbb{E}_{S\sim\mathcal{P}_{\text{task}}}\mathbb{E}_{X\sim P_{s|S}}\left[g_Z(X)\right]\right]}{\mathbb{E}[\det(K_{ZZ})]}$$

defines a weighted expectation over the space $\mathcal{U}^p$.
Now for $K_{ZZ}\succ 0$, the Schur-complement formula gives

$$\begin{aligned}
&\det\begin{pmatrix} K_{ZZ} & k_\theta\left(Z, h_\psi(X)\right) \\ k_\theta\left(h_\psi(X), Z\right) & k_\theta\left(h_\psi(X), h_\psi(X)\right)\end{pmatrix} \\
&\quad = \det\left(K_{ZZ}\right)\left[k_\theta\left(h_\psi(X), h_\psi(X)\right) - k_\theta\left(h_\psi(X), Z\right)K_{ZZ}^{-1}k_\theta\left(Z, h_\psi(X)\right)\right] \\
&\quad = \det\left(K_{ZZ}\right)g_Z(X)
\end{aligned}$$

so that

$$\det\left(K_{ZZ}\right)g_Z(X) = \det\left(K_{Z\cup\{h_\psi(X)\},Z\cup\{h_\psi(X)\}}\right).$$

Straightforwardly,

$$\begin{aligned}
\mathbb{E}\left[\det(K_{ZZ})R(Z)\right] &= \mathbb{E}_{Z,X}[\det(K_{ZZ})g_Z(X)] \\
&= \mathbb{E}\left[\det\left(K_{Z\cup\{h_\psi(X)\},Z\cup\{h_\psi(X)\}}\right)\right] \\
&= (p+1)!\sum_{1\le j_1<\cdots<j_{p+1}}\prod_{l=1}^{p+1}\lambda_{j_l}.
\end{aligned}$$

Note that for every integer $r\in\{0,\cdots,p\}$,

$$(p+1-r)\sum_{1\le j_1<\cdots<j_{p+1}}\prod_{l=1}^{p+1}\lambda_{j_l} \le \left(\sum_{j>r}\lambda_j\right)\left(\sum_{1\le j_1<\cdots<j_p}\prod_{l=1}^{p}\lambda_{j_l}\right)$$

Indeed, every $(p+1)$- element index set contains at least $p+1-r$ indices larger than r, as at most r indices can belong to $\{1,\cdots,r\}$. Its eigenvalue product therefore appears at least $p+1-r$ times. Consequently,

$$\begin{aligned}
\frac{(p+1)!\sum_{1\le j_1<\cdots<j_{p+1}}\prod_{l=1}^{p+1}\lambda_{j_l}}{p!\sum_{1\le j_1<\cdots<j_p}\prod_{l=1}^{p}\lambda_{j_l}} &= (p+1)\frac{\sum_{1\le j_1<\cdots<j_{p+1}}\prod_{l=1}^{p+1}\lambda_{j_l}}{\sum_{1\le j_1<\cdots<j_p}\prod_{l=1}^{p}\lambda_{j_l}} \\
&\le \frac{p+1}{p+1-r}\sum_{j>r}\lambda_j
\end{aligned}$$

Set $r=\lceil p/2\rceil$. Then

$$\begin{aligned}
\frac{\mathbb{E}\left[\mathbf{1}_{\{K_{ZZ}>0\}}\det(K_{ZZ})\mathbb{E}_{S\sim\mathcal{P}_{\text{task}}}\mathbb{E}_{X\sim P_{s|S}}\left[g_Z(X)\right]\right]}{\mathbb{E}[\det(K_{ZZ})]} &= \frac{(p+1)!\sum_{1\le j_1<\cdots<j_{p+1}}\prod_{l=1}^{p+1}\lambda_{j_l}}{p!\sum_{1\le j_1<\cdots<j_p}\prod_{l=1}^{p}\lambda_{j_l}} \\
&\le 2\sum_{j>\lceil p/2\rceil}\lambda_j.
\end{aligned}$$

As $\inf_{Z\in\mathcal{U}^p,K_{ZZ}\succ 0}R(Z)\le R(Z)$, multiplying $\det(K_{ZZ})$ and taking expectation

$$\inf_{Z\in\mathcal{U}^p,K_{ZZ}\succ 0}R(Z)\mathbb{E}_Z\left[\det(K_{ZZ})\right] \le \mathbb{E}_Z\left[\mathbf{1}_{\{K_{ZZ}\succ 0\}}\det(K_{ZZ})R(Z)\right]$$

so that $\inf_{Z\in\mathcal{U}^p,K_{ZZ}\succ 0}R(Z)\le\frac{\mathbb{E}_Z\left[\mathbf{1}_{\{K_{ZZ}\succ 0\}}\det(K_{ZZ})R(Z)\right]}{\mathbb{E}_Z[\det(K_{ZZ})]}$. Consequently,

$$\inf_{Z\in\mathcal{U}^p,K_{ZZ}\succ 0}R(Z)\le 2\sum_{j>\lceil p/2\rceil}\lambda_j.$$

Finally, applying the polynomial decay condition,

$$2\sum_{j>\lceil p/2\rceil}\lambda_j \le 2C_\alpha\sum_{j>\lceil p/2\rceil}j^{-(1+\alpha)}$$

$$\leq 2C_\alpha \int_{\lceil p/2 \rceil}^{\infty} x^{-(1+\alpha)} dx$$
$$= \frac{2C_\alpha}{\alpha} \lceil p/2 \rceil^{-\alpha}$$
$$\leq \frac{2^{\alpha+1} C_\alpha}{\alpha p^\alpha}.$$

□

Corollary 10. *Suppose that a minimizer of R over $\{Z \in \mathcal{U}^p : K_{ZZ} \succ 0\}$ exists, and that the compact subset $\mathcal{Z}_p \subset \mathcal{U}^p$ contains one of them:*

$$\mathcal{Z}_p \cap \arg \min_{Z \in \mathcal{U}^p,\, K_{ZZ} \succ 0} R(Z) \neq \emptyset.$$

By Lemma 9, any such minimizer satisfies $R(Z) \leq 2^{\alpha+1} C_\alpha / (\alpha p^\alpha)$. Now let $\hat{Z}_{\mathrm{tr}}$ be a point satisfying $\hat{Z}_{\mathrm{tr}} \in \arg\min_{Z \in \mathcal{Z}_p} \hat{R}_{M_0,n}(Z)$. Existence of $\hat{Z}_{\mathrm{tr}}$ is guaranteed as $\mathcal{Z}_p$ is compact and $Z \mapsto \hat{R}_{M_0,n}$ is continuous. Then, on the event of Theorem 8, writing $\Delta_{M_0,n}(\epsilon, \delta) := \sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{Z}_p) + \log(4/\delta)}{2M_0}} + \sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{Z}_p) + \log(4/\delta)}{2M_0 n}} + 4L_g \epsilon$, we have that

$$R(\hat{Z}_{\mathrm{tr}}) \leq \frac{2^{\alpha+1} C_\alpha}{\alpha p^\alpha} + 2\Delta_{M_0,n}(\epsilon, \delta).$$

Additionally, if a new task S is drawn independent to the pre-training data and $\|f_S^*\|_{\mathcal{H}_k} \leq B$ for some constant $B > 0$, then

$$\mathbb{E}_S \left[\|r_{S,\hat{Z}_{\mathrm{tr}}}\|^2_{L^2(P_{s|S})} | \mathcal{F}_{\mathrm{VI}} \right] \leq B^2 \left[\frac{2^{\alpha+1} C_\alpha}{\alpha p^\alpha} + 2\Delta_{M_0,n}(\epsilon, \delta) \right].$$

Proof. On the event of the Theorem 8 gives, simultaneously for every $Z \in \mathcal{Z}_p$,

$$\left| R(Z) - \widehat{R}_{M_0,n}(Z) \right| \leq \Delta_{M_0,n}(\epsilon, \delta).$$

Therefore,

$$R(\widehat{Z}_{\mathrm{tr}}) \leq \widehat{R}_{M_0,n}(\widehat{Z}_{\mathrm{tr}}) + \Delta_{M_0,n}(\epsilon, \delta)$$
$$= \inf_{Z \in \mathcal{Z}_p} \widehat{R}_{M_0,n}(Z) + \Delta_{M_0,n}(\epsilon, \delta)$$
$$\leq \inf_{Z \in \mathcal{Z}_p} R(Z) + 2\Delta_{M_0,n}(\epsilon, \delta)$$
$$= \inf_{\substack{Z \in \mathcal{U}^p \\ K_{ZZ} \succ 0}} R(Z) + 2\Delta_{M_0,n}(\epsilon, \delta)$$
$$\leq \frac{2^{\alpha+1} C_\alpha}{\alpha} p^{-\alpha} + 2\Delta_{M_0,n}(\epsilon, \delta).$$

Lemma 6 gives

$$\mathbb{E}_S \left[\|r_{S,\hat{Z}_{\mathrm{tr}}}\|^2_{L^2(P_{\mathrm{s}|S})} \mid \mathcal{F}_{\mathrm{VI}} \right] \leq B^2 R(\hat{Z}_{\mathrm{tr}}),$$

and substitution proves desired inequality. □

Lemma 11. *Assuming $\mathcal{U} \subset \mathbb{R}^d$ and $\mathcal{Z}_p \subset (B_d(0, R))^p$, for every $\epsilon > 0$,*

$$\log \mathcal{N}(\epsilon, \mathcal{Z}_p, \|\cdot\|_F) \leq pd \log \left(1 + \frac{2R\sqrt{p}}{\epsilon} \right)$$

where $d := \dim(\mathcal{U})$. Consequently, if $L_g > 0$, then for every integer $n \geq 1$ and $\epsilon_n = (8L_g n)^{-1}$,

$$\log \mathcal{N}(\epsilon_n, \mathcal{Z}_p) \leq pd \log \left[1 + 32RL_k n \left(\frac{\kappa^2 p}{\Lambda} + \frac{\kappa^4 p^2}{\Lambda^2} \right) \right].$$

Consequently, if $R, \kappa, L_k = O(1)$ and Λ^{-1} grows at most polynomially in p and n, then

$$\log \mathcal{N}(\epsilon_n, \mathcal{Z}_p) + \log \frac{4}{\delta} = \tilde{O}\left(pd + \log(1/\delta) \right)$$

Proof. Identify each inducing points with a vector in $\mathbb{R}^{pd}$. Since $\mathcal{Z}_p \subseteq (B_d(0,R))^p$, it is contained in the Euclidean ball of radius $R\sqrt{p}$ in $\mathbb{R}^{pd}$. Lemma A.1 gives an internal net with

$$\mathcal{N}(\epsilon, \mathcal{Z}_p, \|\cdot\|_F) \le \left(1 + \frac{2R\sqrt{p}}{\epsilon}\right)^{pd}.$$

Taking logarithms proves the first assertion. Substituting $\epsilon_n = (8L_g n)^{-1}$ and

$$L_g = \frac{2\kappa^2 L_k \sqrt{p}}{\Lambda} + \frac{2\kappa^4 L_k p\sqrt{p}}{\Lambda^2}$$

gives the second. Under the stated growth condition, the logarithm in the covering bound is $O(\log(2pn))$, which gives the last assertion after absorbing logarithmic factors. □

Remark. By setting $\epsilon = (8L_g n)^{-1}$, combining the results of Corollary 10 and Lemma 11 yields

$$\mathbb{E}_S\left[\|r_{S,\hat{Z}_{\mathrm{tr}}}\|^2_{L^2(P_{s|S})} | \mathcal{F}_{\mathrm{VI}}\right] = \tilde{O}\left(p^{-\alpha} + \sqrt{\frac{p\dim(\mathcal{U}) + \log(1/\delta)}{M_0}}\left(1 + \frac{1}{\sqrt{n}}\right) + \frac{1}{n}\right).$$

C Derivation of the shared head initialization

In this section we derive the criterion in (8) and interpret one of its statistical term as a weighted least-squares objective across tasks.

C.1 Motivation and finite-step surrogate

Fix the representation. Note that we write $f_\alpha = k(\cdot, Z)\alpha, \alpha = K_{ZZ}^{-1/2}\beta$ as in Section 3.1. Suppose we consider the MAML objective (Finn et al., 2017) with η as the inner loop learning rate. Let $\mathcal{L}_S$ denote the common population squared loss. When $P_{s|S} = P_{q|S}$, for each task, its first-order expansion as $\eta \to 0$ is

$$\mathcal{L}_S\big(f_{\alpha - \eta\nabla_\alpha \mathcal{L}_S(f_\alpha)}\big) = \mathcal{L}_S(f_\alpha) - \eta\|\nabla_\alpha \mathcal{L}_S(f_\alpha)\|_2^2 + O(\eta^2).$$

This expansion motivates a negative squared-gradient term.

To relate the loss to an RKHS distance, suppose that $f_S^* \in \mathcal{H}_k$ and $k(x,x) \le 1$. The squared loss then satisfies

$$\mathcal{L}_S(f) - \mathcal{L}_S(f_S^*) = \frac{1}{2}\|f - f_S^*\|^2_{L^2(P_{q|S})} \le \frac{1}{2}\|f - f_S^*\|^2_{\mathcal{H}_k}.$$

Since f_S^* is unavailable, we suggest a computable surrogate by replacing it with the E-step adapted predictor via functional gradient descent. The substituted predictor depends on the initial coefficient β. Using (6), write

$$\beta_{m,E} = \mathcal{A}_{m,E}(\beta) = A_{m,E}\beta + d_{m,E},$$
$$A_{m,E} = (I - \tau H_m)^E, \qquad d_{m,E} = \tau\sum_{t=0}^{E-1}(I - \tau H_m)^t b_m.$$

The resulting squared RKHS displacement is defined as

$$\left\|k(\cdot, Z)K_{ZZ}^{-1/2}(\beta - \beta_{m,E})\right\|^2_{\mathcal{H}_k} = \|(I - A_{m,E})\beta - d_{m,E}\|_2^2.$$

Its average across tasks is the adaptation energy in Lemma 18, which controls the optimization-error bound under the assumptions of that lemma.

We combine this displacement term with the empirical query-gradient term. Since

$$\nabla_\alpha \widehat{L}_{D_m^q}(f_{\vartheta,\beta}) = K_{ZZ}^{1/2}(G_m\beta - g_m),$$

the criterion takes the form

$$\widehat{\mathcal{E}}_{\mathrm{RKHS}}(\beta) = \frac{1}{2M}\sum_{m=1}^{M}\Big[\|(I - A_{m,E})\beta - d_{m,E}\|_2^2 \\ - \eta(G_m\beta - g_m)^\top K_{ZZ}(G_m\beta - g_m)\Big].$$

Here η is interpreted as the relative weight of the query-gradient term. The population expansion motivates its form; it does not establish that minimizing this finite-step surrogate minimizes the post-adaptation query risk.

Quadratic expansion. Direct expansion gives

$$\widehat{\mathcal{E}}_{\mathrm{RKHS}}(\beta) = \frac{1}{2}\beta^\top P_{M,E}\beta - q_{M,E}^\top\beta + C,$$

where C is independent of β and

$$P_{M,E} = \frac{1}{M}\sum_{m=1}^{M}\big[(I - A_{m,E})^\top(I - A_{m,E}) - \eta G_m K_{ZZ} G_m\big],$$
$$q_{M,E} = \frac{1}{M}\sum_{m=1}^{M}\big[(I - A_{m,E})d_{m,E} - \eta G_m K_{ZZ} g_m\big].$$

We used the symmetry of $A_{m,E}$ and G_m. If $P_{M,E}$ is nonsingular, the undamped criterion has the unique stationary point $P_{M,E}^{-1}q_{M,E}$; this point is the unique minimizer when $P_{M,E} \succ 0$.

Adding $\gamma\beta^\top K_{ZZ}^{-1}\beta/2$ gives the stationary point

$$\beta^\star = (P_{M,E} + \gamma K_{ZZ}^{-1})^{-1}q_{M,E},$$

whenever the coefficient matrix is nonsingular. It is the unique minimizer when $P_{M,E} + \gamma K_{ZZ}^{-1} \succ 0$. The experiments use $\gamma = 10^{-4}$.

Interpretation of the displacement term. The displacement term also admits an interpretation in terms of the least-squares solutions. Since $b_m \in \mathrm{range}(H_m)$, let $\widehat{\beta}_m = H_m^\dagger b_m$ be the minimum-norm least-squares head. Then

$$d_{m,E} = (I - A_{m,E})\widehat{\beta}_m.$$

Consequently,

$$\frac{1}{2M}\sum_{m=1}^{M}\|(I - A_{m,E})\beta - d_{m,E}\|_2^2 \\ = \frac{1}{2M}\sum_{m=1}^{M}(\beta - \widehat{\beta}_m)^\top W_{m,E}(\beta - \widehat{\beta}_m), \qquad W_{m,E} := (I - A_{m,E})^2.$$

For $\eta = \gamma = 0$, the initialization therefore minimizes a weighted sum of squared distances to the task-specific least-squares heads. If $\sum_m W_{m,E} \succ 0$, the unique minimizer is

$$\beta^\star = \left(\sum_{m=1}^{M} W_{m,E}\right)^{-1}\sum_{m=1}^{M} W_{m,E}\widehat{\beta}_m.$$

This is a matrix-weighted average of the task heads. If the sum is singular, the objective leaves some directions undetermined, so the minimizer is not unique. Replacing the inverse by the Moore–Penrose inverse gives the minimum-norm minimizer.

The weights follow from the support updates. Along an eigenvector of H_m with eigenvalue $\lambda_{m,j}$, the difference from $\widehat{\beta}_m$ after E steps is $(1 - \tau\lambda_{m,j})^E$ times the initial difference. The difference

between the initial and adapted coefficients is therefore $1-(1-\tau\lambda_{m,j})^E$ times that initial difference. Squaring this displacement gives the weight

$$\left[1-(1-\tau\lambda_{m,j})^E\right]^2.$$

When $0 \le \tau\lambda_{m,j} \le 1$, a larger weight means that the E support updates remove a larger fraction of the initial difference. For the same initial difference, these directions produce a larger displacement and thus contribute more to the objective. A zero eigenvalue gives zero weight because the support updates leave that direction unchanged.

If every $W_{m,E}$ were I, the expression for $\beta^\star$ would reduce to the ordinary mean of the unregularized least-squares heads. The task-mean baseline in Appendix F.5 averages ridge-regression heads. The weighted-average interpretation above applies to the displacement term alone; the full proposed initialization also includes the query-gradient term and damping.

D Proof of the Conditional Excess Transfer Risk Bound

We use the sampling and conditioning conventions in Appendix A.2. Throughout this section, the final representation is fixed conditional on $\mathcal{F}_{\text{train}}$, and the risk is conditional on $\mathcal{G}_S$. The population coefficient $\bar\beta_S$ and residual r_S are the query projection defined there.

As H_S is symmetric, we diagonalize it as $H_S = U_S\Lambda_S U_S^\top$ with $\Lambda_S = \mathrm{diag}(\lambda_{\mathrm{S},1},\cdots,\lambda_{\mathrm{S,p}})$, $\lambda_{S,i} \ge 0$. With this, the recursion formula of the $t+1$ th step functional gradient descent given by $\beta_{S,t+1} = (I-\tau H_S)\beta_{S,t} + \tau b_S$ becomes

$$\theta_{S,t+1}(i) = (1-\tau\lambda_{S,i})\theta_{S,t}(i) + \tau\mu_{S,i}, \quad i = 1,\cdots,p$$

where $\theta_{S,t} = U_S^\top \beta_{S,t}$ and $\mu_S = U_S^\top b_S$. We also write the error relative to $\bar\beta_S$ as

$$e_{S,t}(i) := \theta_{S,t}(i) - \bar\theta_S(i), \quad \bar\theta_S = U_S^\top\bar\beta_S, \quad i = 1,\cdots,p$$

Lemma 12. *For each task S and its corresponding support Hessian matrix H_S, the empirical update through the functional gradient descent satisfies the recursion formula*

$$e_{S,t+1}(i) = (1-\tau\lambda_{S,i})e_{S,t}(i) + \tau\xi_{S,i}$$

where the drift term $\xi_{S,i}$ is defined to the ith coordinate of vector ξ_S given by

$$\xi_S = \mu_S - \Lambda_S\bar\theta_S = U_S^\top(b_S - H_S\bar\beta_S) \in \mathbb{R}^p, \quad \xi_S = (\xi_{S,1},\cdots,\xi_{S,p}).$$

Moreover, its conditional second moment satisfies

$$\mathbb{E}\left[\xi_{S,i}^2 \mid \mathcal{G}_S\right] \le \frac{\lambda_{S,i}}{n}\left(\|r_S(X_S)\|_2^2 + \sigma^2\right)$$

Proof. First, we note that the matrix H_S can be written as

$$H_S = \frac{1}{n}\Phi_S^\top\Phi_S, \quad \Phi_S := K(X_S,\mathrm{Z})K_{ZZ}^{-1/2}$$

Consequently, from the definition we have

$$\Phi_S\bar\beta_S = K(X_S,\mathrm{Z})K_{ZZ}^{-1/2}\bar\beta_S = \bar f_S(X_S)$$

so that

$$\begin{aligned}
\xi_S = U_S^\top(b_S - H_S\bar\beta_S) &= U_S^\top\left(b_S - \frac{1}{n}\Phi_S^\top\Phi_S\bar\beta_S\right)\\
&= U_S^\top\left(\frac{1}{n}\Phi_S^\top Y_S - \frac{1}{n}\Phi_S^\top\bar f_S(X_S)\right)\\
&= U_S^\top\frac{1}{n}\Phi_S^\top\left(Y_S - (f_S^*(X_S) - r_S(X_S))\right)\\
&= U_S^\top\frac{1}{n}\Phi_S^\top\left(\varepsilon_S + r_S(X_S)\right).
\end{aligned}$$

Let $u_{S,i}$ be the ith column of U_S and define the feature-response vector

$$v_{S,i} := \Phi_S u_{S,i} \in \mathbb{R}^n.$$

Then

$$\xi_{S,i} = \frac{1}{n} u_{S,i}^\top \Phi_S^\top (r_S(X_S) + \varepsilon_S) = \frac{1}{n} v_{S,i}^\top (r_S(X_S) + \varepsilon_S) = \frac{1}{n} \langle v_{S,i},\, r_S(X_S) + \varepsilon_S \rangle.$$

Also note that

$$\|v_{S,i}\|_2^2 = \|\Phi_S u_{S,i}\|_2^2 = u_{S,i}^\top \Phi_S^\top \Phi_S u_{S,i} = n\, u_{S,i}^\top H_S u_{S,i} = n\lambda_{S,i}.$$

Now taking the conditional expectation, applying the Cauchy-Schwarz inequality and the noise condition $\mathbb{E}[\varepsilon_S \varepsilon_S^\top | \mathcal{G}_S] \preceq \sigma^2 I$,

$$\begin{aligned}
\mathbb{E}\left[\xi_{S,i}^2 \mid \mathcal{G}_S\right] &\le \frac{1}{n^2}\|v_{S,i}\|_2^2 \|r_S(X_S)\|_2^2 + \frac{1}{n^2}\mathbb{E}\left[\langle v_{S,i}, \varepsilon_S\rangle^2 | \mathcal{G}_S\right] \\
&= \frac{1}{n^2}\|v_{S,i}\|_2^2 \|r_S(X_S)\|_2^2 + \frac{1}{n^2} v_{S,i}^\top \mathbb{E}[\varepsilon_S \varepsilon_S^\top | \mathcal{G}_S] v_{S,i} \\
&\le \frac{1}{n^2}\|v_{S,i}\|_2^2 \|r_S(X_S)\|_2^2 + \frac{1}{n^2}\sigma^2 \|v_{S,i}\|_2^2 \\
&= \frac{\lambda_{S,i}}{n}\left(\|r_S(X_S)\|_2^2 + \sigma^2\right).
\end{aligned}$$

□

We now obtain a spectral bias-variance decomposition around $\bar{f}_S$.

Lemma 13. *For every integer $t \ge 0$ and every $i \in I_S := \{i : \lambda_{S,i} > 0\}$, we have that for each task S*

$$\mathbb{E}[e_{S,t}^2(i) | \mathcal{G}_S] \le 2(1-\tau\lambda_{S,i})^{2t} e_{S,0}^2(i) + \frac{2}{n}\left(\|r_S(X_S)\|_2^2 + \sigma^2\right) \cdot \frac{(1-(1-\tau\lambda_{S,i})^t)^2}{\lambda_{S,i}}.$$

Summing over i we have

$$\begin{aligned}
\mathbb{E}\left[\|\beta_t - \bar{\beta}_S\|_2^2 | \mathcal{G}_S\right] \le\; & 2\sum_{i=1}^{p} (1-\tau\lambda_{S,i})^{2t} e_{S,0}^2(i) \\
& + \frac{2}{n}\left(\|r_S(X_S)\|_2^2 + \sigma^2\right) \sum_{i \in I_S} \frac{(1-(1-\tau\lambda_{S,i})^t)^2}{\lambda_{S,i}}.
\end{aligned}$$

Proof. First, we claim that in general it is enough to consider the case where $\lambda_{S,i} > 0$. Indeed, if $\lambda_{S,i} = 0$, then $v_{S,i} = 0$ and hence $\xi_{S,i} = 0$. The recursion therefore gives $e_{S,t}(i) = e_{S,0}(i)$, or equivalently $\Pi_{S,0}\beta_{S,t} = \Pi_{S,0}\widehat{\beta}_0$. These coordinates contribute only their unchanged initial error. Now from the recursion $e_{S,t+1}(i) = (1-\tau\lambda_{S,i})e_{S,t}(i) + \tau\xi_{S,i}$ gives the closed form of $e_{S,t}(i)$ as

$$\begin{aligned}
e_{S,t}(i) &= (1-\tau\lambda_{S,i})^t e_{S,0}(i) + \tau \sum_{j=0}^{t-1} (1-\tau\lambda_{S,i})^j \xi_{S,i} \\
&= (1-\tau\lambda_{S,i})^t e_{S,0}(i) + \frac{1-(1-\tau\lambda_{S,i})^t}{\lambda_{S,i}} \xi_{S,i}
\end{aligned}$$

Now using that $(a+b)^2 \le 2a^2 + 2b^2$,

$$e_{S,t}^2(i) \le 2(1-\tau\lambda_{S,i})^{2t} e_{S,0}^2(i) + 2\left(\frac{1-(1-\tau\lambda_{S,i})^t}{\lambda_{S,i}}\right)^2 \xi_{S,i}^2$$

Now taking the conditional expectation and using that $\mathbb{E}\left[\xi_{S,i}^2 \mid \mathcal{G}_S\right] \le \frac{\lambda_{S,i}}{n}\left(\|r_S(X_S)\|_2^2 + \sigma^2\right)$,

$$\mathbb{E}[e_{S,t}^2(i) | \mathcal{G}_S] \le 2(1-\tau\lambda_{S,i})^{2t} e_{S,0}^2(i) + \frac{2}{n}\left(\|r_S(X_S)\|_2^2 + \sigma^2\right) \cdot \frac{(1-(1-\tau\lambda_{S,i})^t)^2}{\lambda_{S,i}}.$$

Then bound of $\mathbb{E}\left[\|\beta_t - \bar{\beta}_S\|_2^2 | \mathcal{G}_S\right]$ follows by summing over i, as $\|\beta_t - \bar{\beta}_S\|_2^2 = \|U_S^\top e_{S,t}\|_2^2 = \sum_{i=1}^p e_{S,t}^2(i)$. □

Theorem 14. *Define the support-conditional (conditional on $\mathcal{G}_S$) excess risk after T steps of functional gradient descent initialized at the point $f(\cdot) = k(\cdot, Z)K_{ZZ}^{-1/2}\hat{\beta}_0$, as*

$$\mathcal{R}_S^{sc}(T) = \mathbb{E}_\varepsilon\left[\mathcal{L}_S(f_{\beta_{S,T}}) - \mathcal{L}_S(f_S^*)|\mathcal{G}_S\right].$$

Let $I_S := \{i : \lambda_{S,i} > 0\}$ and $\Pi_{S,0}$ denote the projection to $\ker(H_S)$. Then

$$\mathcal{R}_S^{sc}(T) \le \frac{1}{2}\|r_S\|_{L^2(P_{q|S})}^2 + \frac{1}{2}\|\Pi_{S,0}(\hat{\beta}_0 - \bar{\beta}_S)\|_2^2 + \sum_{i\in I_S}(1-\tau\lambda_{S,i})^{2T}e_{S,0}^2(i) + \frac{1}{n}\left(\|r_S(X_S)\|_2^2 + \sigma^2\right)\sum_{i\in I_S}\frac{(1-(1-\tau\lambda_{S,i})^T)^2}{\lambda_{S,i}}$$

Proof. From the definition,

$$\mathcal{L}_S(f) = \frac{1}{2}\mathbb{E}_{(X,Y)\sim P_{q|S}}\left[(f(X)-Y)^2\right], \quad Y = f_S^*(X) + \varepsilon$$

so that

$$\begin{aligned}\mathcal{L}_S(f) &= \frac{1}{2}\mathbb{E}\left[(f(X) - f_S^*(X) - \varepsilon)^2\right] \\ &= \frac{1}{2}\|f - f_S^*\|_{L^2(P_{q|S})}^2 - \mathbb{E}\left[(f(X) - f_S^*(X))\cdot\varepsilon\right] + \frac{1}{2}\mathbb{E}\left[\varepsilon^2\right]\end{aligned}$$

Conditioned on $\mathcal{G}_S$,

$$\begin{aligned}&\mathbb{E}_\varepsilon\left[\mathbb{E}_{P_{q|S}}\left[(f(X) - f_S^*(X))\cdot\varepsilon\right]|\mathcal{G}_S\right] \\ &= \mathbb{E}_\varepsilon\left[\mathbb{E}_{P_{q|S}}\left[(f(X) - f_S^*(X))\underbrace{\mathbb{E}[\varepsilon \mid X, S, \mathcal{F}_{\text{train}}, D_S]}_{=0 \text{ for query noise}}\right]\Big|\,\mathcal{G}_S\right] = 0.\end{aligned}$$

and that $\mathcal{L}_S(f_S^*) = \frac{1}{2}\mathbb{E}\left[((f_S^*(X) - Y)^2\right] = \frac{1}{2}\mathbb{E}\left[\varepsilon^2\right]$, we have that

$$\mathcal{R}_S^{sc}(T) = \mathbb{E}_\varepsilon\left[L_S(f_{\beta_{S,T}}) - L_S(f_S^*)|\mathcal{G}_S\right] = \mathbb{E}_\varepsilon\left[\frac{1}{2}\|f_{\beta_{S,T}} - f_S^*\|_{L^2(P_{q|S})}^2 \mid \mathcal{G}_S\right]$$

As $\Sigma_S = \mathbb{E}_{X\sim P_{q|S}}\left[\phi(X)\phi(X)^\top\right] \preceq I$ holds under condition that $\kappa \le 1$,

$$\begin{aligned}\frac{1}{2}\|f_{\beta_{S,T}} - f_S^*\|_{L^2(P_{q|S})}^2 &= \frac{1}{2}\|r_S\|_{L^2(P_{q|S})}^2 + \frac{1}{2}(\beta_{S,T} - \bar{\beta}_S)^\top\Sigma_S(\beta_{S,T} - \bar{\beta}_S) \\ &\le \frac{1}{2}\|r_S\|_{L^2(P_{q|S})}^2 + \frac{1}{2}\|\beta_{S,T} - \bar{\beta}_S\|_2^2\end{aligned}$$

Now as $\Pi_{S,0}\beta_{S,T} = \Pi_{S,0}\hat{\beta}_0$ under the functional gradient descent update, and that

$$\sum_{i\notin I_S} e_{S,T}(i)^2 = \|\Pi_{S,0}(\hat{\beta}_0 - \bar{\beta}_S)\|_2^2,$$

on taking the conditional expectation on $\mathcal{G}_S$

$$\begin{aligned}\mathcal{R}_S^{sc}(T) &= \mathbb{E}_\varepsilon\left[\frac{1}{2}\|f_{\beta_{S,T}} - f_S^*\|_{L^2(P_{q|S})}^2 \mid \mathcal{G}_S\right] \\ &\le \frac{1}{2}\|r_S\|_{L^2(P_{q|S})}^2 + \frac{1}{2}\sum_{i\notin I_S} e_{S,T}(i)^2 + \frac{1}{2}\mathbb{E}\left[\sum_{i\in I_S} e_{S,T}(i)^2|\mathcal{G}_S\right] \\ &\le \frac{1}{2}\|r_S\|_{L^2(P_{q|S})}^2 + \frac{1}{2}\|\Pi_{S,0}(\hat{\beta}_0 - \bar{\beta}_S)\|_2^2 + \frac{1}{2}\mathbb{E}\left[\sum_{i\in I_S} e_{S,T}(i)^2|\mathcal{G}_S\right] \\ &\le \frac{1}{2}\|r_S\|_{L^2(P_{q|S})}^2 + \frac{1}{2}\|\Pi_{S,0}(\hat{\beta}_0 - \bar{\beta}_S)\|_2^2 \\ &\quad + \sum_{i\in I_S}(1-\tau\lambda_{S,i})^{2T}e_{S,0}^2(i) + \frac{1}{n}\left(\|r_S(X_S)\|_2^2 + \sigma^2\right)\sum_{i\in I_S}\frac{(1-(1-\tau\lambda_{S,i})^T)^2}{\lambda_{S,i}}\end{aligned}$$

where the last inequality follows from Lemma 13. □

Corollary 15. *Suppose that* $1-\tau\lambda_{S,i}=\rho_S\in(0,1)$ *for every* $i\in I_S$ *and* $V_S>0$*. Then*

$$\mathcal{R}_S^{sc}(T)\le F_S+B_S\rho_S^{2T}+V_S(1-\rho_S^T)^2$$

where $F_S:=\frac{1}{2}\|r_S\|^2_{L^2(P_{q|S})}+\frac{1}{2}\|\Pi_{S,0}(\hat{\beta}_0-\bar{\beta}_S)\|_2^2$, $B_S:=\sum_{i\in I_S}e^2_{S,0}(i)$ *and* $V_S:=\frac{\|r_S(X_S)\|_2^2+\sigma^2}{n}\cdot\frac{\tau|I_S|}{1-\rho_S}$ *stand for the constant coefficients (independent of* T*) appearing on the conditional excess risk bound each representing the floor, bias and the variance term, respectively. The continuous upper bound is minimized at* $T^*:=\frac{\log(1+B_S/V_S)}{\log(1/\rho_S)}$*, with value*

$$\min_{T\ge0}\{F_S+B_S\rho_S^{2T}+V_S(1-\rho_S^T)^2\}=F_S+\frac{B_SV_S}{B_S+V_S}\le F_S+\min\{B_S,V_S\}.$$

Proof. The upper bound comes directly from the definition of F_S, B_S and V_S under the assumption that $1-\tau\lambda_{S,i}=\rho_S$. Now expanding the upper bound and differentiating by T,

$$\begin{aligned}\frac{d}{dT}\left(F_S+B_S\rho_S^{2T}+V_S(1-\rho_S^T)^2\right)&=\frac{d}{dT}\left((B_S+V_S)\rho_S^{2T}-2V_S\rho_S^T+V_S+F_S\right)\\&=2\log\rho_S\left[(B_S+V_S)\rho_S^{2T}-V_S\rho_S^T\right]\end{aligned}$$

Then solving for $2\log\rho_S\left[(B_S+V_S)\rho_S^{2T}-V_S\rho_S^T\right]=0$, we have that $\rho_S^T=\frac{V_S}{B_S+V_S}$. Hence, $T^*=\frac{\log(1+B_S/V_S)}{\log(1/\rho_S)}$ is the solution of the above, and substituting T^* into the bound gives $F_S+\frac{B_SV_S}{B_S+V_S}$. An integer bound minimizer is obtained by comparing the values at $\lfloor T^*\rfloor$ and $\lceil T^*\rceil$. □

Remark. The conditional bound separates the stopping depth from the residual of the chosen representation. The reference residual estimate in Corollary 10 can be applied to the final representation only under the additional comparison in (23). However, this comparison is not implied by pre-training via full ELBO form objective.

E Proof of the Excess Transfer Risk bound

In this section we derive the uniform excess transfer risk bound after T steps of functional gradient descent, conditional on $\mathcal{F}_{\mathrm{VI}}$ and averaged over meta-training and a fresh task and its support dataset. Recall the definition of the conditional excess transfer risk,

$$\mathcal{R}_S^{sc}(T)=\mathbb{E}_{\varepsilon_S}\left[L_S(f_{\beta_{S,T}})-L_S(f_S^*)|\mathcal{G}_S\right].$$

Throughout this section, we additionally assume that the support and query data distribution coincides: $P_{s|S}\equiv P_{q|S}$. Conditional on $\mathcal{F}_{\mathrm{VI}}$ the meta-training tasks are sampled independently from the same task distribution. The datasets D_m and D_S appearing in the empirical losses below are the support datasets used for adaptation.

Assume that the kernel satisfies $K_{ZZ}\succ0$ and $k(x,x)\le1$, and that

$$|Y|\le B_y,\quad \|\hat{\beta}_0\|_2\le R_0,\quad \|\bar{\beta}_S\|_2\le B,\ \text{and}\quad 0<\tau\le1.$$

The noise and independence conventions are those in Appendix A.2; assume $\sigma^2>0$ in the logarithmic bounds. We define the environment-averaged excess transfer risk as follows:

$$\mathcal{R}^u(T):=\mathbb{E}\left[L_S(f_{\beta_{S,T}})-L_S(f_S^*)\mid\mathcal{F}_{\mathrm{VI}}\right]$$

This expectation includes all meta-training randomness and the fresh task and support sample, conditional on $\mathcal{F}_{\mathrm{VI}}$. Probability statements for the resulting bound concern pre-training. Let D_S be the unseen data sample of task S and $D_1,\cdots,D_M$ as the meta-training support samples of n observations each, respectively. If $\hat{L}_{D_m}(f_\beta)=\frac{1}{2n}\sum_{i=1}^n(f_\beta(X_{m,i})-Y_{m,i})^2$ denotes the empirical loss on the data sample D_m of task m, the excess transfer risk can be further decomposed as (abbreviating the expectation symbol as $\mathbb{E}(\cdot)$):

$$\mathcal{R}^{\mathrm{u}}(T)=\underbrace{\mathbb{E}[L_S(f_{\beta_{S,T}})-\widehat{L}_{D_S}(f_{\beta_{S,T}})]}_{:=\mathcal{R}_{\mathrm{within}}(T)}$$

$$+\underbrace{\mathbb{E}\left[\widehat{L}_{D_S}(f_{\beta_{S,T}}) - \frac{1}{M}\sum_{m=1}^{M}\widehat{L}_{D_m}(f_{\beta_{m,T}})\right]}_{:=\mathcal{R}_{\text{across}}(T)}$$

$$+\underbrace{\mathbb{E}\frac{1}{M}\sum_{m=1}^{M}[\widehat{L}_{D_m}(f_{\beta_{m,T}}) - \widehat{L}_{D_m}(f_{\widehat{\beta}_{D_m}})]}_{:=\mathcal{O}(T)}$$

$$+\underbrace{\mathbb{E}\frac{1}{M}\sum_{m=1}^{M}[\widehat{L}_{D_m}(f_{\widehat{\beta}_{D_m}}) - \widehat{L}_{D_m}(f^{\star}_{S_m})]}_{:=\mathcal{A}^{\star}_{\text{sub}}}$$

$$+\underbrace{\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\widehat{L}_{D_m}(f^{\star}_{S_m}) - L_S(f^{\star}_S)\right]}_{=0}. \tag{20}$$

Lemma 16. *(Within-task generalization.) Under the common assumptions of Appendix E, initialize adaptation from $\widehat{\beta}_0$ with $\|\widehat{\beta}_0\|_2 \le R_0$. Let $\rho_p = \mathbb{E}[\|r_S\|^2_{L^2(P_{s|S})} \mid \mathcal{F}_{\text{VI}}]$, including meta-training randomness in this expectation. Then, for every integer $T \ge 1$,*

$$\mathcal{R}_{\text{within}}(T) \le \frac{(1+B_y^2)\left[1+(R_0+B)^2\right]}{\sqrt{n}} + 4(1+B_y^2)\sqrt{1+\log\max\left\{1, \frac{(B_y+B)^2}{\sigma^2/n}\right\}}\left(\frac{\tau T\rho_p}{\sqrt{n}} + \frac{\tau^2T^2\sigma^2}{n^{3/2}}\right).$$

If $\rho_p \lesssim p^{-\alpha} + n^{-1/2}$ given that M_0 is sufficiently large, then

$$\mathcal{R}_{\text{within}}(T) \lesssim \frac{1}{\sqrt{n}} + \sqrt{\log(n)}\left(\frac{T}{p^{\alpha}\sqrt{n}} + \frac{T}{n} + \frac{T^2}{n^{3/2}}\right).$$

Proof. Let $D_S = \{(X_i, Y_i)\}_{i=1}^n$ denote the support dataset of task S. Then writing

$$\Phi_S = \begin{pmatrix}\phi(X_1)^\top \\ \vdots \\ \phi(X_n)^\top\end{pmatrix}, \quad H_S = \frac{1}{n}\Phi_S^\top\Phi_S, \quad A_S = I - \tau H_S,$$

Let $\bar{\beta}_S$ be the population projected coefficient defined in Appendix A.2. Then $Y = \Phi_S\bar{\beta}_S + r + \varepsilon_S$ with $r_i = f^*_S(X_i) - \phi(X_i)^\top\bar{\beta}_S$. From the recursion formula of the functional gradient descent, the T-step head of the function for task S over the induced space becomes

$$\beta_{S,T} = \underbrace{A_S^T\hat{\beta}_0 + (I - A_S^T)\bar{\beta}_S}_{:=a_S} + \underbrace{\frac{\tau}{n}\sum_{t=0}^{T-1}A_S^t\Phi_S^\top(r+\varepsilon_S)}_{:=u_S}$$

It is straightforward that $\|a_S\|_2 \le \|A_S^T\hat{\beta}_0\|_2 + \|(I - A_S^T)\bar{\beta}_S\|_2 \le R_0 + B$ as A_S^T and $I - A_S^T$ are contractions. Now to bound the term u_S, consider the eigen-decomposition of the matrix $H_S = U_S\text{diag}(\lambda_{\text{S,1}}, \cdots, \lambda_{\text{S,p}})U_S^\top$, where $U_S = (u_{S,1}, \cdots, u_{S,p})$ is the matrix of eigenvectors. For $\lambda_{S,j} > 0$,

$$u_{S,j}^\top u_S = \frac{1-(1-\tau\lambda_{S,j})^T}{\lambda_{S,j}}u_{S,j}^\top\frac{1}{n}\Phi_S^\top(r+\varepsilon_S)$$

Writing $\bar{v}_j = \frac{\Phi_S u_{S,j}}{\sqrt{n\lambda_{S,j}}}$ for $\lambda_{S,j} > 0$, $\{\bar{v}_j\}$ are orthonormal and hence

$$u_{S,j}^\top u_S = \frac{1-(1-\tau\lambda_{S,j})^T}{\lambda_{S,j}}\sqrt{\frac{\lambda_{S,j}}{n}}\bar{v}_j^\top(r+\varepsilon_S),$$

$$\|u_S\|_2^2 = \frac{1}{n} \sum_{\lambda_{S,j}>0} \frac{\left(1-(1-\tau\lambda_{S,j})^T\right)^2}{\lambda_{S,j}} (\bar{v}_j^\top (r+\varepsilon_S))^2.$$

The scalar bounds of Lemma A.9 apply. As $0 \le 1-(1-\tau\lambda_{S,j})^T \le 1, 1-(1-\tau\lambda_{S,j})^T \le T\tau\lambda_{S,j}$ by Bernoulli inequality and hence

$$\frac{\left(1-(1-\tau\lambda_{S,j})^T\right)^2}{\lambda_{S,j}} \le \min\{\tau T, \tau^2 T^2 \lambda_{S,j}\}$$

Note that $\lambda_{S,j} \le \operatorname{tr} H_S \le 1$ as $\operatorname{tr} H_S \le 1$. Applying the above with Bessel's inequality we have

$$\begin{aligned}\|u_S\|_2^2 &\le \frac{\min\{\tau T, \tau^2 T^2\}}{n} \sum_{\lambda_{S,j}>0} (\bar{v}_j^\top (r+\varepsilon_S))^2 \\ &\le \frac{\min\{\tau T, \tau^2 T^2\}}{n} \|r+\varepsilon_S\|_2^2 \\ &\le \min\{\tau T, \tau^2 T^2\}(B_y + B)^2.\end{aligned}$$

The last inequality follows from that $|r_i + \varepsilon_{S,i}| \le B_y + B$, as $r_i + \varepsilon_{S,i} = Y_i - \phi(X_i)^\top \bar{\beta}_S$. For the expectation bound, we condition on $\mathcal{G}_S$ and average over ε_S. The residual vector and the feature matrix are fixed under this conditioning. The notation $\mathbb{E}_{\varepsilon_S}$ below denotes this conditional expectation

$$\mathbb{E}_{\varepsilon_S}\left[(\bar{v}_j^\top (r+\varepsilon_S))^2\right] = (\bar{v}_j^\top r)^2 + \mathbb{E}_{\varepsilon_S}\left[(\bar{v}_j^\top \varepsilon_S)^2\right] \le (\bar{v}_j^\top r)^2 + \sigma^2$$

and hence

$$\begin{aligned}\mathbb{E}_{\varepsilon_S}\|u_S\|_2^2 &\le \frac{\tau T}{n} \sum_{\lambda_{S,j}>0} (\bar{v}_j^\top r)^2 + \frac{\sigma^2}{n} \sum_{\lambda_{S,j}>0} \frac{(1-(1-\tau\lambda_{S,j})^T)^2}{\lambda_{S,j}} \\ &\le \frac{\tau T}{n}\|r\|_2^2 + \frac{\tau^2 T^2 \sigma^2}{n}.\end{aligned}$$

The last step uses the trace sum in Lemma A.9. Averaging next over the support inputs, meta-training data, and task S,

$$\mathbb{E}\left[\frac{1}{n}\|r\|_2^2 \mid \mathcal{F}_{\mathrm{VI}}\right] = \mathbb{E}\left[\|r_S\|^2_{L^2(P_{s|S})} \mid \mathcal{F}_{\mathrm{VI}}\right] = \rho_p$$

we have that

$$\begin{aligned}\mathbb{E}\left[\|u_S\|_2^2 \mid \mathcal{F}_{\mathrm{VI}}\right] &\le \tau T \mathbb{E}\left[\|r_S\|^2_{L^2(P_{s|S})} \mid \mathcal{F}_{\mathrm{VI}}\right] + \frac{\tau^2 T^2 \sigma^2}{n} \\ &= \tau T \rho_p + \frac{\tau^2 T^2 \sigma^2}{n}.\end{aligned}$$

Now we define

$$V := \begin{pmatrix}\phi(X) \\ -Y\end{pmatrix}, \quad w_\beta = \begin{pmatrix}\beta \\ 1\end{pmatrix}$$

Then $w_\beta^\top V = \phi(X)^\top \beta - Y$ and hence

$$L_S(f_\beta) - \hat{L}_{D_S}(f_\beta) = \frac{1}{2} w_\beta^\top \left(\mathbb{E}\left[VV^\top \mid S, \mathcal{F}_{\mathrm{train}}\right] - \frac{1}{n}\sum_{i=1}^n V_i V_i^\top\right) w_\beta.$$

Let $\Delta_S := \frac{1}{2}\|\mathbb{E}\left[VV^\top \mid S, \mathcal{F}_{\mathrm{train}}\right] - \frac{1}{n}\sum_{i=1}^n V_i V_i^\top\|_F$. Then for all β,

$$\left[L_S(f_\beta) - \hat{L}_{D_S}(f_\beta)\right]_+ \le \Delta_S(1 + \|\beta\|_2^2)$$

Using that $\beta_{S,T} = a_S + u_S$,

$$\left[L_S(f_{\beta_{S,T}}) - \hat{L}_{D_S}(f_{\beta_{S,T}})\right]_+ \le 2\Delta_S\left[1 + (R_0 + B)^2 + \|u_S\|_2^2\right].$$

We use Lemma A.5 in the Hilbert space of matrices with the Frobenius norm, conditionally on $S, \mathcal{F}_{\text{train}}$. Since $\|VV^\top\|_F = \|V\|_2^2 \le 1 + B_y^2$, $\mathbb{E}\left[\Delta_S | S, \mathcal{F}_{\text{train}}\right] \le \frac{1+B_y^2}{2\sqrt{n}}$. Moreover, Δ_S differs at most by $\frac{1+B_y^2}{n}$ when a single observation is interchanged. Lemma A.4 therefore gives

$$\mathbb{P}\left(\Delta_S > \frac{1+B_y^2}{2\sqrt{n}} + t \mid S, \mathcal{F}_{\text{train}}\right) \le \exp\left(-\frac{2nt^2}{(1+B_y^2)^2}\right), \quad t \ge 0$$

As the right hand side does not depend on $\mathcal{F}_{\text{train}}$, by the tower property from $\mathcal{F}_{\text{VI}} \subset \mathcal{F}_{\text{train}}$

$$\mathbb{P}\left(\Delta_S > \frac{1+B_y^2}{2\sqrt{n}} + t \mid S, \mathcal{F}_{\text{VI}}\right) \le \exp\left(-\frac{2nt^2}{(1+B_y^2)^2}\right)$$

The tail bound is uniform in S, so averaging over the task gives the same inequality conditional only on $\mathcal{F}_{\text{VI}}$. We now apply Lemma A.10 with $F = \Delta_S$ and $U = \|u_S\|_2^2$. For $a := (1+B_y^2)/(2\sqrt{n})$,

$$\begin{aligned}\mathbb{E}\left[\Delta_S \|u_S\|_2^2 \mid \mathcal{F}_{\text{VI}}\right] &\le \mathbb{E}\left[(a + (\Delta_S - a)_+)\|u_S\|_2^2 \mid \mathcal{F}_{\text{VI}}\right] \\ &\le a\mathbb{E}\left[\|u_S\|_2^2 \mid \mathcal{F}_{\text{VI}}\right] + \int_0^\infty \mathbb{E}\left[\|u_S\|_2^2 \mathbf{1}\{\Delta_S > a + t\} | \mathcal{F}_{\text{VI}}\right] dt.\end{aligned}$$

Since

$$\begin{aligned}&\mathbb{E}\left[\|u_S\|_2^2 \mathbf{1}\{\Delta_S > a + t\} | \mathcal{F}_{\text{VI}}\right] \\ &\le \min\left\{\tau T \rho_p + \frac{\tau^2 T^2 \sigma^2}{n}, \min\{\tau T, \tau^2 T^2\}(B_y + B)^2 \exp\left(-\frac{2nt^2}{(1+B_y^2)^2}\right)\right\},\end{aligned}$$

taking

$$t^* = \frac{1+B_y^2}{\sqrt{2n}}\sqrt{1 + \log\max\left\{1, \left(\frac{(B_y+B)^2}{\sigma^2/n}\right)\right\}},$$

we have

$$\mathbb{E}\left[\|u_S\|_2^2 \mathbf{1}\{\Delta_S > a + t\} | \mathcal{F}_{\text{VI}}\right] \le \begin{cases} \tau T \rho_p + \frac{\tau^2 T^2 \sigma^2}{n} & 0 \le t \le t^*, \\ \min\{\tau T, \tau^2 T^2\}(B_y + B)^2 e^{-2nt^2/(1+B_y^2)^2} & t \ge t^*. \end{cases}$$

Integrating the expectation above yields

$$\begin{aligned}\int_0^\infty \mathbb{E}\left[\|u_S\|_2^2 \mathbf{1}\{\Delta_S > a + t\} | \mathcal{F}_{\text{VI}}\right] dt &= \int_0^{t^*} \mathbb{E}\left[\|u_S\|_2^2 \mathbf{1}\{\Delta_S > a + t\} | \mathcal{F}_{\text{VI}}\right] dt \\ &\quad + \int_{t^*}^\infty \mathbb{E}\left[\|u_S\|_2^2 \mathbf{1}\{\Delta_S > a + t\} | \mathcal{F}_{\text{VI}}\right] dt \\ &\le \left(\tau T \rho_p + \frac{\tau^2 T^2 \sigma^2}{n}\right) \int_0^{t^*} dt + \min\{\tau T, \tau^2 T^2\}(B_y + B)^2 \int_{t^*}^\infty e^{-2nt^2/(1+B_y^2)^2} dt \\ &\le \left(\tau T \rho_p + \frac{\tau^2 T^2 \sigma^2}{n}\right) t^* \\ &\quad + \min\{\tau T, \tau^2 T^2\}(B_y + B)^2 \frac{(1+B_y^2)^2 e^{(-(1+\log\max\{1,(B_y+B)^2/(\sigma^2/n)\}))}}{4nt^*} \\ &\le \frac{1+B_y^2}{\sqrt{n}}\left(\tau T \rho_p + \frac{\tau^2 T^2 \sigma^2}{n}\right)\left(\frac{\sqrt{1 + \log\max\left\{1, \left(\frac{(B_y+B)^2}{\sigma^2/n}\right)\right\}}}{\sqrt{2}}\right. \\ &\quad \left. + \frac{1}{2e\sqrt{2(1 + \log\max\left\{1, \left(\frac{(B_y+B)^2}{\sigma^2/n}\right)\right\})}}\right).\end{aligned}$$

Consequently,

$$\mathbb{E}\left[\Delta_S \|u_S\|_2^2 | \mathcal{F}_{\mathrm{VI}}\right] \le \frac{1+B_y^2}{\sqrt{n}} \left(\tau T \rho_p + \frac{\tau^2 T^2 \sigma^2}{n}\right) \left(\frac{1}{2} + \frac{\sqrt{1+\log\max\left\{1, \left(\frac{(B_y+B)^2}{\sigma^2/n}\right)\right\}}}{\sqrt{2}} + \frac{1}{2e\sqrt{2(1+\log\max\left\{1, \left(\frac{(B_y+B)^2}{\sigma^2/n}\right)\right\})}}\right)$$
$$\le \frac{2(1+B_y^2)\sqrt{1+\log\max\left\{1, \left(\frac{(B_y+B)^2}{\sigma^2/n}\right)\right\}}}{\sqrt{n}} \left(\tau T \rho_p + \frac{\tau^2 T^2 \sigma^2}{n}\right)$$

where we used that $1 + \log\max\left\{1, \left(\frac{(B_y+B)^2}{\sigma^2/n}\right)\right\} \ge 1$ in the last line. Finally we obtain

$$\begin{aligned}
\mathcal{R}_{\mathrm{within}}(T) &\le \mathbb{E}\left[\left[L_S(f_\beta) - \hat{L}_{D_S}(f_\beta)\right]_+\right] \\
&\le 2\mathbb{E}\left[\Delta_S\left[1 + (R_0+B)^2 + \|u_S\|_2^2\right]\right] \\
&= 2\left((1+(R_0+B)^2)\mathbb{E}\left[\Delta_S\right] + \mathbb{E}\left[\Delta_S \|u_S\|_2^2\right]\right) \\
&\le 2\Bigg(\frac{1+B_y^2}{2\sqrt{n}} \cdot (1+(R_0+B)^2) \\
&\quad + \frac{2(1+B_y^2)\sqrt{1+\log\max\left\{1, \left(\frac{(B_y+B)^2}{\sigma^2/n}\right)\right\}}}{\sqrt{n}} \left(\tau T \rho_p + \frac{\tau^2 T^2 \sigma^2}{n}\right)\Bigg) \\
&\le \frac{(1+B_y^2)\left[1+(R_0+B)^2\right]}{\sqrt{n}} \\
&\quad + 4(1+B_y^2)\sqrt{1+\log\max\left\{1, \frac{(B_y+B)^2}{\sigma^2/n}\right\}} \left(\frac{\tau T \rho_p}{\sqrt{n}} + \frac{\tau^2 T^2 \sigma^2}{n^{3/2}}\right).
\end{aligned}$$

yielding the desired inequality. The rate follows by substituting the stated bound on ρ_p. Its conditions are given in Corollary 10, Lemma 11, and (23). □

Lemma 17. *(Across-task generalization.) Under the assumptions of Lemma 16, let $(\mathcal{V}, d_{\mathcal{V}})$ be the admissible representation class, fixed conditional on $\mathcal{F}_{\mathrm{VI}}$. Let $D_{1:M}$ denote the support samples of M independent meta-training tasks, each of size n. Define the meta-level empirical loss function $j_{\vartheta,\beta,T}(D) := \frac{1}{2n}\|\Phi_D^\vartheta \beta_{D,T} - Y_D\|_2^2$. Assume the Lipschitz bound $|j_{\vartheta,\beta,T}(D) - j_{\vartheta',\beta,T}(D)| \le L_{\mathcal{V}} d_{\mathcal{V}}(\vartheta,\vartheta')$ uniformly over $\vartheta, \vartheta' \in \mathcal{V}$, $\|\beta\|_2 \le R_0$, and D, at the depths under consideration. Then, for every $T \ge 1$ and $\epsilon > 0$,*

$$\mathcal{R}_{\mathrm{across}}(T) \le \frac{2R_0 B_y}{\sqrt{M}} + \frac{C_K R_0^2 \sqrt{\log(2p)}}{2eT\tau\sqrt{M}} + (R_0+B_y)^2\sqrt{\frac{2\log\mathcal{N}(\epsilon,\mathcal{V},d_{\mathcal{V}})}{M}} + 2L_{\mathcal{V}}\epsilon.$$

Proof. In this proof we add the superscript ϑ as

$$A_D^\vartheta := I - \tau H_D^\vartheta, \quad d_{D,T}^\vartheta := \tau \sum_{t=0}^{T-1} (A_D^\vartheta)^t b_D^\vartheta$$

then we have $\beta_{D,T}^{\vartheta,\beta} = (A_D^\vartheta)^T \beta + d_{D,T}^\vartheta$. Recall the empirical loss

$$\hat{L}_D^\vartheta(\beta) = \frac{1}{2}\beta^\top H_D^\vartheta \beta - (b_D^\vartheta)^\top \beta + \frac{1}{2n}\|Y_D\|_2^2.$$

Also from that

$$\begin{aligned} H_D^\vartheta d_{D,T}^\vartheta &= \tau H_D^\vartheta \sum_{t=0}^{T-1} (A_D^\vartheta)^t b_D^\vartheta \\ &= (I - A_D^\vartheta) \sum_{t=0}^{T-1} (A_D^\vartheta)^t b_D^\vartheta \\ &= (I - (A_D^\vartheta)^T) b_D^\vartheta, \end{aligned}$$

we have

$$\begin{aligned} j_{\vartheta,\beta,T}(D) &= \frac{1}{2}\beta^\top (A_D^\vartheta)^T H_D^\vartheta (A_D^\vartheta)^T \beta + \beta^\top (A_D^\vartheta)^T (H_D^\vartheta d_{D,T}^\vartheta - b_D^\vartheta) + \underbrace{j_{\vartheta,0,T}(D)}_{:=c_{\vartheta,D,T}} \\ &= \frac{1}{2}\beta^\top (A_D^\vartheta)^T H_D^\vartheta (A_D^\vartheta)^T \beta + \beta^\top (A_D^\vartheta)^T (-(A_D^\vartheta)^T b_D^\vartheta) + c_{\vartheta,D,T} \\ &= \frac{1}{2}\beta^\top H_D^\vartheta (A_D^\vartheta)^{2T} \beta - \left((A_D^\vartheta)^{2T} b_D^\vartheta\right)^\top \beta + c_{\vartheta,D,T}. \end{aligned}$$

We provide upper bounds for the coefficients appearing above. As $0 \le \tau\lambda_{D,i} \le 1$ for every eigenvalue $\lambda_{D,i}$ of H_D^ϑ, we have $(1-\tau\lambda_{D,i})^{2T} \le e^{-2T\tau\lambda_{D,i}}$, and hence

$$\|\frac{1}{2} H_D^\vartheta (A_D^\vartheta)^{2T}\|_{\mathrm{op}} \le \frac{1}{2} \sup_{\lambda_{D,i} \ge 0} \lambda_{D,i} e^{-2\tau T \lambda_{D,i}} = \frac{1}{4eT\tau}.$$

Also,

$$\|b_D^\vartheta\|_2 = \|\frac{1}{n}\sum_{i=1}^n \phi_\vartheta(X_i) Y_i\|_2 \le \frac{1}{n}\sum_{i=1}^n \|\phi_\vartheta(X_i)\|_2 |Y_i| \le B_y$$

From that $\|A_D^\vartheta\|_{\mathrm{op}} \le 1$, we have

$$\|(A_D^\vartheta)^{2T} b_D^\vartheta\|_2 \le B_y.$$

From the definition of $\mathcal{R}_{\mathrm{across}}(T)$ we easily have

$$\begin{aligned} \mathcal{R}_{\mathrm{across}}(T) &= \mathbb{E}\left[\mathbb{E}_{S|\mathcal{F}_{\mathrm{VI}}}\left[j_{\vartheta,\beta,T}(D_S)\right] - \frac{1}{M}\sum_{m=1}^M j_{\vartheta,\beta,T}(D_m) \mid \mathcal{F}_{\mathrm{VI}}\right] \\ &\le \mathbb{E}\left[\sup_{\vartheta\in\mathcal{V}, \|\beta\|_2 \le R_0} \left\{\mathbb{E}_{S|\mathcal{F}_{\mathrm{VI}}}\left[j_{\vartheta,\beta,T}(D_S)\right] - \frac{1}{M}\sum_{m=1}^M j_{\vartheta,\beta,T}(D_m)\right\} \mid \mathcal{F}_{\mathrm{VI}}\right] \end{aligned}$$

Let $\{\vartheta_1, \cdots, \vartheta_N\}$ be the ϵ-net of $\mathcal{V}$: $N := \mathcal{N}(\epsilon, \mathcal{V}, d_{\mathcal{V}})$. By Lemma A.1, for each $\vartheta \in \mathcal{V}$ there exists ϑ_k such that $d_{\mathcal{V}}(\vartheta, \vartheta_k) \le \epsilon$. From the assumption that $|j_{\vartheta,\beta,T}(D) - j_{\vartheta',\beta,T}(D)| \le L_{\mathcal{V}} d_{\mathcal{V}}(\vartheta, \vartheta')$ we have

$$\begin{aligned} &\mathbb{E}_{S|\mathcal{F}_{\mathrm{VI}}}\left[j_{\vartheta,\beta,T}(D_S)\right] - \frac{1}{M}\sum_{m=1}^M j_{\vartheta,\beta,T}(D_m) \\ &\qquad \le \mathbb{E}_{S|\mathcal{F}_{\mathrm{VI}}}\left[j_{\vartheta_k,\beta,T}(D_S)\right] - \frac{1}{M}\sum_{m=1}^M j_{\vartheta_k,\beta,T}(D_m) + 2L_{\mathcal{V}}\epsilon \end{aligned}$$

so that

$$\begin{aligned} &\sup_{\vartheta\in\mathcal{V}, \|\beta\|_2 \le R_0} \left\{\mathbb{E}_{S|\mathcal{F}_{\mathrm{VI}}}\left[j_{\vartheta,\beta,T}(D_S)\right] - \frac{1}{M}\sum_{m=1}^M j_{\vartheta,\beta,T}(D_m)\right\} \\ &\qquad \le \max_{1\le k\le N} \sup_{\|\beta\|_2 \le R_0} \left\{\mathbb{E}_{S|\mathcal{F}_{\mathrm{VI}}}\left[j_{\vartheta_k,\beta,T}(D_S)\right] - \frac{1}{M}\sum_{m=1}^M j_{\vartheta_k,\beta,T}(D_m)\right\} + 2L_{\mathcal{V}}\epsilon. \end{aligned}$$

Let $\omega = (\omega_m)_{m=1}^M$ be independent Rademacher signs, independent of the task data. Conditional symmetrization (Lemma A.6) gives

$$\mathbb{E}\left[\max_{1\le k\le N}\sup_{\|\beta\|_2\le R_0}\left\{\mathbb{E}_{S|\mathcal{F}_{\mathrm{VI}}}\left[j_{\vartheta_k,\beta,T}(D_S)\right]-\frac{1}{M}\sum_{m=1}^M j_{\vartheta_k,\beta,T}(D_m)\right\}\mid\mathcal{F}_{\mathrm{VI}}\right]$$
$$\le 2\mathbb{E}\left[\mathbb{E}_\omega\max_{1\le k\le N}\underbrace{\sup_{\|\beta\|\le R_0}\frac{1}{M}\sum_{m=1}^M\omega_m j_{\vartheta_k,\beta,T}(D_m)}_{:=Z_k(\omega)}\mid\mathcal{F}_{\mathrm{VI}}\right].$$

Lemma A.9 implies, uniformly for $\|\beta\|_2\le R_0$,

$$0\le j_{\vartheta,\beta,T}(D)\le\widehat{L}_D^\vartheta(\beta)\le\frac{1}{2}(R_0+B_y)^2.$$

We next apply Lemma A.7 to the N classes indexed by the net points with $b=(R_0+B_y)^2/2$ to get

$$\mathbb{E}_\omega\max_{1\le k\le N}\sup_{\|\beta\|\le R_0}\frac{1}{M}\sum_{m=1}^M\omega_m j_{\vartheta_k,\beta,T}(D_m)\le\max_{1\le k\le N}\mathbb{E}_\omega\sup_{\|\beta\|_2\le R_0}\frac{1}{M}\sum_{m=1}^M\omega_m j_{\vartheta_k,\beta,T}(D_m)$$
$$+\frac{1}{2}(R_0+B_y)^2\sqrt{\frac{2\log N}{M}}.$$

Now for a fixed ϑ_k,

$$\mathbb{E}_\omega\sup_{\|\beta\|_2\le R_0}\frac{1}{M}\sum_{m=1}^M\omega_m j_{\vartheta_k,\beta,T}(D_m)$$
$$=\mathbb{E}_\omega\sup_{\|\beta\|_2\le R_0}\frac{1}{M}\sum_{m=1}^M\omega_m\left(\frac{1}{2}\beta^\top H_{D_m}^{\vartheta_k}(A_{D_m}^{\vartheta_k})^{2T}\beta-((A_{D_m}^{\vartheta_k})^{2T}b_{D_m}^{\vartheta_k})^\top\beta+c_{\vartheta_k,D_m,T}\right)$$
$$\le\frac{R_0^2}{M}\mathbb{E}_\omega\|\sum_{m=1}^M\omega_m\frac{1}{2}H_{D_m}^{\vartheta_k}(A_{D_m}^{\vartheta_k})^{2T}\|_{\mathrm{op}}+\frac{R_0}{M}\mathbb{E}_\omega\|\sum_{m=1}^M\omega_m(A_{D_m}^{\vartheta_k})^{2T}b_{D_m}^{\vartheta_k}\|_2.$$

For a fixed net point and observed task samples, $c_{\vartheta_k,D_m,T}$ does not depend on β or the signs. It therefore leaves the supremum as an additive term and has zero Rademacher expectation. By the matrix Khintchine bound (Lemma A.8),

$$\mathbb{E}_\omega\|\sum_{m=1}^M\omega_m\frac{1}{2}H_{D_m}^{\vartheta_k}(A_{D_m}^{\vartheta_k})^{2T}\|_{\mathrm{op}}\le C_K\sqrt{\log(2p)}\|\left[\sum_{m=1}^M\left(\frac{1}{2}H_{D_m}^{\vartheta_k}(A_{D_m}^{\vartheta_k})^{2T}\right)^2\right]^{1/2}\|_{\mathrm{op}}$$
$$\le C_K\sqrt{\log(2p)}\frac{\sqrt{M}}{4eT\tau}.$$

For the linear term, the Rademacher part of Lemma A.5 gives

$$\mathbb{E}_\omega\|\sum_{m=1}^M\omega_m(A_{D_m}^{\vartheta_k})^{2T}b_{D_m}^{\vartheta_k}\|_2\le\left(\mathbb{E}_\omega\|\sum_{m=1}^M\omega_m(A_{D_m}^{\vartheta_k})^{2T}b_{D_m}^{\vartheta_k}\|_2^2\right)^{1/2}$$
$$\le\left(\sum_{m=1}^M\|(A_{D_m}^{\vartheta_k})^{2T}b_{D_m}^{\vartheta_k}\|_2^2\right)^{1/2}$$
$$\le\sqrt{M}B_y.$$

Thus for each ϑ_k, we obtain

$$\mathbb{E}_\omega\sup_{\|\beta\|_2\le R_0}\frac{1}{M}\sum_{m=1}^M\omega_m j_{\vartheta_k,\beta,T}(D_m)\le\frac{R_0B_y}{\sqrt{M}}+\frac{C_KR_0^2\sqrt{\log(2p)}}{4eT\tau\sqrt{M}}.$$

Plugging all the bounds, we finally obtain the desired bound as follows.

$$\mathcal{R}_{\mathrm{across}}(T) \le \frac{2R_0 B_y}{\sqrt{M}} + \frac{C_K R_0^2 \sqrt{\log(2p)}}{2eT\tau\sqrt{M}} + (R_0 + B_y)^2 \sqrt{\frac{2\log \mathcal{N}(\epsilon, \mathcal{V}, d_{\mathcal{V}})}{M}} + 2L_{\mathcal{V}}\epsilon.$$

□

Note that, to match with the rate later, we choose $\epsilon = \frac{1}{(1+L_{\mathcal{V}})\sqrt{M}}$.

Lemma 18. *(Bound of optimization error in $\mathcal{H}_{k,Z}$: $\mathcal{O}(T)$). Assume the common conditions stated in Appendix E. Let $E > 0$ denote the number of inner loop training steps used for meta-training. Define $\hat{\beta}_m \in \arg\min_\beta \hat{L}_{D_m}(f_{\hat{\vartheta},\beta})$ and $\Pi_{m,+}$ as the orthogonal projector onto $\mathrm{range}(\mathrm{H_m})$. Then, writing $\hat{\mathcal{P}}_{E,M}(\hat{\vartheta}, \hat{\beta}_0) := \frac{1}{M}\sum_{m=1}^{M} \|\hat{\beta}_0 - \mathcal{A}_E^{\hat{\vartheta}}(\hat{\beta}_0; D_m)\|_2^2$, we have*

$$\mathcal{O}(T) \le \frac{\mathbb{E}\hat{\mathcal{P}}_{E,M}(\hat{\vartheta}, \hat{\beta}_0)}{4eT\tau(1-(1-\tau\mu)^E)^2}.$$

Conditional on $\mathcal{F}_{\mathrm{VI}}$, we assume that there is a fixed $\mu > 0$ such that almost surely over meta-training, every positive eigenvalue of H_m, $m = 1, \dots, M$ is at least μ.

Proof. From the simple calculation, we have that

$$\hat{L}_{D_m}(\hat{\vartheta}, \beta_{m,T}) - \hat{L}_{D_m}(\hat{\vartheta}, \hat{\beta}_m) = \frac{1}{2}(\hat{\beta}_0 - \hat{\beta}_m)^\top H_m (I - \tau H_m)^{2T} (\hat{\beta}_0 - \hat{\beta}_m).$$

Diagonalizing H_m as $H_m = U_m \mathrm{diag}(\lambda_{\mathrm{m},1}, \cdots, \lambda_{\mathrm{m,p}}) U_m^\top$ yields

$$\frac{1}{2}(\hat{\beta}_0 - \hat{\beta}_m)^\top H_m (I - \tau H_m)^{2T} (\hat{\beta}_0 - \hat{\beta}_m) = \frac{1}{2} \sum_{\lambda_{m,i}>0} \lambda_{m,i}(1-\tau\lambda_{m,i})^{2T} \left[u_{m,i}^\top (\hat{\beta}_0 - \hat{\beta}_m)\right]^2$$

From $\lambda(1-\tau\lambda)^{2T} \le \lambda e^{-2T\tau\lambda} \le \frac{1}{2eT\tau}$,

$$\hat{L}_{D_m}(\hat{\vartheta}, \beta_{m,T}) - \hat{L}_{D_m}(\hat{\vartheta}, \hat{\beta}_m) \le \frac{\|\Pi_{m,+}(\hat{\beta}_0 - \hat{\beta}_m)\|_2^2}{4eT\tau}$$

If $\widehat{\beta}_m$ is an empirical minimizer, then $H_m \widehat{\beta}_m = b_m$ and $\widehat{\beta}_0 - \mathcal{A}_E^{\hat{\vartheta}}(\widehat{\beta}_0; D_m) = [I - (I - \tau H_m)^E](\widehat{\beta}_0 - \widehat{\beta}_m)$. The last assertion of Lemma A.9 therefore gives the second inequality given below. Taking expectations, we obtain

$$\begin{aligned}
\mathcal{O}(T) &\le \frac{1}{4eT\tau}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} \|\Pi_{m,+}(\hat{\beta}_0 - \hat{\beta}_m)\|_2^2\right] \\
&\le \frac{1}{4eT\tau}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} \frac{\|\hat{\beta}_0 - \mathcal{A}_E^{\hat{\vartheta}}(\hat{\beta}_0; D_m)\|_2^2}{(1-(1-\tau\mu)^E)^2}\right] \\
&\le \frac{\mathbb{E}\hat{\mathcal{P}}_{E,M}(\hat{\vartheta}, \hat{\beta}_0)}{4eT\tau(1-(1-\tau\mu)^E)^2}.
\end{aligned}$$

□

Corollary 19. *For every $T \ge 1$, the uniform excess transfer risk satisfies*

$$\begin{aligned}
\mathcal{R}^u(T) \le\ & T^2 \left(\frac{4(1+B_y^2)\tau^2\sigma^2}{n^{3/2}} \sqrt{1 + \log\max\left\{1, \frac{(B_y+B)^2}{\sigma^2/n}\right\}}\right) \\
& + T \cdot \left(\frac{4(1+B_y^2)\tau\rho_p}{\sqrt{n}} \sqrt{1 + \log\max\left\{1, \frac{(B_y+B)^2}{\sigma^2/n}\right\}}\right) \\
& + \frac{1}{T} \cdot \left(\frac{C_K R_0^2 \sqrt{\log(2p)}}{2e\tau\sqrt{M}} + \frac{\mathbb{E}\hat{\mathcal{P}}_{E,M}(\hat{\vartheta}, \hat{\beta}_0)}{4e\tau(1-(1-\tau\mu)^E)^2}\right)
\end{aligned}$$

$$
\begin{aligned}
&+ (R_0 + 2B_y)^2 \sqrt{\frac{2\log \mathcal{N}(\frac{1}{(1+L_\mathcal{V})\sqrt{M}}, \mathcal{V}, d_\mathcal{V})}{M}} \\
&+ \frac{2 + 2R_0 B_y}{\sqrt{M}} + \frac{(1 + B_y^2)[1 + (R_0 + B)^2]}{\sqrt{n}} \\
&+ \mathcal{A}^*_{\text{sub}}.
\end{aligned}
$$

The bound contains terms proportional to T^{-1}, T, and T^2, together with statistical terms independent of T and the empirical comparison $\mathcal{A}^*_{\text{sub}}$. The latter is retained as defined in (20).

To simplify the notation, we rewrite the above bound as

$$\mathcal{R}^u(T) \le g(T) + C_0$$

where

$$
\begin{aligned}
g(T) &:= \frac{a_0}{T} + b_1 T + b_2 T^2, \\
A' &:= \frac{C_K R_0^2 \sqrt{\log(2p)}}{2e\sqrt{M}} + \frac{\mathbb{E}\hat{\mathcal{P}}_{E,M}}{4e(1-(1-\tau\mu)^E)^2}, \quad a_0 := \frac{A'}{\tau}, \\
L_n &:= 1 + \log \max\left\{1, \frac{(B_y + B)^2}{\sigma^2/n}\right\}, \\
b_1 &:= \frac{4(1 + B_y^2)\sqrt{L_n}\tau\rho_p}{\sqrt{n}}, \quad b_2 := \frac{4(1 + B_y^2)\sqrt{L_n}\tau^2\sigma^2}{n^{3/2}}, \\
C_0 &:= \mathcal{A}^*_{\text{sub}} + (R_0 + 2B_y)^2 \sqrt{\frac{2\log \mathcal{N}\left(\frac{1}{(1+L_\mathcal{V})\sqrt{M}}, \mathcal{V}, d_\mathcal{V}\right)}{M}} \\
&\quad + \frac{2 + 2R_0 B_y}{\sqrt{M}} + \frac{(1 + B_y^2)[1 + (R_0 + B)^2]}{\sqrt{n}}.
\end{aligned} \tag{21}
$$

Let T_{lin} and T_{quad} be the balancing steps between the decreasing term a_0/T and increasing terms $b_1 T$ and $b_2 T^2$, respectively:

$$T_{\text{lin}} = \sqrt{\frac{a_0}{b_1}}, \quad T_{\text{quad}} = \left(\frac{a_0}{b_2}\right)^{1/3}.$$

Let $T_0 = \min\{T_{\text{lin}}, T_{\text{quad}}\}$. Then

$$
\begin{aligned}
b_1 T_0 \le \sqrt{a_0 b_1}, \quad b_2 T_0^2 \le a_0^{2/3} b_2^{1/3}, \\
\frac{a_0}{T_0} = \max\{\sqrt{a_0 b_1}, a_0^{2/3} b_2^{1/3}\} \le \sqrt{a_0 b_1} + a_0^{2/3} b_2^{1/3}.
\end{aligned}
$$

Thus,

$$g(T_0) \le 2\sqrt{a_0 b_1} + 2a_0^{2/3} b_2^{1/3}.$$

Now let $T^* = \max\{1, \lceil T_0 \rceil\}$. If $T_0 \ge 1$ Then $T_0 \le T^* \le T_0 + 1 \le 2T_0$, so that

$$
\begin{aligned}
\frac{a_0}{T^*} \le \sqrt{a_0 b_1} + a_0^{2/3} b_2^{1/3}, \quad b_1 T^* \le 2\sqrt{a_0 b_1}, \\
b_2 (T^*)^2 \le 4 a_0^{2/3} b_2^{1/3}.
\end{aligned}
$$

If $0 \le T_0 < 1$ then $g(T^*) = g(1) = a_0 + b_1 + b_2$ so that, considering all the cases yields

$$g(T^*) \le 3\sqrt{a_0 b_1} + 5a_0^{2/3} b_2^{1/3} + b_1 + b_2.$$

Direct substitution gives

$$a_0 b_1 = \frac{A'}{\tau} \frac{4(1 + B_y^2)\sqrt{L_n}\,\tau\rho_p}{\sqrt{n}} = \frac{4(1 + B_y^2)\sqrt{L_n}\,A'\rho_p}{\sqrt{n}},$$

$$\sqrt{a_0 b_1} = 2\sqrt{1+B_y^2}\, L_n^{1/4} \frac{\sqrt{A'\rho_p}}{n^{1/4}},$$
$$a_0^2 b_2 = \frac{A'^2}{\tau^2} \frac{4(1+B_y^2)\sqrt{L_n}\,\tau^2\sigma^2}{n^{3/2}} = \frac{4(1+B_y^2)\sqrt{L_n}\,A'^2\sigma^2}{n^{3/2}},$$
$$a_0^{2/3} b_2^{1/3} = [4(1+B_y^2)]^{1/3} L_n^{1/6} \frac{A'^{2/3}\sigma^{2/3}}{\sqrt{n}}.$$

Therefore, the bound at T^* takes the form of

$$\mathcal{R}^u(T^*) \leq 6\sqrt{1+B_y^2}\, L_n^{1/4} \frac{\sqrt{A'\rho_p}}{n^{1/4}} + 5[4(1+B_y^2)]^{1/3} L_n^{1/6} \frac{A'^{2/3}\sigma^{2/3}}{\sqrt{n}} + \frac{4(1+B_y^2)\sqrt{L_n}\tau\rho_p}{\sqrt{n}} + \frac{4(1+B_y^2)\sqrt{L_n}\tau^2\sigma^2}{n^{3/2}} + C_0.$$

Now using AM-GM inequality, we have

$$\frac{\sqrt{A'\rho_p}}{n^{1/4}} \leq \frac{1}{2}\left(\rho_p + \frac{A'}{\sqrt{n}}\right).$$

For convenience, we make an assumption that $\frac{\mathbb{E}\left[\hat{\mathcal{P}}_{E,M}\right]}{(1-(1-\tau\mu)^E)^2} = O(1)$ so that $A' = O(1 + \frac{\sqrt{\log 2p}}{\sqrt{M}})$. Thus, substituting this inequality yields

$$\mathcal{R}^u(T^*) \leq C_0 + O\Bigg(L_n^{1/4}\rho_p + \frac{L_n^{1/4}}{\sqrt{n}}\left[1 + \frac{\sqrt{\log(2p)}}{\sqrt{M}}\right] + \frac{L_n^{1/6}}{\sqrt{n}}\left[1 + \frac{[\log(2p)]^{1/3}}{M^{1/3}}\right] + \sqrt{L_n}\left(\frac{\rho_p}{\sqrt{n}} + \frac{1}{n^{3/2}}\right)\Bigg). \tag{22}$$

For the rate of the residual, we assume the hypotheses of Corollary 10 and Lemma 11. Here $r_{S,\hat{Z}_{\mathrm{tr}}}$ denote the residual associated with the reference kernel and empirical trace minimizer of Appendix B. We additionally assume the following residual comparison between the final representation and the fixed-kernel empirical trace minimizer:

$$\rho_p = \mathbb{E}\left[\|r_S\|^2_{L^2(P_{s|S})} \mid \mathcal{F}_{\mathrm{VI}}\right] \leq \mathbb{E}_S\left[\|r_{S,\hat{Z}_{\mathrm{tr}}}\|^2_{L^2(P_{s|S})} \mid \mathcal{F}_{\mathrm{VI}}\right]. \tag{23}$$

Consequently, on the VI event of probability at least $1-\delta$, the bound of the residual takes the form

$$\rho_p = \tilde{O}\left(p^{-\alpha} + \sqrt{\frac{pd + \log(1/\delta)}{M_0}}\left(1 + \frac{1}{\sqrt{n}}\right) + \frac{1}{n}\right).$$

We also know that

$$C_0 \leq \mathcal{A}^*_{\mathrm{sub}} + \tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{M}} + \sqrt{\frac{pd + D_\psi + D_{\mathrm{hyp}}}{M}}\right)$$

where D_ψ and D_{hyp} denote the dimension of the feature-extractor parameter space and kernel hyperparameter space, respectively. Indeed, with the bounded representation class $\mathcal{V} \subset \mathbb{R}^{pd+D_\psi+D_{\mathrm{hyp}}}$ with the Euclidean metric $d_{\mathcal{V}}$, Lemma A.1 gives

$$\log \mathcal{N}(\epsilon, \mathcal{V}, d_{\mathcal{V}}) \leq (pd + D_\psi + D_{\mathrm{hyp}}) \log\left(1 + \frac{2\sup_{\vartheta\in\mathcal{V}} \|\vartheta\|_2}{\epsilon}\right).$$

Taking $\epsilon = \frac{1}{(1+L_{\mathcal{V}})\sqrt{M}}$ yields

$$\log \mathcal{N}\left(\frac{1}{(1+L_{\mathcal{V}})\sqrt{M}}, \mathcal{V}, d_{\mathcal{V}}\right) \leq (pd + D_\psi + D_{\mathrm{hyp}}) \log\left[1 + 2(1+L_{\mathcal{V}})\sqrt{M} \sup_{\vartheta\in\mathcal{V}} \|\vartheta\|_2\right].$$

Substituting the residual and covering bounds yields

$$\mathcal{R}^u(T^*) \le \mathcal{A}^*_{\text{sub}} + \widetilde{O}\left(p^{-\alpha} + \sqrt{\frac{pd}{M_0}} + \sqrt{\frac{pd + D_\psi + D_{\text{hyp}}}{M}} + \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{M}}\right).$$

The approximation term and the two inducing points estimation terms are balanced, up to logarithmic factors, by

$$p^* \asymp \left(\frac{\min\{M_0, M\}}{d}\right)^{1/(2\alpha+1)}.$$

When this choice is admissible, substitution gives

$$\mathcal{R}^u(T^*) \le \mathcal{A}^*_{\text{sub}} + \widetilde{O}\left[\left(\frac{d}{\min\{M_0, M\}}\right)^{\alpha/(2\alpha+1)} + \sqrt{\frac{D_\psi + D_{\text{hyp}}}{M}} + \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{M}}\right].$$

F Experimental details and additional results

This section describes the details on the experimental setups including task generation, the training and evaluation procedures for each method, and the ablations summarized in Section 5.

F.1 Dataset description and preparation

Synthetic datasets. In all the synthetic datasets used for the experiments, the inputs are drawn uniformly on the interval $[-5, 5]$ and labels with Gaussian noise with standard deviation $\sigma = 0.1$. Sinusoidal tasks are $A\sin(x + \phi)$ with $A \sim U[0.1, 5]$ and $\phi \sim U[0, 2\pi]$. Omega tasks replace the frequency by $\omega \sim U[0.7, 1.3]$. Misfit tasks add $0.5\sin(7x + \phi_2)$ where $\phi_2 \sim U[0, 2\pi]$ to the sinusoidal task. This dataset tests the effect of an additional high-frequency component. Oracle tasks are $\sum_{j=1}^{32} \alpha_j k(x, z_j^0)$ where k is the RBF kernel with length scale 1, the points z_j^0 are selected from a uniform grid of $[-5, 5]$, and the coefficients are independent standard Gaussian variables rescaled to standard deviation from $U[0.5, 2.5]$. Structured-Oracle tasks are similar to the Oracle task where the weights α are defined as $\alpha = \bar{\alpha} + (s/\sqrt{k})U\gamma$, $\gamma \sim \mathcal{N}(0, I_k)$. The vector $\bar{\alpha}$ and the matrix $U \in \mathbb{R}^{32\times k}$ are drawn once with a fixed seed and shared by all tasks. The columns of U are chosen so that the functions $\sum_j U_{j,i} k(\cdot, z_j^0)$ are orthonormal. The mean function $\sum_j \bar{\alpha}_j k(\cdot, z_j^0)$ has a standard deviation of 1.5. In our experiment we use $k = 8$ and $s = 1$. This task is devised to construct a synthetic task with a nonzero mean function, as the first four synthetic tasks have zero mean function. All the meta-training, validation and test tasks are independent draws from the same generator.

Real world datasets. To validate our method in the real world datasets, we use three well-known datasets: Beijing air quality (Zhang et al., 2017), METR-LA and PEMS-BAY (Li et al., 2018) datasets. Beijing air quality contains hourly records from 12 monitoring stations. Each task is a segment of 50 consecutive records of one station, taken after removing records with missing values, and 200 segments are drawn per station at uniformly random positions. The inputs are the position within the window and ten measurements (including PM10, SO2, NO2, CO, O3, temperature, pressure, dew point, rainfall and wind speed), and the target is the PM2.5 concentration. METR-LA contains speed readings of 207 loop detectors every five minutes. Each task is a segment of 50 time steps of one detector. For each detector 100 segments are drawn randomly. The inputs are the five-minute time slot within the day and the six previous readings, and the target is the current speed. PEMS-BAY contains speed readings of 325 detectors every five minutes and is converted with the same input features and target. For each detector, 23 segments of 56 consecutive steps, six for lags and 50 for targets, are placed evenly along the recording. On METR-LA and PEMS-BAY, segments with missing readings are dropped, which gives the numbers of tasks in Table 3. In all the real-world datasets the input features are standardized and targets are rescaled to the range $[0, 10]$ by a min-max transform. Table 3 contains the detailed numerical information of the number of input features and tasks.

Support and query sizes. Support datasets have $n = 10$ points on the synthetic datasets and $n = 5$ on the real world datasets. During meta-training every method receives 40 query points per task. Our proposed method then splits them into two halves of 20 points, one for the shared initial head and one for calculating the query losses. Validation and test tasks use 20 query points. On the real world datasets, each validation or test task is drawn uniformly at random from its group and its 50

Table 3: Number of input features and number of tasks in each group of the real world datasets. Every task has 50 observations.

Datasets	Input features	Stage I	Meta-training	Validation	Test	Total
Air quality	11	720	960	360	360	2,400
METR-LA	7	5,233	6,978	2,616	2,619	17,446
PEMS-BAY	7	2,238	2,984	1,119	1,120	7,461

observations are split at random into the support and query sets. The test error of the proposed model is averaged over 200 tasks per run.

F.2 Proposed method

Model architecture. The feature map h_ψ is a multilayer feedforward network with layers of widths input-40-40-32 and ReLU activations. The weights of the last layer are initialized with mean zero and standard deviation 0.01. The kernel on features is an RBF kernel with a learned length scale and output scale. By default, we set $p = 32$ as the number of inducing points.

Stage I (VI pre-training). In Stage I, the feature map, the inducing points and the kernel parameters are pre-trained with the sparse variational objective (Equation (7)) on $M_0 = 1000$ tasks with $n_0 = 10$ points each. We run 5,000 iterations with Adam optimizer with initial learning rate 0.001 with a batch of 64 tasks at each step. The result is stored for each dataset and shared by every run and every ablation of that dataset.

Stage II (Meta-training) In the outer-loop, we run 3,000 iterations with Adam optimizer with initial learning rate 0.001 and 64 tasks per step. Gradient norm is clipped at 1. For the inner-loop functional gradient descent, we set the inner step as $E = 10$, step size $\tau = 0.5$ and query-gradient weight of $\eta = 10^{-5}$. At every step the shared head is computed in closed form from the support dataset and the first half of the query dataset of each task. We use damping 10^{-4} to the linear system as stated in Equation (10).

Stage III (Testing). For the (meta) testing, we run the depth at most $T = 30$, as default. To compute the early stopping rules, we compute the necessary quantities using the separate 300 validation tasks for each training run.

F.3 Baselines

MAML. In the implementation of MAML, we use a multilayer feedforward network with layers of widths input-40-40-1 and ReLU activations. During training, we use inner learning rate 0.01 and 10 inner-loop updates, to match with the training depth of our method. We also separately evaluate the case with one inner-loop update, which matches with the original MAML setup. For the outer loop, we use learning rate of 0.001, 64 tasks per step with 3,000 iterations. The inner loop uses standard gradient descent and the outer loop uses Adam optimizer.

DKT. DKT uses the feature map with the same architecture as our proposed model; multilayer feedforward network with layers input-40-40-32 with ReLU activations, followed by an RBF kernel with learned length scale and output scale. Adam optimizer is used with initial learning rate of 0.001, 64 tasks per step and 3,000 iterations. At test time, it predicts the Gaussian process posterior mean with the given support dataset.

R2-D2 and iKML. Both baselines use the official implementation of Falk et al. (2022) with the hyperparameters of its Beijing air quality configuration. Both are evaluated on 1,000 validation tasks given every 250 iterations. We state here the iKML predictors returned zero in two of three runs on Sinusoidal and Misfit and in one of three runs on Omega and the Structured-Oracle task. We suspect the failure of iKML on the synthetic data (in few seeds) is related to the initialization of the frequency network, as one-dimensional inputs generate frequencies concentrated near a single value at initialization. We report the configuration used in official implementation in Table 5 and its corresponding query MSE values per runs in Table 4.

Evaluation of the baselines. On the synthetic datasets, each baseline is evaluated on the 300 test tasks per run and on real world datasets, 200 test tasks per run.

Table 4: Test query MSE of iKML using the official implementation for each training run, on the test tasks of Table 1.

Dataset	Run 1	Run 2	Run 3
Sinusoidal	4.3096	4.3065	0.0126
Omega	0.0395	4.1615	0.0386
Misfit	4.3462	4.3407	0.1728
Oracle	0.2985	0.2982	0.2981
Structured-Oracle	0.2895	3.4200	0.2842
Air quality	0.0741	0.0716	0.1230
METR-LA	0.5687	0.4427	0.6434
PEMS-BAY	0.1055	0.0719	0.0729

Table 5: Experimental setting and hyperparameters used for Table 1.

Tasks	
Synthetic inputs	$x \sim U[-5, 5]$, label noise $\sigma = 0.1$
Sinusoidal	$A \sin(x + \phi)$, $A \sim U[0.1, 5]$, $\phi \sim U[0, 2\pi]$
Omega	$A \sin(\omega x + \phi)$, $\omega \sim U[0.7, 1.3]$
Misfit	Sinusoidal $+ 0.5 \sin(7x + \phi_2)$, $\phi_2 \sim U[0, 2\pi]$
Oracle	$\sum_{j=1}^{32} \alpha_j k(x, z_j^0)$, RBF with $\ell = 1$ on a uniform grid, rescaled to standard deviation $U[0.5, 2.5]$
Structured-Oracle	$\sum_{j=1}^{32} \alpha_j k(x, z_j^0)$ with $\alpha = \bar{\alpha} + (s/\sqrt{k})U\gamma$, $\gamma \sim \mathcal{N}(0, I_k)$; $\bar{\alpha}$ and $U \in \mathbb{R}^{32\times k}$ fixed across tasks, $k = 8$; the mean function has standard deviation 1.5 and the task-specific part has expected variance s^2 over the input range, $s = 1$
Real world	Beijing air quality, METR-LA, PEMS-BAY; disjoint task split 30/40/15/15 (VI / meta-train / validation / test)
Support / query	$n = 10$ (synthetic), $n = 5$ (real); 40 query points per meta-training task; 20 query points per test task
Meta-testing	three training runs; 300 test tasks per run on the synthetic datasets, 200 test tasks per run on the real world datasets
Ours	
Stage I (VI)	$M_0 = 1000$ tasks, $n_0 = 10$ points, 5,000 iterations, Adam 10^{-3}, 64 tasks per step
Representation	$p = 32$ inducing points, feature dimension 32, MLP width 40, RBF kernel
Stage II	3,000 iterations, Adam 10^{-3}, 64 tasks per step, $E = 10$, $\tau = 0.5$, $\eta = 10^{-5}$, ridge 10^{-4} on the shared-head system, gradient-norm clip 1.0; query split 20 / 20 between the head criterion and the outer loss
Test-time adaptation	$\tau_S = \min\{\tau, 1/\lambda_{\max}(H_S)\}$; T^* selected on 300 validation tasks over $0 \le T \le 30$; fixed τ during meta-training
Baselines (d=input feature dimension)	
MAML	MLP d–40–40–1 (ReLU), inner rate 0.01, second order, $E = 10$; Adam 10^{-3}, 64 tasks per step, 3,000 iterations
DKT	MLP d–40–40–32 (ReLU), RBF kernel with learned length scale, output scale and noise; marginal likelihood on support and query; Adam 10^{-3}, 64 tasks per step, 3,000 iterations
R2-D2	MLP d–64–(64–64)×3–64 (ReLU), linear kernel, learned $\log \lambda$ (initial $\lambda = 10^{-3}$); Adam 10^{-5}, 4 tasks per step, 10,000 iterations; validation MSE on 1,000 tasks every 250 iterations
iKML	Pushforward MLP 64–64–(64–64)×3–d (ReLU) on $z \sim \mathcal{N}(0, I_{64})$; $D =$ 20,000 random features $\sqrt{2/D}\cos(\omega^\top x + b)$, $b \sim U[0, 2\pi]$, resampled per task; learned $\log \lambda$ (initial $\lambda = 10^{-3}$); Adam 10^{-4}, 4 tasks per step, 10,000 iterations; validation MSE on 1,000 tasks every 250 iterations

F.4 ADAPTATION DEPTH

For the bound-selected values (bound(T^*)) in Table 1, we use 300 independent validation tasks, separate from the ones used for validation-MSE selection. Each task provides n support datasets and query datasets for estimating the quantities in (15). The query set has 200 observations for synthetic tasks and $50 - n$ for real world tasks: 45 at the default $n = 5$, 40 at $n = 10$, and 30 at $n = 20$.

For each validation task, let Φ_s and Φ_q be its whitened support and query feature matrices, with feature rows $k(x, Z)K_{ZZ}^{-1/2}$. We use $\beta_0 = K_{ZZ}^{1/2}\alpha$, $H_S = \Phi_s^\top \Phi_s/n = U_S \operatorname{diag}(\lambda_{S,j})U_S^\top$, and $\tau_S = \min\{\tau, 1/\lambda_{\max}(H_S)\}$. The projected coefficient is estimated by ridge regression using only the validation query labels:

$$\widehat{\widetilde{\beta}}_S = (\Phi_q^\top \Phi_q + \lambda_{\mathrm{rid}} I)^{-1} \Phi_q^\top y_q, \qquad \lambda_{\mathrm{rid}} = 10^{-3} \lambda_{\max}(\Phi_q^\top \Phi_q). \tag{24}$$

Support labels are used for adaptation and the support-residual estimate below, not for fitting $\widehat{\widetilde{\beta}}_S$.

On synthetic tasks we use the known noise variance $\widehat{\sigma}^2 = 0.1^2$. On real world tasks we use the likelihood variance learned in Stage I as a plug-in value. Each dataset shares one Stage I model across the three runs, and we fix a single variance per dataset across all tasks and runs. In our experiment the logged likelihood standard deviations are $\widehat{\sigma} = 0.0635$, 0.2275, and 0.0955 for Air quality, METR-LA, and PEMS-BAY, respectively, giving variances approximately 0.0040, 0.0518, and 0.0091.

Writing n_q for the number of validation query observations, we estimate the query and support residual quantities by subtracting the noise contribution and clipping at zero:

$$\begin{aligned} \widehat{r}_{q,S}^2 &= \max\left\{ \frac{\|y_q - \Phi_q \widehat{\widetilde{\beta}}_S\|_2^2}{n_q} - \widehat{\sigma}^2,\, 0 \right\}, \\ \widehat{r}_{s,S}^2 &= \max\left\{ \|y_s - \Phi_s \widehat{\widetilde{\beta}}_S\|_2^2 - n\widehat{\sigma}^2,\, 0 \right\}. \end{aligned} \tag{25}$$

The query quantity estimates a mean squared residual, whereas the support quantity estimates the squared norm of the residual vector. Let $\widehat{e}_{S,0} = U_S^\top(\beta_0 - \widehat{\widetilde{\beta}}_S)$ and $q_{S,j}(T) = (1 - \tau_S \lambda_{S,j})^T$. We evaluate

$$\begin{aligned} \widehat{B}_S(T) = \frac{1}{2}\widehat{r}_{q,S}^2 + \frac{1}{2}\sum_{\lambda_{S,j}=0} \widehat{e}_{S,0,j}^2 + \sum_{\lambda_{S,j}>0} q_{S,j}(T)^2 \widehat{e}_{S,0,j}^2 \\ + \frac{\widehat{r}_{s,S}^2 + \widehat{\sigma}^2}{n} \sum_{\lambda_{S,j}>0} \frac{(1 - q_{S,j}(T))^2}{\lambda_{S,j}}. \end{aligned} \tag{26}$$

The estimated bound curve is computed for depths $0 \le T \le 200$. For Table 1, a single depth minimizes the average of $\widehat{B}_S(T)$ over the 300 validation tasks, restricted to $0 \le T \le 30$. In real world datasets, the bound-selected entries use 200 test-task evaluations with 20 query observations per task at the selected depth. Note that $\widehat{B}_S(T)$ is a plug-in surrogate for the conditional bound; this stopping rule is distinct from the task-specific diagnostic in Appendix F.7 and the uniform stopping rule in (18).

Test-time step and support size. Table 6 reports query MSE for the proposed model at fixed adaptation depths with several support sizes, on 300 test tasks per run on the synthetic datasets and 200 test tasks per run on the real world datasets. Table 7 reports the same evaluation restricted to $T \le 30$, the range used for the comparison with MAML in Table 9 and for the selection in Table 1. For every dataset the numerically computed limit as $T \to \infty$ has larger error than finite depth whenever $n \le 10$, supporting the use of the early stopping in such settings. The early stopping depth T^* is computed using the separate validation tasks, by minimizing the average estimated bound appearing at Equation (15). We also report the case where the maximum adaptation step is increased to 200.

Table 6: Test query MSE of the proposed model as a function of the depth T and of the support size n, mean over three runs, with 300 test tasks per run on the synthetic datasets and 200 test tasks per run on the real world datasets. Both selections use $0 \leq T \leq 200$. The column val(T^*) reports the error at the depth T^* that minimizes the mean query error on 300 validation tasks, and the column bound(T^*) denotes the error at the depth that minimizes the bound of Theorem 2 averaged over the separate validation tasks, whose query labels are used to estimate $\bar{\beta}_S$ and r_S. T^* of the three runs is given in parentheses. The last column is the numerically computed limit of the iteration ($T \to \infty$).

Dataset	n	$T = 1$	$T = 10$	$T = 50$	$T = 200$	val (T^*)	bound (T^*)	converged
Sinusoidal	5	1.324	0.287	0.127	0.087	0.087 (200)	0.117 (41/73/87)	50.1
	10	0.955	0.067	0.030	0.027	0.027 (84/103/200)	0.044 (14/19/30)	0.951
	20	0.555	0.019	0.016	0.016	0.016 (34/39/60)	0.017 (11/14/35)	0.150
	40	0.494	0.016	0.014	0.014	0.014 (54/44/64)	0.015 (8/12/28)	0.030
Omega	5	2.293	1.237	0.972	0.850	0.850 (200)	0.901 (67/117/110)	240
	10	1.479	0.297	0.152	0.120	0.122 (116/187/200)	0.165 (28/40/45)	29.6
	20	1.278	0.093	0.044	0.040	0.040 (200)	0.050 (25/36/37)	0.706
	40	1.107	0.063	0.033	0.030	0.030 (200)	0.038 (23/39/31)	0.072
Misfit	5	1.691	0.486	0.382	0.503	0.381 (43/49/51)	0.584 (6/7/7)	147
	10	1.116	0.248	0.226	0.258	0.224 (20/19/25)	0.343 (4/5/5)	6.735
	20	0.812	0.175	0.174	0.184	0.171 (15/17/25)	0.243 (3/3/4)	1.872
	40	0.670	0.159	0.158	0.161	0.157 (15/34/42)	0.194 (3)	0.606
Oracle	5	2.081	1.378	1.292	1.275	1.274 (132/182/200)	1.296 (52/45/39)	224
	10	1.645	0.548	0.400	0.382	0.381 (99/115/99)	0.447 (29/23/21)	98.6
	20	1.309	0.227	0.126	0.112	0.118 (95/70/84)	0.155 (29/20/18)	3.319
	40	1.172	0.121	0.058	0.049	0.055 (200/46/47)	0.072 (200/17/16)	0.091
Structured	5	0.830	0.710	0.690	0.685	0.687 (83/96/122)	0.691 (51/38/62)	12.5
	10	0.668	0.389	0.353	0.353	0.351 (120/200/128)	0.359 (36/23/73)	153
	20	0.583	0.184	0.130	0.119	0.120 (134/92/200)	0.129 (62/24/162)	109
	40	0.471	0.076	0.048	0.039	0.039 (200/156/200)	0.041 (126/27/198)	11.2
Air quality	5	0.058	0.051	0.050	0.056	0.050 (16/18/19)	0.053 (6/7/6)	0.130
	10	0.045	0.036	0.034	0.034	0.033 (112/78/80)	0.036 (11/15/13)	0.145
	20	0.057	0.040	0.034	0.031	0.031 (200)	0.041 (9)	0.197
METR-LA	5	0.371	0.383	0.393	0.434	0.346 (0)	0.386 (16/26/22)	73.3
	10	0.321	0.324	0.328	0.346	0.321 (0/10/2)	0.325 (15/23/19)	246
	20	0.355	0.354	0.353	0.360	0.355 (1)	0.353 (23/31/24)	19.5
PEMS-BAY	5	0.052	0.051	0.055	0.060	0.051 (7/24/12)	0.051 (10/15/7)	0.527
	10	0.043	0.042	0.041	0.042	0.041 (12/19/14)	0.041 (200/200/14)	1.358
	20	0.044	0.043	0.043	0.043	0.043 (19/38/17)	0.043 (200/200/23)	1.524

Table 7: Same as Table 6, restricted to $0 \leq T \leq 30$, the range used in Tables 1 and 9.

Dataset	n	$T=1$	$T=10$	$T=20$	$T=30$	val (T^*)	bound (T^*)	converged
Sinusoidal	5	1.324	0.287	0.193	0.158	0.158 (30)	0.158 (30)	50.1
	10	0.955	0.067	0.042	0.035	0.035 (30)	0.044 (14/19/30)	0.951
	20	0.555	0.019	0.017	0.016	0.016 (30)	0.017 (11/14/30)	0.150
	40	0.494	0.016	0.015	0.014	0.014 (30)	0.015 (8/12/28)	0.030
Omega	5	2.293	1.237	1.104	1.041	1.041 (30)	1.041 (30)	240
	10	1.479	0.297	0.209	0.178	0.178 (30)	0.179 (28/30/30)	29.6
	20	1.278	0.093	0.059	0.050	0.050 (30)	0.052 (25/30/30)	0.706
	40	1.107	0.063	0.044	0.038	0.038 (30)	0.039 (23/30/30)	0.072
Misfit	5	1.691	0.486	0.398	0.380	0.380 (30)	0.584 (6/7/7)	147
	10	1.116	0.248	0.225	0.222	0.224 (20/19/25)	0.343 (4/5/5)	6.735
	20	0.812	0.175	0.171	0.172	0.171 (15/17/25)	0.243 (3/3/4)	1.872
	40	0.670	0.159	0.157	0.157	0.157 (15/30/30)	0.194 (3)	0.606
Oracle	5	2.081	1.378	1.331	1.311	1.311 (30)	1.311 (30)	224
	10	1.645	0.548	0.463	0.430	0.430 (30)	0.447 (29/23/21)	98.6
	20	1.309	0.227	0.159	0.140	0.140 (30)	0.155 (29/20/18)	3.319
	40	1.172	0.121	0.076	0.065	0.065 (30)	0.079 (25/17/16)	0.091
Structured	5	0.830	0.710	0.699	0.694	0.694 (30)	0.694 (30)	12.5
	10	0.668	0.389	0.367	0.359	0.359 (30)	0.361 (30/23/30)	153
	20	0.583	0.184	0.153	0.141	0.141 (30)	0.143 (30/24/30)	109
	40	0.471	0.076	0.059	0.053	0.053 (30)	0.053 (30/27/30)	11.2
Air quality	5	0.058	0.051	0.050	0.050	0.050 (16/18/19)	0.053 (6/7/6)	0.130
	10	0.045	0.036	0.035	0.034	0.034 (30)	0.036 (11/15/13)	0.145
	20	0.057	0.040	0.037	0.036	0.036 (30)	0.041 (9)	0.197
METR-LA	5	0.371	0.383	0.385	0.388	0.346 (0)	0.386 (16/26/22)	73.3
	10	0.321	0.324	0.325	0.326	0.321 (0/10/2)	0.325 (15/23/19)	246
	20	0.355	0.354	0.353	0.353	0.355 (1)	0.353 (23/30/24)	19.5
PEMS-BAY	5	0.052	0.051	0.052	0.053	0.051 (7/24/12)	0.051 (10/15/7)	0.527
	10	0.043	0.042	0.041	0.041	0.041 (12/19/14)	0.041 (30/30/14)	1.358
	20	0.044	0.043	0.043	0.043	0.043 (19/30/17)	0.043 (30/30/23)	1.524

Table 8: Summary of the test query MSE of the proposed model when the depth is selected on validation tasks over $0 \leq T \leq 30$ (as in Table 1) or over $0 \leq T \leq 100$, by the validation MSE (val) or by the bound (bound). Mean over three training runs with the standard deviation in subscript; the selected depths of the three runs are given in the last columns. The test tasks are those of Table 1.

	val		bound		T^*, $T \leq 100$ (val, bound)	
Dataset	$T \leq 30$	$T \leq 100$	$T \leq 30$	$T \leq 100$	val	bound
Sinusoidal	$0.0355_{\pm 0.0069}$	$0.0272_{\pm 0.0070}$	$0.0439_{\pm 0.0152}$	$0.0439_{\pm 0.0152}$	84/100/100	14/19/30
Omega	$0.1780_{\pm 0.0012}$	$0.1296_{\pm 0.0071}$	$0.1792_{\pm 0.0019}$	$0.1655_{\pm 0.0136}$	100	28/40/45
Misfit	$0.2236_{\pm 0.0057}$	$0.2236_{\pm 0.0057}$	$0.3432_{\pm 0.0263}$	$0.3432_{\pm 0.0263}$	20/19/25	4/5/5
Oracle	$0.4297_{\pm 0.0121}$	$0.3808_{\pm 0.0100}$	$0.4467_{\pm 0.0047}$	$0.4467_{\pm 0.0047}$	99/100/99	29/23/21
Structured	$0.3592_{\pm 0.0161}$	$0.3501_{\pm 0.0138}$	$0.3615_{\pm 0.0191}$	$0.3586_{\pm 0.0223}$	100	36/23/73
Air quality	$0.0503_{\pm 0.0021}$	$0.0503_{\pm 0.0021}$	$0.0527_{\pm 0.0025}$	$0.0527_{\pm 0.0025}$	16/18/19	6/7/6
METR-LA	$0.3463_{\pm 0.0186}$	$0.3463_{\pm 0.0186}$	$0.3857_{\pm 0.0110}$	$0.3857_{\pm 0.0110}$	0	16/26/22
PEMS-BAY	$0.0514_{\pm 0.0012}$	$0.0514_{\pm 0.0012}$	$0.0513_{\pm 0.0012}$	$0.0513_{\pm 0.0012}$	7/24/12	10/15/7

Table 9: Test query MSE of gradient-based models as a function of the test-time depth T, averaged over three training runs. On the synthetic datasets, Ours and MAML are evaluated on the same 300 test tasks per run. On the real world datasets, we used 200 test tasks per run. The last column reports the depth selected on separate held-out tasks: 300 validation tasks for Ours, and two-fold cross-fitting over the test tasks for MAML. Best value per row in bold.

Dataset	$T = 1$	$T = 10$	$T = 20$	$T = 30$	selected T
Ours ($E = 10$); MSE at selected T by validation tasks					
Sinusoidal	0.9553	0.0673	0.0425	**0.0355**	**0.0355**
Omega	1.4791	0.2975	0.2086	**0.1780**	**0.1780**
Misfit	1.1162	0.2483	0.2247	**0.2220**	0.2236
Oracle	1.6446	0.5484	0.4634	**0.4297**	**0.4297**
Structured	0.6680	0.3892	0.3674	**0.3592**	**0.3592**
Air quality	0.0580	0.0515	0.0501	**0.0497**	0.0503
METR-LA	0.3706	0.3830	0.3855	0.3879	**0.3463**
PEMS-BAY	0.0515	**0.0513**	0.0519	0.0528	0.0514
MAML ($E = 1$); MSE at selected T by two-fold cross-fitting					
Sinusoidal	1.0887	0.4998	0.4545	**0.4380**	0.4469
Omega	1.4640	0.7773	0.7253	**0.6893**	0.6971
Misfit	1.2508	0.6670	0.6265	**0.6148**	0.6185
Oracle	1.5513	0.7737	0.7325	**0.6944**	0.7227
Structured	0.7363	0.6339	0.6419	0.6394	**0.6306**
Air quality	0.0519	0.0459	0.0425	**0.0423**	0.0432
METR-LA	**0.3677**	0.6507	0.7111	0.6428	**0.3677**
PEMS-BAY	0.0615	0.0623	0.0642	0.0689	**0.0539**
MAML ($E = 10$); MSE at selected T by two-fold cross-fitting					
Sinusoidal	1.6803	0.3776	0.3144	**0.2938**	0.2949
Omega	1.8701	0.6115	0.5390	**0.5101**	**0.5101**
Misfit	2.1159	0.5088	0.4500	**0.4383**	0.4431
Oracle	7.8177	1.2006	1.0012	**0.9254**	0.9257
Structured	0.7416	0.5140	0.5007	0.4961	**0.4950**
Air quality	0.1778	0.0479	0.0447	**0.0440**	0.0442
METR-LA	1.3239	0.4078	0.4198	0.4347	**0.4030**
PEMS-BAY	1.8973	**0.0555**	0.0578	0.0584	0.0565

Table 10: Training depth E and test depth T (one run per value, $n = 10$, 300 test tasks). Evaluation follows the methods in Table 6.

Dataset	E	$T = 1$	$T = 10$	$T = 50$	$T = 200$	val (T^*)	bound (T^*)	converged
Sinusoidal	1	0.879	0.136	0.082	0.071	0.071 (200)	0.092 (31)	3.562
	5	0.837	0.111	0.070	0.059	0.059 (163)	0.084 (23)	0.874
	10	0.964	0.083	0.052	0.059	0.052 (40)	0.089 (9)	1.571
	20	1.293	0.096	0.042	0.039	0.040 (69)	0.078 (13)	2.174
	30	1.498	0.105	0.035	0.024	0.024 (165)	0.044 (32)	1.703
Omega	1	1.360	0.419	0.347	0.374	0.347 (57)	0.351 (35)	52.1
	5	1.186	0.297	0.176	0.137	0.137 (200)	0.196 (33)	57.4
	10	1.433	0.306	0.171	0.141	0.141 (190)	0.197 (29)	28.2
	20	1.647	0.343	0.195	0.162	0.162 (200)	0.164 (173)	41.2
	30	1.689	0.317	0.138	0.103	0.103 (200)	0.115 (97)	42.6

F.5 SHARED INITIAL HEAD.

Figure 4 starts the same T-step adaptation from four different head initializations on the learned model of each main run, on 300 test tasks. The proposed head is the Stage II head. The task-mean head averages ridge heads, fitted using 256 meta-training support datasets with ridge term $10^{-3}\lambda_{\max}(\Phi_S^\top \Phi_S)$. The zero head sets all coefficients to zero and the random head is a Gaussian vector rescaled to the norm of the proposed head.

Results. On the first four synthetic datasets (Sinusoidal, Omega, Misfit and Oracle) the error curves for the three heads (Proposed, task-mean and zero) almost coincide, as expected for task distributions with a zero mean function. The random head is indistinguishable on Sinusoidal and Misfit but is worse on Omega and Oracle. On Structured-Oracle, the task-mean head reaches 0.26, the proposed head is 0.35 and the zero head 0.57 at the depth with lowest error on test tasks. On the real world datasets, our proposed head reaches 0.048 on Air quality, 0.36 on METR-LA and 0.042 on PEMS-BAY, against 0.13, 0.95, 0.10 for the task-mean head and 0.16, 1.27 and 0.42 for the zero head. The gap is largest at small depths, for example 0.062 against 0.20 and 0.36 on Air quality at $T = 1$. The task-mean head is the best initialization on Structured-Oracle, but it yields higher query MSE on the real world datasets than our proposed head.

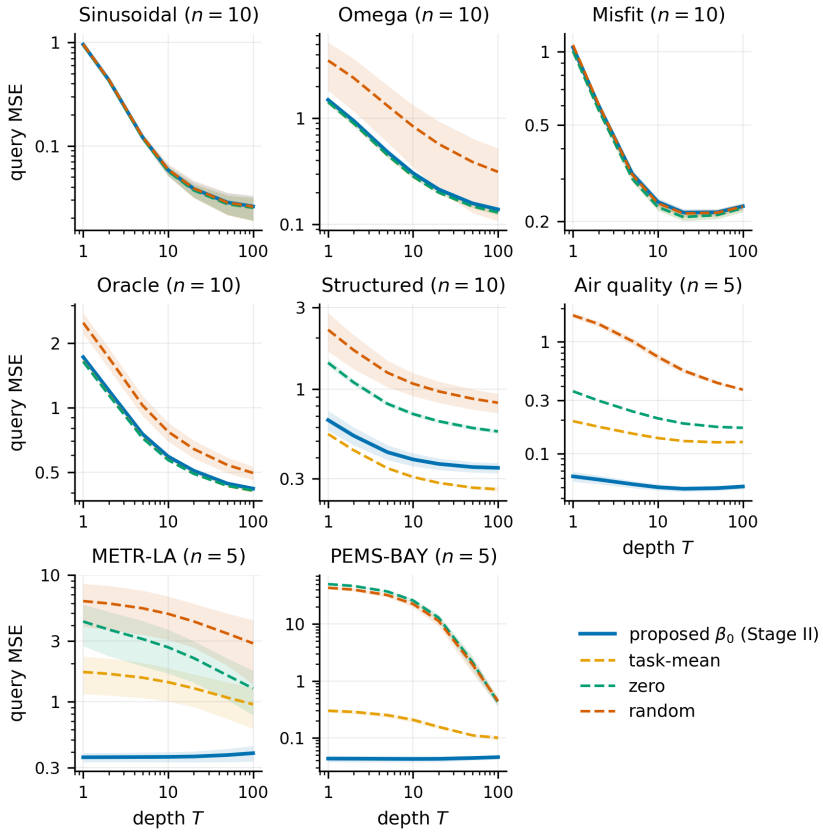

Figure 4: Initialization ablation on all families. The same T-step update is started from the Stage II head, the task-mean head, the zero head or a random head of the same norm, with the learned kernel and inducing points fixed.

F.6 STAGE I PRE-TRAINING

Table 11 compares the pre-trained inducing points with inducing points sampled randomly from Gaussian distribution. After constructing the inducing points, each model undergoes the meta-training and testing. The ablation uses the same setup as described in Table 5. Stage I gives lower query MSE

at the reported depths and under both depth-selection rules, compared to randomly selected inducing points.

Table 11: Effect of Stage I on the final model (one run per variant, 300 test tasks, the same tasks for both variants of a family). Test query MSE at fixed depths and at the depth selected on 300 validation tasks over $0 \leq T \leq 200$ by the validation error (val) and by the bound (bound).

Dataset	Inducing points	$T = 10$	$T = 20$	$T = 30$	val (T^*)	bound (T^*)
Sinusoidal	Stage I	0.053	0.035	0.032	0.030 (84)	0.037 (18)
	random, no Stage I	2.535	2.567	2.605	2.536 (11)	3.366 (1)
Omega	Stage I	0.337	0.236	0.203	0.157 (116)	0.203 (30)
	random, no Stage I	2.921	2.972	3.018	2.924 (11)	3.620 (1)
Misfit	Stage I	0.234	0.218	0.219	0.218 (20)	0.344 (4)
	random, no Stage I	2.701	2.729	2.770	2.700 (11)	3.522 (1)
Oracle	Stage I	0.552	0.465	0.430	0.374 (99)	0.435 (28)
	random, no Stage I	2.781	2.798	2.800	2.797 (7)	3.437 (2)
Structured	Stage I	0.394	0.370	0.360	0.345 (120)	0.356 (36)
	random, no Stage I	0.923	0.939	0.961	0.923 (10)	0.988 (1)

F.7 Depth selected by the bound

In this subsection we check whether the minimizer of the bound in Theorem 2 follows the actual error curve. The bound depends on unknown task quantities, which we estimate using additional labeled observations. The bound requires the reference $\bar{\beta}_S$ and the residual r_S with known noise level σ. For each test task, we instead estimate $\bar{\beta}_S$ and r_S with 200 additional query observations and score the selected depth on a separate query dataset of 200 observations. We compare this depth with a single depth tuned on another 300 validation tasks (which we used for the evaluation of our proposed model), and with fixed depths in $\{1, 5, 10, 20, 50, 75, 100\}$. The term regret refers to the ratio where its error on the separate query dataset divided by the lowest error over $1 \leq T \leq 200$ on that query dataset. The real world datasets are excluded here as the noise level is unknown and each task has only limited number of query points.

Results. Table 12 and Figure 5 summarize the regret values. The bound-selected depth has a median regret between 1.05 and 1.13 on synthetic datasets, except on Misfit dataset. On Sinusoidal, Oracle, and Structured-Oracle, the bound-selected depth gives a median regret ratio similar to the one used validation query MSE for the depth selection. $T = 1$ gives substantially higher median regret in all five settings.

Table 12: Median per-task regret of the depth T^* selected by minimizing plug-in conditional excess transfer risk bound (bound), the depth selected using the validation query MSE (val) and fixed depths.

Dataset	bound	tuned	$T = 1$	$T = 5$	$T = 10$	$T = 20$	$T = 50$	$T = 75$	$T = 100$
Sinusoidal	1.07	1.07	31.38	2.25	1.26	1.13	1.07	1.06	1.07
Omega	1.13	1.05	13.84	3.43	1.93	1.40	1.18	1.13	1.11
Oracle	1.12	1.13	8.70	2.51	1.66	1.30	1.17	1.14	1.13
Structured	1.05	1.04	2.15	1.28	1.14	1.10	1.07	1.06	1.05
Misfit	1.30	1.05	3.16	1.11	1.05	1.05	1.09	1.13	1.15

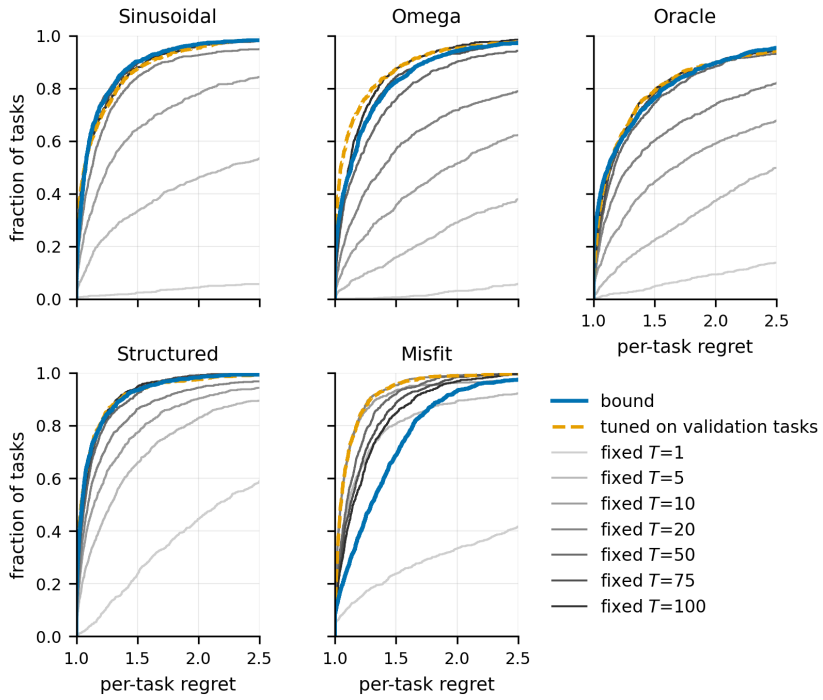

Figure 5: Empirical distribution of the per-task regret of the depth T^* selected by minimizing plug-in conditional excess transfer risk bound (bound), the depth selected using the validation query MSE (val) and fixed depths (900 test tasks per dataset, $n = 10$). A curve further to the upper left is better.

F.8 Number of inducing points.

Figure 6 compares the proposed model with $p \in \{8, 16, 32, 64, 128\}$ on Sinusoidal and Omega datasets. The error at the selected depth (val(T^*)) is lowest at the default setting $p = 32$, and is higher for larger p. This result is qualitatively consistent with an approximation-estimation trade-off in Corollary 5.

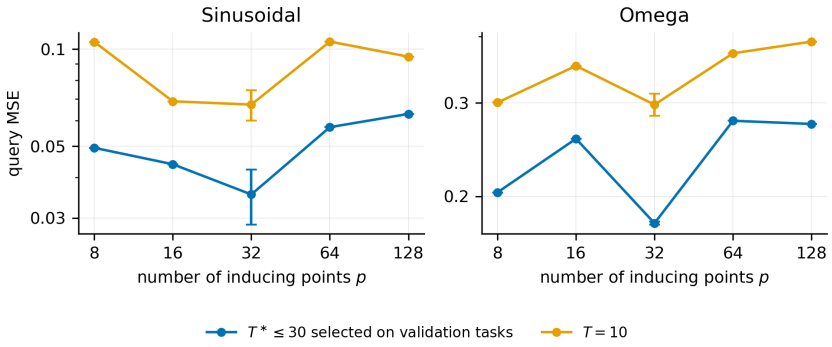

Figure 6: Test query MSE against the number of inducing points. Error bars show the standard deviation over the three runs at $p = 32$; other values use one run.

G LLM USAGE STATEMENT

In this work, we used generative AI tools to assist with the scientific writing and plotting figures. We have not used generative AI tools to generate the theoretical framework, or to formulate mathematical claims appearing in the appendix except for LaTeX typesetting and language editing. All AI-assisted writings were reviewed and inspected by the authors, and all edited texts were checked for correctness of the scientific claims. We also used generative AI tools for parts of the code to assist with implementation, refactoring, and evaluation scripts. All code and its outputs were reviewed and tested by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.