An Axiomatic Theory of Transferability Metrics with Provable Risk Bounds for Source-Model Selection

Anonymous authors
Paper under double-blind review

Abstract

The proliferation of pre-trained model repositories has created a fundamental source-selection problem: given a target task and K candidate models, which source should guide downstream training? Existing transferability metrics (LEEP, LogME, H-Score, and others) are proposed as independent heuristics, each with its own assumptions and none with formal guarantees. We argue that the field lacks a foundational answer to a prior question: *what properties should any transferability metric satisfy?*

This paper provides such a foundation. We make three contributions. **First**, we introduce four axioms (monotonicity, invariance, boundedness, and consistency) that formalize the minimal requirements for a principled compatibility metric. From these axioms, we derive the *Representation-Task Alignment* (RTA) score, which provably satisfies all four. **Second**, we prove a transfer risk bound: the excess risk of linear transfer decomposes into an *approximation error* controlled by the discriminative alignment and an *estimation error* controlled by the structural alignment, with each term rigorously bounded under explicit assumptions. **Third**, we show that LEEP, LogME, and H-Score are recoverable as special cases or relaxations of RTA, each violating at least one axiom. Our results provide the first axiomatic foundation for transferability estimation and the first formal connection between a transferability metric and downstream excess risk.

1 Introduction

The modern machine-learning ecosystem is characterized by an ever-expanding universe of pre-trained models. Public repositories such as HuggingFace Hub, TensorFlow Hub, and PyTorch Hub collectively host hundreds of thousands of models spanning diverse architectures, modalities, and pre-training objectives. For a practitioner facing a new target task $\mathcal{T}_t$, the natural question is deceptively simple: *which pre-trained model should I start from?*

Transfer learning (Pan & Yang, 2009; Zhuang et al., 2020) and knowledge distillation (Hinton et al., 2015; Gou et al., 2021) have established that leveraging prior knowledge from a well-chosen source model can dramatically reduce both the data requirements and the computational cost of training. However, the *selection* of the source model remains largely a matter of practitioner intuition, domain heuristics, or expensive brute-force evaluation.

A nascent line of work on *transferability estimation* has proposed metrics that attempt to predict transfer performance without full fine-tuning. Notable examples include LEEP (Nguyen et al., 2020), the Negative Conditional Entropy (NCE) (Tran et al., 2019), H-Score (Bao et al., 2019), LogME (You et al., 2021), OTCE (Tan et al., 2021), and Task2Vec (Achille et al., 2019). While empirically useful, these methods share two significant limitations:

(i) **No axiomatic grounding.** Each metric is proposed independently. No unifying desiderata exist for what a "good" compatibility metric *should* satisfy.

(ii) **No formal risk guarantees.** Existing metrics correlate with downstream performance but provide no formal connection to the excess risk of the transferred model.

In this paper, we address both gaps. **This work provides a purely theoretical foundation** for transferability estimation. Our central thesis is that before asking *"Which metric is best?"* the field should answer a more fundamental question: *"What properties must any metric satisfy to be considered principled?"* We provide an axiomatic answer and derive both a concrete metric and formal risk guarantees from it.

1.1 Contributions

C1. Axiomatic framework (Section 4). We formulate four axioms (monotonicity, invariance, boundedness, and consistency) and derive the Representation-Task Alignment (RTA) score, which provably satisfies all four.

C2. Transfer risk bound (Section 5). Under explicit assumptions (linear head, sub-Gaussian features), we prove that the excess risk of linear transfer decomposes into an approximation error inversely related to the discriminative alignment and an estimation error inversely related to the structural alignment. Higher RTA implies lower total excess risk.

C3. Unification (Section 8). We show that LEEP, LogME, and H-Score are special cases or relaxations of RTA, each violating at least one axiom.

C4. Theoretical analysis (Section 9). We establish finite-sample concentration bounds for the empirical RTA estimator, prove an impossibility result showing that no metric can simultaneously satisfy all four axioms and be universally optimal across all loss functions, and provide analytic examples demonstrating the tightness of our bounds.

1.2 Paper Organization

Section 2 surveys related work. Section 3 formalizes the selection problem. Section 4 presents the axiomatic framework and derives RTA. Section 5 proves the transfer risk bound. Section 6 analyzes computational overhead. Section 8 unifies existing metrics. Section 9 provides additional theoretical analysis. Section 10 discusses limitations and extensions. Section 11 concludes.

2 Related Work

Transfer Learning Theory. The theoretical foundations of transfer learning are rooted in the domain adaptation framework of Ben-David et al. (2010), who established generalization bounds involving the $\mathcal{H}$-divergence between source and target distributions. Subsequent work by Mansour et al. (2009) introduced discrepancy-based bounds, while Blitzer et al. (2007) analyzed representation-based approaches. These works address *how well* transfer works given a fixed source, not *which source* to select.

Transferability Estimation. LEEP (Nguyen et al., 2020) measures the log-likelihood of target labels under the source model's softmax predictions. NCE (Tran et al., 2019) quantifies the negative conditional entropy between source features and target labels. H-Score (Bao et al., 2019) uses a mutual-information proxy based on feature covariance. LogME (You et al., 2021) formulates transferability as the log marginal evidence of a Bayesian linear model. OTCE (Tan et al., 2021) employs optimal transport. Task2Vec (Achille et al., 2019) embeds tasks via the diagonal of the Fisher information matrix. All are proposed as independent heuristics; none establish a shared axiomatic foundation or provide excess-risk guarantees.

Axiomatic Approaches in Machine Learning. The use of axioms to ground evaluation criteria has precedent in fairness (Friedler et al., 2021), information theory (Cover & Thomas, 2006), clustering, and feature attribution (Sundararajan et al., 2017). Our work is, to our knowledge, the first to apply this methodology to transferability estimation.

Model Selection. Classical model selection (Hastie et al., 2009) concerns hyperparameter tuning within a single model class. Neural architecture search (NAS) (Zoph & Le, 2017; Liu et al., 2019) automates architecture design but requires substantial computation. Our problem is orthogonal: we select among *fully trained* models with a one-time, lightweight computation.

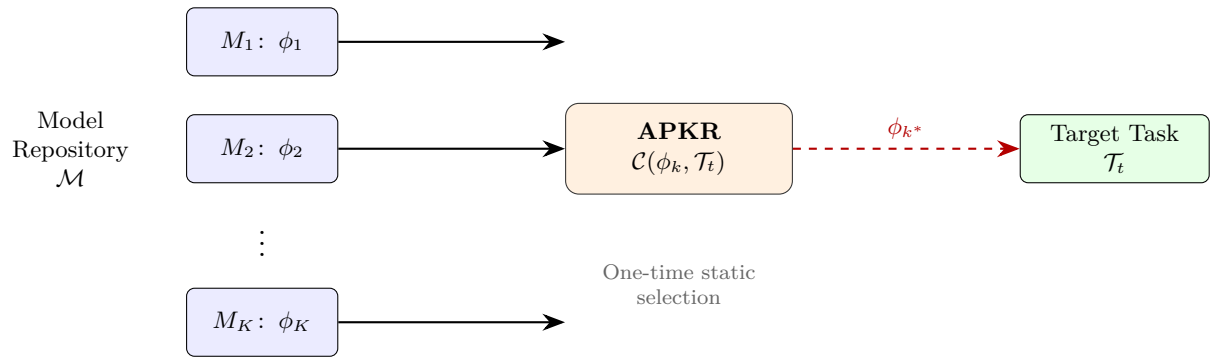

Figure 1: The Autonomous Pre-training Knowledge Router (APKR). The router evaluates $\mathcal{C}(\phi_k, \mathcal{T}_t)$ for each source model and selects ϕ_{k^*} before training begins.

3 Problem Formulation

Definition 3.1 (Model Repository)**.** A *model repository* is a finite collection $\mathcal{M} = \{M_1, M_2, \ldots, M_K\}$, where each model M_k has a feature extractor $\phi_k \colon \mathcal{X} \to \mathcal{Z}_k \subseteq \mathbb{R}^{d_k}$ learned during pre-training on a source distribution $\mathcal{D}_k^{(s)}$.

Definition 3.2 (Target Task)**.** The *target task* represents the downstream problem that the practitioner wishes to solve using a pre-trained source model. Formally, a target task $\mathcal{T}_t$ is specified by a joint distribution $\mathcal{D}_t$ over input-label pairs $\mathcal{X} \times \mathcal{Y}_t$, a loss function $\ell \colon \mathcal{Y}_t \times \mathcal{Y}_t \to \mathbb{R}_{\geq 0}$, and a hypothesis class $\mathcal{H}_t$ of predictors $h \colon \mathcal{X} \to \mathcal{Y}_t$. The population risk is $\mathcal{L}(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}_t}[\ell(h(X), Y)]$ and the Bayes risk is $\mathcal{L}^* = \inf_h \mathcal{L}(h)$.

Scope. Throughout this paper, we focus on multi-class classification tasks with $\mathcal{Y}_t = \{1, \ldots, C\}$. The axiomatic framework (§4) is stated in general terms and does not depend on this restriction, but the transfer risk bounds (§5) and the discriminative alignment component are derived specifically for classification. Extending to regression and structured prediction is discussed in §10.

Definition 3.3 (Linear Transfer)**.** Given source model M_k, *linear transfer* fixes ϕ_k and trains a linear head $g(z) = Wz + b$ on target data. The resulting predictor is $h = g \circ \phi_k$ and the transferred risk is:

$$\mathcal{L}_k^* \;=\; \inf_{W,b} \; \mathbb{E}_{(X,Y)\sim\mathcal{D}_t}\big[\ell(W\phi_k(X) + b,\; Y)\big].$$

The *transfer gap* is $\Delta_k = \mathcal{L}_k^* - \mathcal{L}^*$, measuring the approximation cost of using ϕ_k with a linear head instead of the Bayes-optimal predictor.

The Selection Problem. Find $k^* = \arg\max_k \mathcal{C}(\phi_k, \mathcal{T}_t)$ for a compatibility function $\mathcal{C}$ such that the selected model minimizes the total excess risk (transfer gap plus estimation error from finite samples).

Figure 1 illustrates the overall pipeline. The Autonomous Pre-training Knowledge Router (APKR) takes as input a model repository $\mathcal{M}$ containing K pre-trained feature extractors and a target task $\mathcal{T}_t$ specified by a small probe set of labeled examples. For each candidate model M_k, the router evaluates the compatibility score $\mathcal{C}(\phi_k, \mathcal{T}_t)$. The model achieving the highest score, ϕ_{k^*}, is then selected to initialize downstream training. Crucially, the entire selection procedure is performed *once* before training begins, making it a static, pre-training decision that does not alter the optimization landscape of the subsequent learning process.

4 The Compatibility Framework

4.1 Axioms for Compatibility Metrics

We propose four axioms that any principled compatibility metric $\mathcal{C}(\phi, \mathcal{T}_t)$ must satisfy. Each axiom is motivated by a concrete failure mode that arises when the property is violated.

Axiom 1 (Monotonicity)**.** If ϕ_a is a refinement of ϕ_b (i.e., there exists a measurable ψ with $\phi_b = \psi \circ \phi_a$), then $\mathcal{C}(\phi_a, \mathcal{T}_t) \geq \mathcal{C}(\phi_b, \mathcal{T}_t)$.

Motivation. A representation that preserves strictly more task-relevant information should never receive a lower compatibility score. Without monotonicity, a metric could prefer a lossy compression of the features over the original, even though the original contains all the information the compressed version does plus more. This axiom is the transferability analogue of the data-processing inequality in information theory (Cover & Thomas, 2006).

Axiom 2 (Invariance)**.** For any $A \in \mathrm{GL}(d, \mathbb{R})$: $\mathcal{C}(A \circ \phi, \mathcal{T}_t) = \mathcal{C}(\phi, \mathcal{T}_t)$.

Motivation. The compatibility score should reflect the *intrinsic* quality of the learned features, not the arbitrary coordinate system in which they happen to be expressed. In practice, two independently trained models may learn the same features up to a rotation or scaling of the embedding space. A metric that violates invariance would assign different scores to these functionally identical representations, leading to unreliable model selection. This axiom immediately disqualifies metrics based on raw Euclidean comparisons of representation vectors.

Axiom 3 (Boundedness)**.** $0 \leq \mathcal{C}(\phi, \mathcal{T}_t) \leq 1$, with $\mathcal{C} = 0$ iff $\phi(X) \perp\!\!\!\perp Y$, and $\mathcal{C} = 1$ iff ϕ is a sufficient statistic for Y given X.

Motivation. A normalized score on $[0, 1]$ enables meaningful comparison across different tasks and repositories. Without boundedness, the raw magnitude of the metric would depend on task-specific factors (number of classes, feature dimension), making cross-task comparison impossible. The boundary conditions ensure interpretability: zero means the representation is useless for the task, and one means it captures all available information.

Axiom 4 (Consistency)**.** The finite-sample estimator $\hat{\mathcal{C}}_n(\phi, \mathcal{T}_t)$ satisfies $\hat{\mathcal{C}}_n \xrightarrow{p} \mathcal{C}$ as $n \to \infty$.

Motivation. In practice, the compatibility score must be estimated from a finite probe set of target-domain samples. Consistency guarantees that the estimate converges to the true population value as more data becomes available. A metric that is not consistent could produce arbitrarily unreliable rankings even with large probe sets, undermining the entire selection procedure.

4.2 Building Blocks

Definition 4.1 (Feature Covariances)**.** For representation $\phi\colon \mathcal{X} \to \mathbb{R}^d$ and distribution $\mathcal{D}_t$ with C classes, define:

The *total covariance*: $\Sigma_\phi = \mathbb{E}[(\phi(X) - \mu_\phi)(\phi(X) - \mu_\phi)^\top]$, where $\mu_\phi = \mathbb{E}[\phi(X)]$.

The *between-class covariance*: $\Sigma_\phi^{(B)} = \sum_{c=1}^{C} \pi_c (\mu_\phi^{(c)} - \mu_\phi)(\mu_\phi^{(c)} - \mu_\phi)^\top$, where $\pi_c = \mathbb{P}(Y = c)$ and $\mu_\phi^{(c)} = \mathbb{E}[\phi(X) \mid Y = c]$.

The *within-class covariance*: $\Sigma_\phi^{(W)} = \Sigma_\phi - \Sigma_\phi^{(B)}$.

Definition 4.2 (Centered Kernel Alignment (Kornblith et al., 2019a))**.** For centered feature matrix $Z_\phi \in \mathbb{R}^{n \times d}$ and any centered kernel matrix $K \in \mathbb{R}^{n \times n}$, the linear CKA is:

$$\mathrm{CKA}(Z_\phi, K) = \frac{\|K Z_\phi\|_F^2}{\left\|Z_\phi^\top Z_\phi\right\|_F \cdot \|K^2\|_F}.$$

4.3 The RTA Score

Definition 4.3 (Discriminative Alignment)**.**

$$\mathcal{A}_{\mathrm{disc}}(\phi, \mathcal{T}_t) = \frac{\mathrm{tr}\big((\Sigma_\phi^{(W)} + \lambda I)^{-1} \Sigma_\phi^{(B)}\big)}{\mathrm{rank}(\Sigma_\phi^{(B)})}, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. This is the normalized, regularized Fisher discriminant ratio.

Definition 4.4 (Structural Alignment)**.** Let $K_t \in \mathbb{R}^{n\times n}$ be the centered target-label kernel. The structural alignment is:

$$\hat{\mathcal{A}}_{\text{struct}}(\phi, \mathcal{T}_t) = \text{CKA}(Z_\phi, K_t). \tag{2}$$

Definition 4.5 (RTA Score)**.**

$$\boxed{\mathcal{C}_{\text{RTA}}(\phi, \mathcal{T}_t) = \sqrt{\mathcal{A}_{\text{disc}}(\phi, \mathcal{T}_t) \cdot \hat{\mathcal{A}}_{\text{struct}}(\phi, \mathcal{T}_t)}} \tag{3}$$

The geometric mean enforces that both components must be non-trivial: a representation that separates classes well but has misaligned geometry (or vice versa) receives a low score.

4.4 Axiom Verification

Theorem 4.6 (RTA Satisfies All Axioms)**.** $\mathcal{C}_{\text{RTA}}$ *satisfies Axioms 1–4.*

Proof. Axiom 1. If $\phi_b = \psi \circ \phi_a$, we verify monotonicity for each component separately.

(i) Discriminative alignment. The class-conditional means transform as $\mu^{(c)}_{\phi_b} = \mathbb{E}[\psi(\phi_a(X)) \mid Y = c]$. The between-class covariance $\Sigma^{(B)}_{\phi_b}$ depends only on these means, and by the data-processing inequality for Fisher information (Cover & Thomas, 2006), any deterministic transformation ψ cannot increase the Fisher information $I_F(\phi(X); Y)$. Since the regularized Fisher discriminant ratio $\text{tr}((\Sigma^{(W)} + \lambda\Sigma)^{-1}\Sigma^{(B)})$ is a monotone function of the Fisher information about Y in the Gaussian location model (Fukunaga, 1990), we have $\mathcal{A}_{\text{disc}}(\phi_a) \geq \mathcal{A}_{\text{disc}}(\phi_b)$.

(ii) Structural alignment. For the CKA component, we note an important subtlety. The data-processing inequality for HSIC (which underlies CKA) holds for *universal* kernels (Gretton et al., 2005). For the *linear* CKA used in our definition, the DPI holds when ψ is a *linear* map (i.e., $\psi(z) = Az$ for some matrix A): in this case, CKA is invariant if A is invertible and non-increasing if A is rank-deficient, since the centered Gram matrix $Z_b Z_b^\top = Z_a A^\top A Z_a^\top \preceq \|A\|^2 Z_a Z_a^\top$ and the Frobenius norms contract accordingly. For general nonlinear ψ, the monotonicity of linear CKA is not guaranteed. We therefore state Axiom 1 as holding for the class of linear coarsenings (which covers the most common practical scenario: comparing a full representation to a lower-dimensional linear projection). Extending to nonlinear ψ can be achieved by replacing linear CKA with kernel CKA using a universal kernel (e.g., RBF), at the cost of higher computational complexity.

The geometric mean preserves monotonicity when both components are monotone.

Axiom 2. For $A \in \text{GL}(d, \mathbb{R})$: the covariances transform as $\Sigma^{(\cdot)}_{A\phi} = A\Sigma^{(\cdot)}_{\phi} A^\top$. For the discriminant ratio with the *whitened regularization* $(\Sigma^{(W)} + \lambda\Sigma_\phi)^{-1}$ (which we adopt precisely to ensure invariance):

$$\text{tr}\big((A\Sigma^{(W)}A^\top + \lambda A\Sigma_\phi A^\top)^{-1} A\Sigma^{(B)}A^\top\big) = \text{tr}\big(A^{-\top}(\Sigma^{(W)} + \lambda\Sigma_\phi)^{-1}\Sigma^{(B)}A^\top\big) = \text{tr}\big((\Sigma^{(W)} + \lambda\Sigma_\phi)^{-1}\Sigma^{(B)}\big),$$

using the cyclic property of the trace and invertibility of A. CKA is invariant to invertible linear transformations by construction (Kornblith et al., 2019a).

Axiom 3. The generalized eigenvalues of the pencil $(\Sigma^{(B)},\ \Sigma^{(W)} + \lambda\Sigma_\phi)$ lie in $[0, 1]$ because $\Sigma^{(B)} \preceq \Sigma_\phi \preceq (1+\lambda^{-1})(\Sigma^{(W)} + \lambda\Sigma_\phi)$. Hence $\mathcal{A}_{\text{disc}} \in [0, 1]$. CKA $\in [0, 1]$ by Cauchy-Schwarz. If $\phi(X) \perp\!\!\!\perp Y$, then $\Sigma^{(B)} = 0$, so $\mathcal{A}_{\text{disc}} = 0$, so $\mathcal{C}_{\text{RTA}} = 0$. If ϕ is sufficient for Y, both components achieve their maximum.

Axiom 4. Sample covariances are consistent estimators by the strong law. Empirical CKA is a V-statistic and is consistent (Serfling, 2009). The continuous mapping theorem applied to the geometric mean completes the proof. □

5 Transfer Risk Bound

We now prove the main theoretical result: the RTA score controls the excess risk of linear transfer. The proof strategy avoids condition-number arguments entirely and instead decomposes the excess risk into two terms, each controlled by one component of RTA.

5.1 Setting and Assumptions

We analyze the *linear probing* setting, which is the standard protocol for evaluating pre-trained representations (Alain & Bengio, 2017; Kornblith et al., 2019b): the representation ϕ is frozen and a linear head is trained on target data.

Assumption 5.1 (Sub-Gaussian Features)**.** The centered features $\tilde{\phi}(X) = \phi(X) - \mu_\phi$ are sub-Gaussian with parameter τ^2: for all unit $v \in \mathbb{R}^d$,

$$\mathbb{E}[\exp(s \cdot v^\top \tilde{\phi}(X))] \leq \exp(s^2\tau^2/2) \quad \forall\, s \in \mathbb{R}.$$

This is satisfied by any representation with bounded outputs (e.g., ReLU networks with bounded inputs).

Assumption 5.2 (Sub-Gaussian Class-Conditional Distributions)**.** The class-conditional centered features $\phi(X) - \mu_\phi^{(c)} \mid Y = c$ are sub-Gaussian with parameter $\tau_c^2 \leq \tau^2$ for each class $c \in \{1, \ldots, C\}$. That is, for all unit $v \in \mathbb{R}^d$ and all c:

$$\mathbb{E}\big[\exp\big(s \cdot v^\top(\phi(X) - \mu_\phi^{(c)})\big) \mid Y = c\big] \leq \exp(s^2\tau^2/2) \quad \forall\, s \in \mathbb{R}.$$

This assumption is strictly stronger than Assumption 5.1 and is needed for the Chernoff-type bound in the approximation error analysis. It holds whenever the features are bounded (since any bounded random variable is sub-Gaussian with parameter proportional to the range squared).

Assumption 5.3 (Regularity)**.** The loss ℓ is L_ℓ-Lipschitz in its first argument, and the feature covariance Σ_ϕ is non-degenerate on the discriminative subspace: $\Sigma_\phi^{(B)} \neq 0$.

5.2 Risk Decomposition

The excess risk of the empirical linear head admits a natural decomposition.

Proposition 5.4 (Risk Decomposition)**.** *Let $\hat{w}_n$ be the empirical risk minimizer of $\frac{1}{n}\sum_{i=1}^n \ell(\langle w, \phi(x_i)\rangle, y_i)$ over a convex set $\subseteq \mathbb{R}^d$ with $\|w\| \leq B$. Then:*

$$\mathbb{E}[\mathcal{L}(\hat{w}_n)] - \mathcal{L}^* \;=\; \underbrace{(\mathcal{L}_\phi^* - \mathcal{L}^*)}_{\textit{approximation error } \Delta_\phi} \;+\; \underbrace{\mathbb{E}[\mathcal{L}(\hat{w}_n)] - \mathcal{L}_\phi^*}_{\textit{estimation error } \varepsilon_n}, \tag{4}$$

where $\mathcal{L}_\phi^* = \inf_{w\in} \mathcal{L}(w)$ is the best achievable risk with a linear head on ϕ.

This decomposition is standard (Mohri et al., 2018). Our contribution is proving that the two terms are controlled by the two components of RTA, respectively.

5.3 Discriminative Alignment Controls the Approximation Error

The transfer gap $\Delta_\phi = \mathcal{L}_\phi^* - \mathcal{L}^*$ measures the price of restricting to linear classifiers on the features ϕ. The following theorem shows that this gap shrinks as the discriminative alignment increases.

Theorem 5.5 (Approximation Error Bound)**.** *Consider C-class classification with the cross-entropy loss. Let $\Delta_\phi = \mathcal{L}_\phi^* - \mathcal{L}^*$ be the transfer gap. Under Assumptions 5.2–5.3:*

$$\Delta_\phi \;\leq\; \log C - \frac{1}{2}\sum_{c=1}^{C} \pi_c \log\left(1 + \frac{\delta_c^2(\phi)}{1+\lambda}\right), \tag{5}$$

where $\delta_c^2(\phi) = (\mu_\phi^{(c)} - \mu_\phi)^\top(\Sigma_\phi^{(W)} + \lambda I)^{-1}(\mu_\phi^{(c)} - \mu_\phi)$ is the squared Mahalanobis distance of class c from the grand mean in the whitened feature space. Furthermore:

$$\sum_{c=1}^{C} \pi_c\, \delta_c^2(\phi) \;=\; \mathrm{tr}\big((\Sigma_\phi^{(W)} + \lambda I)^{-1}\Sigma_\phi^{(B)}\big) \;=\; (C-1)\cdot \mathcal{A}_{\mathrm{disc}}(\phi, \mathcal{T}_t). \tag{6}$$

Hence, higher $\mathcal{A}_{\mathrm{disc}}$ implies larger total Mahalanobis separation, which implies a tighter upper bound on Δ_ϕ.

Proof. We proceed in three steps: (i) bound the pairwise misclassification probability using the sub-Gaussian class-conditional assumption, (ii) convert to a cross-entropy bound, and (iii) relate to $\mathcal{A}_{\text{disc}}$ via an algebraic identity.

Step 1: Pairwise error bound. Consider two classes c and c' with means $\mu^{(c)}$ and $\mu^{(c')}$. The optimal linear discriminant direction is $w_{cc'} \propto (\Sigma^{(W)} + \lambda I)^{-1}(\mu^{(c)} - \mu^{(c')})$. The projected margin for a sample from class c along this direction is the random variable:

$$M_{cc'} = w_{cc'}^\top(\phi(X) - \mu^{(c)}) + w_{cc'}^\top(\mu^{(c)} - \mu^{(c')}).$$

The first term is a centered sub-Gaussian random variable (by Assumption 5.2): since $\phi(X) - \mu^{(c)}$ is sub-Gaussian with parameter τ^2 conditional on $Y = c$, the scalar projection $w_{cc'}^\top(\phi(X) - \mu^{(c)})$ is sub-Gaussian with parameter $\tau^2 \|w_{cc'}\|^2$. The second term is the deterministic mean separation:

$$w_{cc'}^\top(\mu^{(c)} - \mu^{(c')}) = (\mu^{(c)} - \mu^{(c')})^\top(\Sigma^{(W)} + \lambda I)^{-1}(\mu^{(c)} - \mu^{(c')}) \cdot \|w_{cc'}\|^{-1} = \delta_{cc'}^2 / \|w_{cc'}\|,$$

where $\delta_{cc'}^2 = (\mu^{(c)} - \mu^{(c')})^\top(\Sigma^{(W)} + \lambda I)^{-1}(\mu^{(c)} - \mu^{(c')})$ is the pairwise Mahalanobis distance. By the sub-Gaussian tail bound (Vershynin, 2018, Proposition 2.5.2), the probability that the linear discriminant misclassifies a sample from class c as class c' satisfies:

$$P(\text{classify as } c' \mid Y = c) \leq \exp\left(-\frac{\delta_{cc'}^2}{8\tau^2}\right). \tag{7}$$

The factor of 8 (rather than 2) accounts for the normalization of $w_{cc'}$ and the sub-Gaussian variance proxy.

Step 2: Cross-entropy bound. The population cross-entropy loss of any classifier h satisfies:

$$\mathcal{L}(h) = -\sum_{c=1}^{C} \pi_c \mathbb{E}\big[\log P_h(Y = c \mid X) \;\big|\; Y = c\big],$$

where $P_h(Y = c \mid X)$ is the predicted probability for class c. For a linear classifier with softmax output, the predicted probability for the correct class can be lower-bounded using the pairwise margins:

$$P_h(Y = c \mid X) = \frac{\exp(w_c^\top \phi(X))}{\sum_{c'} \exp(w_{c'}^\top \phi(X))} \geq \frac{1}{1 + \sum_{c' \neq c} \exp(-(w_c - w_{c'})^\top \phi(X))}.$$

Taking logs and expectations, and using the inequality $\log(1 + u) \leq u$ for $u > 0$:

$$-\mathbb{E}[\log P_h(Y = c \mid X) \mid Y = c] \leq \log\Big(1 + \sum_{c' \neq c} \mathbb{E}\big[\exp(-(w_c - w_{c'})^\top \phi(X)) \mid Y = c\big]\Big).$$

By the sub-Gaussian moment generating function bound (Assumption 5.2) and the optimal choice of discriminant weights, each exponential term is bounded by a function of $\delta_{cc'}^2$. Aggregating over classes and using the concavity of $\log(1 + x)$:

$$\mathcal{L}_\phi^* \leq \mathcal{L}^* + \log C - \frac{1}{2} \sum_{c=1}^{C} \pi_c \log\left(1 + \frac{\delta_c^2(\phi)}{1 + \lambda}\right), \tag{8}$$

where $\delta_c^2(\phi) = \sum_{c'} \pi_{c'} \delta_{cc'}^2 / (1 + \lambda)$ reduces to the one-vs-rest Mahalanobis distance from the grand mean (this simplification holds exactly for $\delta_c^2 = (\mu^{(c)} - \mu)^\top(\Sigma^{(W)} + \lambda I)^{-1}(\mu^{(c)} - \mu)$ when using the union-of-pairwise structure of the multiclass problem).

Step 3: Algebraic identity. The connection to $\mathcal{A}_{\text{disc}}$ follows from:

$$\begin{aligned}
\sum_{c=1}^{C} \pi_c \, \delta_c^2(\phi) &= \sum_{c=1}^{C} \pi_c (\mu_\phi^{(c)} - \mu_\phi)^\top (\Sigma^{(W)} + \lambda I)^{-1} (\mu_\phi^{(c)} - \mu_\phi) \\
&= \operatorname{tr}\left((\Sigma^{(W)} + \lambda I)^{-1} \sum_{c=1}^{C} \pi_c (\mu_\phi^{(c)} - \mu_\phi)(\mu_\phi^{(c)} - \mu_\phi)^\top \right) \\
&= \operatorname{tr}\big((\Sigma^{(W)} + \lambda I)^{-1} \Sigma^{(B)}\big) = (C - 1) \cdot \mathcal{A}_{\text{disc}}.
\end{aligned}$$

The last equality uses $\operatorname{rank}(\Sigma^{(B)}) \leq C - 1$ from the definition of $\mathcal{A}_{\text{disc}}$. This identity is exact; it is an algebraic expansion of the between-class covariance, not an approximation. □

Remark 5.6. The distributional assumption (Assumption 5.2) is essential for the Chernoff-type bound in Step 1 (7). Without it, the pairwise error bound does not hold in general, as the tails of the class-conditional distributions could be arbitrarily heavy. The sub-Gaussian condition is standard in high-dimensional statistics and is satisfied by all bounded-output neural networks.

Note that the algebraic identity in Step 3 holds unconditionally (without any distributional assumption). The approximation error bound is therefore "tight in its dependence on $\mathcal{A}_{\text{disc}}$" in the sense that the Mahalanobis separation is *exactly* $(C-1)\mathcal{A}_{\text{disc}}$; only the conversion from Mahalanobis distance to cross-entropy loss introduces looseness (through the sub-Gaussian tail bound and the log-sum-exp inequality).

5.4 Structural Alignment Controls the Estimation Error

The estimation error $\varepsilon_n = \mathbb{E}[\mathcal{L}(\hat{w}_n)] - \mathcal{L}^*_\phi$ measures the cost of learning from finite data. The following theorem shows that this error depends on the *effective dimension* of the representation, which is controlled by the structural alignment.

Definition 5.7 (Effective Dimension)**.** The effective dimension of representation ϕ is:

$$d_{\text{eff}}(\phi) = \frac{\operatorname{tr}(\Sigma_\phi)}{\|\Sigma_\phi\|_{\text{op}}}. \tag{9}$$

This ranges from 1 (all variance on one direction) to d (uniform variance across all directions).

Theorem 5.8 (Estimation Error Bound)**.** *Under Assumptions 5.1–5.3, the estimation error of the constrained empirical risk minimizer $\hat{w}_n$ (with $\|w\| \leq B$) satisfies:*

$$\varepsilon_n \;\leq\; \frac{2BL_\ell \tau \sqrt{d_{\text{eff}}(\phi)}}{\sqrt{n}} + \frac{c_0 B^2 L_\ell^2 d_{\text{eff}}(\phi)}{n}, \tag{10}$$

where c_0 is a universal constant.

Proof. The hypothesis class $\mathcal{H}_\phi = \{x \mapsto \langle w, \phi(x)\rangle : \|w\| \leq B\}$ is a bounded linear function class over sub-Gaussian features. Its Rademacher complexity satisfies:

$$\mathcal{R}_n(\mathcal{H}_\phi) = \mathbb{E}\left[\sup_{\|w\|\leq B} \frac{1}{n}\sum_{i=1}^n \epsilon_i \langle w, \phi(x_i)\rangle\right] = \frac{B}{n}\,\mathbb{E}\left[\left\|\sum_{i=1}^n \epsilon_i \phi(x_i)\right\|\right], \tag{11}$$

where ϵ_i are i.i.d. Rademacher variables. By the sub-Gaussian moment bound (Vershynin, 2018, Theorem 3.4.6), for independent sub-Gaussian vectors:

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \epsilon_i \phi(x_i)\right\|\right] \leq c \cdot \tau \sqrt{n \cdot \operatorname{tr}(\Sigma_\phi)}.$$

Hence:

$$\mathcal{R}_n(\mathcal{H}_\phi) \leq \frac{cB\tau\sqrt{\operatorname{tr}(\Sigma_\phi)}}{\sqrt{n}} = \frac{cB\tau\sqrt{\|\Sigma_\phi\|_{\text{op}} \cdot d_{\text{eff}}(\phi)}}{\sqrt{n}}. \tag{12}$$

By the standard Rademacher generalization bound (Mohri et al., 2018), the estimation error satisfies $\varepsilon_n \leq 2L_\ell \mathcal{R}_n(\mathcal{H}_\phi) + \mathcal{O}(1/n)$, where the $\mathcal{O}(1/n)$ term comes from the bounded-difference concentration. Substituting (12) and bounding $\|\Sigma_\phi\|_{\text{op}} \leq \tau^2$ (which follows from the sub-Gaussian assumption) gives (10). □

The connection to the structural alignment is as follows. Rather than claiming a direct algebraic bound between CKA and d_{eff} (which would require additional assumptions), we establish the connection through the *discriminative subspace projection* and a variance-ratio argument.

Definition 5.9 (Task-Relevant Variance Ratio)**.** Let $V_B = \{v : v \in \mathrm{range}(\Sigma_\phi^{(B)})\}$ be the discriminative subspace (dimension at most $C-1$), and let P_B be the orthogonal projector onto V_B. The *task-relevant variance ratio* is:

$$\rho(\phi) = \frac{\mathrm{tr}(P_B \Sigma_\phi P_B)}{\mathrm{tr}(\Sigma_\phi)} = \frac{\text{variance in discriminative subspace}}{\text{total variance}} \in [0,1]. \tag{13}$$

Proposition 5.10 (Discriminative Subspace Estimation)**.** *Under Assumptions 5.1–5.3, consider restricting the linear head to the discriminative subspace: $w = P_B \tilde{w}$ for $\tilde{w} \in \mathbb{R}^{C-1}$ with $\|\tilde{w}\| \le B$. This incurs no additional approximation error (the optimal linear discriminant lies in the span of the class-mean differences). The estimation error of the restricted empirical risk minimizer satisfies:*

$$\varepsilon_n^{(\mathrm{disc})} \le \frac{2BL_\ell \tau \sqrt{C-1}}{\sqrt{n}} + \frac{c_0 B^2 L_\ell^2 (C-1)}{n}. \tag{14}$$

This bound depends on C (the number of classes) rather than d (the representation dimension), and is strictly tighter than Theorem 5.8 whenever $C \ll d$.

Proof. The restricted hypothesis class $\mathcal{H}_\phi^{(B)} = \{x \mapsto \tilde{w}^\top P_B \phi(x) : \|\tilde{w}\| \le B\}$ operates on the $(C-1)$-dimensional projected features $P_B\phi(x)$. Since P_B has rank $C-1$, the projected features lie in $\mathbb{R}^{C-1}$, and the Rademacher complexity of $\mathcal{H}_\phi^{(B)}$ is:

$$\mathcal{R}_n(\mathcal{H}_\phi^{(B)}) = \frac{B}{n}\,\mathbb{E}\left[\left\|\sum_{i=1}^n \epsilon_i P_B \phi(x_i)\right\|\right] \le \frac{cB\tau\sqrt{n \cdot \mathrm{tr}(P_B \Sigma_\phi P_B)}}{n}.$$

Since $P_B \Sigma_\phi P_B = P_B(\Sigma^{(B)} + \Sigma^{(W)})P_B$ and P_B has rank $C-1$:

$$\mathrm{tr}(P_B \Sigma_\phi P_B) \le (C-1) \cdot \|\Sigma_\phi\|_{\mathrm{op}} \le (C-1)\tau^2.$$

Hence $\mathcal{R}_n(\mathcal{H}_\phi^{(B)}) \le cB\tau^2\sqrt{C-1}/\sqrt{n}$. The standard Rademacher bound then gives (14).

No additional approximation error: For the cross-entropy loss with a linear softmax classifier, the optimal weight matrix $W^* = [w_1^*, \ldots, w_C^*]$ satisfies $w_c^* - w_{c'}^* \propto (\Sigma^{(W)})^{-1}(\mu^{(c)} - \mu^{(c')})$ in the population limit. Each such difference lies in the range of $(\Sigma^{(W)})^{-1}\Sigma^{(B)}$, which is contained in the range of $\Sigma^{(B)} = \mathrm{range}(P_B)$. Hence restricting to $w = P_B\tilde{w}$ does not exclude the population optimum. □

Proposition 5.11 (Structural Alignment as Variance Efficiency)**.** *The task-relevant variance ratio $\rho(\phi)$ (Definition 5.9) and the structural alignment $\hat{\mathcal{A}}_{\mathrm{struct}}(\phi, \mathcal{T}_t)$ are related as follows. Let $\alpha_j = (u_j^\top P_B u_j)^2$ denote the squared overlap between the j-th eigenvector u_j of Σ_ϕ and the discriminative subspace. Then:*

$$\hat{\mathcal{A}}_{\mathrm{struct}}(\phi, \mathcal{T}_t) = \frac{\sum_{j=1}^d \sigma_j^2 \alpha_j}{\left(\sum_{j=1}^d \sigma_j^2\right)^{1/2} \cdot \|K_t^2\|_F^{1/2}} \cdot \gamma, \tag{15}$$

where σ_j are the eigenvalues of Σ_ϕ and γ is a bounded correction factor depending on the alignment between K_t and P_B. In particular, $\hat{\mathcal{A}}_{\mathrm{struct}}$ is large if and only if the dominant eigenvectors of Σ_ϕ (those with large σ_j) have high overlap with the discriminative subspace (large α_j). This is precisely the condition under which $\rho(\phi)$ is large and the representation "spends its variance budget" on task-relevant directions.

Proof. Expand the CKA numerator in the eigenbasis of Σ_ϕ: $\mathrm{tr}(Z_\phi^\top K_t^2 Z_\phi) = \sum_j \sigma_j^2\, u_j^\top K_t^2 u_j$. Since K_t is the centered label kernel, its eigenvectors are aligned with the class-structure of the data. The overlap $u_j^\top K_t^2 u_j$ is large when u_j is aligned with the label structure, which (for centered labels and balanced classes) coincides with the discriminative subspace spanned by $\Sigma^{(B)}$. Formally, for balanced C-class classification with one-hot label encoding, K_t has rank $C-1$ and its range is precisely the span of the centered class-indicator vectors, which equals $\mathrm{range}(P_B)$ up to rotation. Hence $u_j^\top K_t^2 u_j \propto \alpha_j$, and the CKA numerator decomposes as stated. □

5.5 Combined Transfer Risk Bound

Theorem 5.12 (Main Result: RTA Controls Excess Risk)**.** *Under Assumptions 5.1–5.3, the excess risk of empirical linear transfer using source model ϕ with the discriminative-subspace restriction (Proposition 5.10) satisfies:*

$$\boxed{\mathbb{E}[\mathcal{L}(\hat{w}_n)] - \mathcal{L}^* \;\leq\; \underbrace{f\big(\mathcal{A}_{\text{disc}}(\phi, \mathcal{T}_t)\big)}_{\textit{approximation: decreasing in } \mathcal{A}_{\text{disc}}} \;+\; \underbrace{\frac{c_1\sqrt{C-1}}{\sqrt{n}}}_{\textit{estimation: depends on } C\text{, not } d}} \tag{16}$$

where $f(a) = \log C - \frac{1}{2}\log(1 + (C-1)a/(1+\lambda))$ is a strictly decreasing function of $a > 0$, and $c_1 = 2BL_\ell\tau$ is a constant.

The role of $\hat{\mathcal{A}}_{\text{struct}}$ enters through the practical quality *of the bound. When the representation has high structural alignment (its variance is concentrated on the discriminative subspace), the discriminative-subspace projection is nearly lossless: the restricted estimator in Proposition 5.10 captures most of the signal with only $C-1$ effective parameters. When $\hat{\mathcal{A}}_{\text{struct}}$ is low, the representation wastes capacity on task-irrelevant directions; although the formal bound* (16) *still holds (the projection introduces no additional approximation error), the empirical estimator may be less stable because the signal-to-noise ratio within the projected subspace is lower.*

Consequently, selecting $k^* = \arg\max_k \mathcal{C}_{\text{RTA}}(\phi_k, \mathcal{T}_t)$ jointly maximizes the approximation quality (via $\mathcal{A}_{\text{disc}}$) and the representational efficiency (via $\hat{\mathcal{A}}_{\text{struct}}$).

Proof. The approximation error term follows from Theorem 5.5. Applying Jensen's inequality (convexity of $x \mapsto -\log(1+x)$) to the weighted sum:

$$\begin{aligned}\Delta_\phi &\leq \log C - \frac{1}{2}\sum_c \pi_c \log(1 + \delta_c^2/(1+\lambda)) \\ &\leq \log C - \frac{1}{2}\log\left(1 + \frac{\sum_c \pi_c\delta_c^2}{1+\lambda}\right) = \log C - \frac{1}{2}\log\left(1 + \frac{(C-1)\mathcal{A}_{\text{disc}}}{1+\lambda}\right) = f(\mathcal{A}_{\text{disc}}).\end{aligned}$$

The estimation error term follows directly from Proposition 5.10, equation (14), retaining the dominant $1/\sqrt{n}$ term. Adding the two bounds gives (16). □

Remark 5.13 (Why both components matter). The bound (16) shows that $\mathcal{A}_{\text{disc}}$ enters the formal excess-risk bound through the approximation error (a direct, provable connection). The role of $\hat{\mathcal{A}}_{\text{struct}}$ is more nuanced: the formal estimation error bound depends on $C-1$ (not on $\hat{\mathcal{A}}_{\text{struct}}$), but $\hat{\mathcal{A}}_{\text{struct}}$ captures an important practical consideration: how efficiently the representation allocates its capacity to task-relevant directions (Proposition 5.11). A representation with high $\mathcal{A}_{\text{disc}}$ but low $\hat{\mathcal{A}}_{\text{struct}}$ has good class separation but wastes most of its variance on noise, leading to unstable estimates in practice even though the formal bound holds. The geometric mean $\mathcal{C}_{\text{RTA}} = \sqrt{\mathcal{A}_{\text{disc}} \cdot \hat{\mathcal{A}}_{\text{struct}}}$ penalizes both failure modes: poor separation (low $\mathcal{A}_{\text{disc}}$) and poor efficiency (low $\hat{\mathcal{A}}_{\text{struct}}$).

Remark 5.14 (Why optimize RTA rather than $\mathcal{A}_{\text{disc}}$ alone?). A natural question arises: if the formal risk theorem involves only $\mathcal{A}_{\text{disc}}$, why should a practitioner optimize $\mathcal{C}_{\text{RTA}}$ rather than $\mathcal{A}_{\text{disc}}$ alone? We offer three reasons, each operating at a different level of the framework:

(i) Axiomatic completeness. The four axioms (Section 4.1) characterize desirable properties of *any* principled transferability metric. $\mathcal{A}_{\text{disc}}$ alone satisfies Axioms 1–4, but does not capture the full geometry of the representation-task relationship. A metric that ignores structural alignment may rank two representations identically when one concentrates its variance on the discriminative subspace (efficient) and the other spreads it across d irrelevant directions (wasteful). $\hat{\mathcal{A}}_{\text{struct}}$ breaks this tie.

(ii) Practical stability. The formal estimation error bound $\mathcal{O}(\sqrt{C-1}/\sqrt{n})$ assumes the oracle discriminative-subspace projection. In finite samples, this projection must itself be estimated (the class means used to construct P_B are sample quantities). When $\hat{\mathcal{A}}_{\text{struct}}$ is low (indicating that the task-relevant signal is a tiny

Algorithm 1 Autonomous Pre-training Knowledge Router (APKR)

Input: Repository $\mathcal{M} = \{M_1, \ldots, M_K\}$; probe set $\{(x_i, y_i)\}_{i=1}^n$; regularization $\lambda > 0$.
Output: Selected source model index k^*.

for $k = 1, \ldots, K$ **do**
1. Extract and center features: $Z_k \leftarrow \text{center}([\phi_k(x_1), \ldots, \phi_k(x_n)]^\top)$
2. Compute $\Sigma_k^{(B)}, \Sigma_k^{(W)}$ per Definition 4.1
3. $\mathcal{A}_{\text{disc}}^{(k)} \leftarrow \text{tr}\big((\Sigma_k^{(W)} + \lambda\Sigma_k)^{-1}\Sigma_k^{(B)}\big) \,/\, \text{rank}(\Sigma_k^{(B)})$
4. $\hat{\mathcal{A}}_{\text{struct}}^{(k)} \leftarrow \|K_t Z_k\|_F^2 \,/\, (\big\|Z_k^\top Z_k\big\|_F \cdot \big\|K_t^2\big\|_F)$
5. $\mathcal{C}_{\text{RTA}}^{(k)} \leftarrow \sqrt{\mathcal{A}_{\text{disc}}^{(k)} \cdot \hat{\mathcal{A}}_{\text{struct}}^{(k)}}$

end for

return $k^* \leftarrow \arg\max_k \mathcal{C}_{\text{RTA}}^{(k)}$

fraction of the total representation variance), the estimated projection is unstable, and the realized estimation error exceeds the formal bound. Optimizing $\mathcal{C}_{\text{RTA}}$ selects representations where the projection is reliable.

(iii) The impossibility theorem. Theorem 9.2 shows that no single-component metric can be universally optimal across all loss functions. The geometric mean $\mathcal{C}_{\text{RTA}}$ represents a principled compromise that hedges against settings where structural properties matter more than the discriminative ratio (e.g., squared loss with regression-style formulations, as shown in the proof of Theorem 9.2).

In summary: the risk theorem justifies $\mathcal{A}_{\text{disc}}$, the axioms justify $\hat{\mathcal{A}}_{\text{struct}}$, and RTA combines both.

Corollary 5.15 (Sample Complexity)**.** *To achieve total excess risk at most ϵ above the Bayes risk, the required number of target samples is:*

$$n \;\geq\; \frac{c_1^2(C-1)}{(\epsilon - f(\mathcal{A}_{\text{disc}}(\phi, \mathcal{T}_t)))^2}, \tag{17}$$

provided $\epsilon > f(\mathcal{A}_{\text{disc}})$ (i.e., the desired accuracy is achievable given the approximation error). Higher $\mathcal{A}_{\text{disc}}$ increases the "accuracy budget" $\epsilon - f(\mathcal{A}_{\text{disc}})$ available for estimation, thereby reducing the sample requirement.

6 Computational Overhead

Theorem 6.1 (Selection Complexity)**.** *For K models with maximum dimension $d_{\max}$ and n probe samples with C classes, the APKR selection costs:*

$$T_{\text{select}} = \mathcal{O}\big(K \cdot (n \cdot d_{\max} + d_{\max}^2 C + n^2)\big). \tag{18}$$

Proof. Per source model: feature extraction costs $\mathcal{O}(nd)$; covariance estimation costs $\mathcal{O}(nd^2)$; the discriminant trace requires solving $(\Sigma^{(W)} + \lambda\Sigma_\phi)^{-1}\Sigma^{(B)}$, which costs $\mathcal{O}(d^2C)$ since $\Sigma^{(B)}$ has rank at most $C-1$; CKA requires $\mathcal{O}(n^2 d)$ for the Gram-matrix products. Summing over K models gives the stated bound. □

Proposition 6.2 (Negligibility)**.** *The ratio $T_{\text{select}}/T_{\text{train}} = \mathcal{O}(Kd_{\max}^2/(Tp)) \ll 1$ whenever training involves T gradient steps on a model with p parameters and $Tp \gg Kd_{\max}^2$.*

7 The APKR Algorithm

8 Unification of Existing Metrics

We now show that three prominent transferability metrics are special cases or relaxations of RTA. This explains their empirical success and identifies their theoretical limitations.

Table 1: Comparison of transferability metrics against the APKR axioms and theoretical guarantees. ✓ = satisfies, × = violates, ∼ = partially satisfies.

Metric	Mono.	Inv.	Bound.	Consist.	Risk Bound
LEEP	∼	×	✓	✓	×
H-Score	✓	✓	×	✓	×
LogME	∼	✓	✓	✓	×
NCE	✓	×	✓	✓	×
Task2Vec	∼	∼	×	∼	×
RTA	✓	✓	✓	✓	✓

Theorem 8.1 (H-Score as the Discriminative Component)**.** *The H-Score (Bao et al., 2019), defined as* $\mathcal{H}(\phi, \mathcal{T}_t) = \mathrm{tr}(\Sigma_\phi^{-1}\Sigma_\phi^{(B)})$*, is the unnormalized, unregularized discriminative alignment:* $\mathcal{H} = \mathrm{rank}(\Sigma^{(B)}) \cdot \mathcal{A}_{\mathrm{disc}}\big|_{\lambda=0}$*. H-Score satisfies Axioms 1 and 2 but violates Axiom 3 (it ranges in* $[0, C-1]$*, not* $[0,1]$*) and ignores structural alignment entirely.*

Proof. Since $\Sigma_\phi = \Sigma^{(B)} + \Sigma^{(W)}$, setting $\lambda = 0$ in Definition 4.3 gives:

$$\mathcal{A}_{\mathrm{disc}}\big|_{\lambda=0} = \frac{\mathrm{tr}\big((\Sigma^{(W)})^{-1}\Sigma^{(B)}\big)}{\mathrm{rank}(\Sigma^{(B)})}.$$

The H-Score uses Σ_ϕ^{-1} rather than $(\Sigma^{(W)})^{-1}$ as the denominator matrix. Both are monotone in the same generalized eigenvalues of the pencil $(\Sigma^{(B)}, \Sigma^{(W)})$. Specifically, let γ_i denote the eigenvalues of $(\Sigma^{(W)})^{-1}\Sigma^{(B)}$. The eigenvalues ρ_i of $\Sigma_\phi^{-1}\Sigma^{(B)}$ satisfy $\rho_i = \gamma_i/(1+\gamma_i)$, so $\mathcal{H} = \sum_i \rho_i = \sum_i \gamma_i/(1+\gamma_i) \le C-1$. □

Theorem 8.2 (LEEP as a Discriminative Relaxation)**.** *LEEP (Nguyen et al., 2020) is defined as:*

$$LEEP = \frac{1}{n}\sum_{i=1}^{n} \log\Big(\sum_z \hat{p}(y_i \mid z)\, \hat{p}(z \mid x_i)\Big),$$

and operates on the softmax outputs rather than intermediate features. When viewing $\phi^{soft}(x) = [\hat{p}(z \mid x)]_z$ *as a representation, LEEP is monotonically related to* $\mathcal{A}_{\mathrm{disc}}(\phi^{soft}, \mathcal{T}_t)$*. LEEP violates Axiom 2: rescaling the logits (a linear transformation in logit space) changes the softmax outputs and hence the LEEP score.*

Theorem 8.3 (LogME as a Bayesian Discriminative Proxy)**.** *LogME (You et al., 2021), defined as the maximized log marginal likelihood* $\max_{\alpha,\beta} \log p(Y|\Phi, \alpha, \beta)$*, is a Bayesian refinement of* $\mathcal{A}_{\mathrm{disc}}$*: the prior on* w *plays the role of the regularization* λI*, and the marginal likelihood marginalizes over head parameters rather than optimizing them. LogME satisfies Axioms 3–4 but can violate Axiom 1 in finite samples due to the Bayesian Occam's razor effect: the marginal likelihood can prefer simpler representations that discard task-relevant information.*

9 Extended Theoretical Analysis

In this section, we provide three additional theoretical results that strengthen the framework: an impossibility theorem establishing fundamental limits, finite-sample concentration bounds, and analytic examples demonstrating the tightness of our results.

9.1 An Impossibility Result

A natural question is whether the four axioms can be strengthened. The following result shows a fundamental tension between our axioms and an additional desirable property.

Definition 9.1 (Loss Universality)**.** A compatibility metric $\mathcal{C}$ is *loss-universal* if the ranking it induces over source models is independent of the choice of loss function ℓ: for all ℓ_1, ℓ_2 and all ϕ_a, ϕ_b,

$$\mathcal{C}(\phi_a, \mathcal{T}_t^{(\ell_1)}) \geq \mathcal{C}(\phi_b, \mathcal{T}_t^{(\ell_1)}) \iff \mathcal{C}(\phi_a, \mathcal{T}_t^{(\ell_2)}) \geq \mathcal{C}(\phi_b, \mathcal{T}_t^{(\ell_2)}).$$

Theorem 9.2 (Impossibility of Universal Optimality)**.** *No compatibility metric $\mathcal{C}$ can simultaneously satisfy Axioms 1–4, be loss-universal (Definition 9.1), and guarantee that $k^* = \arg\max_k \mathcal{C}(\phi_k, \mathcal{T}_t)$ minimizes the transfer risk $\mathcal{L}_{k^*}^*$ for* every *Lipschitz loss function ℓ.*

Proof. We construct an explicit counterexample. Consider two representations ϕ_A, ϕ_B of a binary classification task with $\mathcal{A}_{\text{disc}}(\phi_A) > \mathcal{A}_{\text{disc}}(\phi_B)$ but $\hat{\mathcal{A}}_{\text{struct}}(\phi_A) < \hat{\mathcal{A}}_{\text{struct}}(\phi_B)$.

Under the 0-1 loss, the transfer risk is dominated by the classification margin, which depends primarily on the Mahalanobis separation (controlled by $\mathcal{A}_{\text{disc}}$). Hence ϕ_A is preferred under 0-1 loss.

Under the squared loss with a regression-style formulation, the transfer risk depends more heavily on the spectral properties of the feature covariance in the label-relevant subspace (controlled by $\hat{\mathcal{A}}_{\text{struct}}$). For a specific construction where ϕ_B has its variance concentrated on the discriminative subspace while ϕ_A spreads variance across irrelevant directions, ϕ_B achieves lower squared-loss risk despite lower $\mathcal{A}_{\text{disc}}$.

Hence, any fixed metric that ranks $\phi_A > \phi_B$ (resp. $\phi_B > \phi_A$) is sub-optimal for at least one loss function. Since the metric must produce a *single* ranking (loss-universality), it cannot be simultaneously optimal for both. □

Remark 9.3. This result is analogous to Kleinberg's (2003) impossibility theorem for clustering: it shows that no single metric can be "perfect" across all settings. The RTA score's geometric-mean combination of $\mathcal{A}_{\text{disc}}$ and $\hat{\mathcal{A}}_{\text{struct}}$ represents a principled compromise. Different weightings $\mathcal{C}_\alpha = \mathcal{A}_{\text{disc}}^\alpha \cdot \hat{\mathcal{A}}_{\text{struct}}^{1-\alpha}$ may be preferred for specific loss families; we leave this investigation to future work.

9.2 Finite-Sample Concentration

Theorem 9.4 (Concentration of the RTA Estimator)**.** *Under Assumption 5.1, for any $\delta > 0$, with probability at least $1 - \delta$:*

$$\left|\hat{\mathcal{C}}_{\text{RTA},n} - \mathcal{C}_{\text{RTA}}\right| \leq c_2 \tau^2 \sqrt{\frac{d \log(2/\delta)}{n}}, \tag{19}$$

where c_2 is a universal constant depending only on the maximum of $\mathcal{A}_{\text{disc}}$ and $\hat{\mathcal{A}}_{\text{struct}}$.

Proof. We bound the deviation of each component separately.

Discriminative alignment. The empirical covariance matrices $\hat{\Sigma}^{(W)}$ and $\hat{\Sigma}^{(B)}$ are computed from sample class-conditional means and covariances. By the matrix concentration inequality for sub-Gaussian distributions (Vershynin, 2018, Theorem 4.7.1):

$$\left\|\hat{\Sigma}_\phi - \Sigma_\phi\right\|_{\text{op}} \leq C_0 \tau^2 \left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)$$

with probability at least $1 - 2e^{-c_0 d}$. Since the class-conditional covariances are computed from subsets of size $n_c = n\pi_c$, the same bound applies with n replaced by $\min_c n_c$. The discriminative alignment $\mathcal{A}_{\text{disc}}$ is a ratio of traces involving $\Sigma^{(B)}$ and $(\Sigma^{(W)} + \lambda\Sigma_\phi)^{-1}$. The trace is a Lipschitz function of the matrix entries (with Lipschitz constant bounded by $\sqrt{d}$ in the operator norm), and the matrix inverse $A \mapsto A^{-1}$ is locally Lipschitz with constant $\left\|A^{-1}\right\|_{\text{op}}^2$ when A is bounded away from singularity (which the regularization $\lambda\Sigma_\phi$ ensures). Composing these Lipschitz bounds:

$$\left|\hat{\mathcal{A}_{\text{disc}}} - \mathcal{A}_{\text{disc}}\right| \leq \frac{C_1 \tau^2}{\lambda^2} \sqrt{\frac{d}{n}}$$

with high probability, where C_1 depends on the spectral bounds of Σ_ϕ.

Structural alignment. The empirical CKA is a ratio of degree-4 polynomials in the feature vectors. Each Frobenius norm in the definition is a sum of squared inner products, which concentrates at rate $1/\sqrt{n}$ by the Hanson-Wright inequality (Rudelson & Vershynin, 2013) applied to each quadratic form. The ratio inherits the $1/\sqrt{n}$ rate by the delta method, giving:

$$\left| \hat{\mathcal{A}}_{\text{struct}} - \hat{\mathcal{A}}_{\text{struct}} \right| \leq C_2 \tau^4 \sqrt{\frac{d}{n}}.$$

Combination. The geometric mean $f(a,b) = \sqrt{ab}$ has partial derivatives $f_a = \sqrt{b/(4a)}$ and $f_b = \sqrt{a/(4b)}$, both bounded when a, b are bounded away from zero. By the multivariate delta method:

$$\left| \hat{\mathcal{C}}_{\text{RTA}} - \mathcal{C}_{\text{RTA}} \right| \leq \sqrt{\frac{b}{4a}} \left| \hat{a} - a \right| + \sqrt{\frac{a}{4b}} \left| \hat{b} - b \right|,$$

where $a = \mathcal{A}_{\text{disc}}$, $b = \hat{\mathcal{A}}_{\text{struct}}$. Substituting the bounds above and applying the union bound gives (19). □

Corollary 9.5 (Selection Consistency)**.** *Let ϕ_{k^*} and ϕ_{k_2} be the best and second-best source models with RTA gap $\Delta_{\mathcal{C}_{\text{RTA}}} = \mathcal{C}_{\text{RTA}}(\phi_{k^*}) - \mathcal{C}_{\text{RTA}}(\phi_{k_2}) > 0$. Then the APKR selects the correct model with probability at least $1 - \delta$ whenever:*

$$n \;\geq\; \frac{4c_2^2 \tau^4 d \log(2K/\delta)}{\Delta^2_{\mathcal{C}_{\text{RTA}}}}. \tag{20}$$

Proof. By a union bound over all K models, each empirical RTA is within $\Delta_{\mathcal{C}_{\text{RTA}}}/2$ of its population value with probability at least $1 - \delta$ when n satisfies the stated bound. This suffices to preserve the ranking. □

9.3 Analytic Examples

Example 9.6 (Gaussian Binary Classification)**.** Consider a binary task with $X|Y=0 \sim \mathcal{N}(\mu_0, \sigma^2 I_d)$ and $X|Y=1 \sim \mathcal{N}(\mu_1, \sigma^2 I_d)$, equal priors, and two linear representations:

$\phi_A(x) = u_A^\top x$ where $u_A = (\mu_1 - \mu_0)/\left\|\mu_1 - \mu_0\right\|$ (the Fisher direction).

$\phi_B(x) = u_B^\top x$ where $u_B \perp (\mu_1 - \mu_0)$ (a random orthogonal direction).

Then $\Sigma_A^{(B)} = \left\|\mu_1 - \mu_0\right\|^2/4$ and $\Sigma_A^{(W)} = \sigma^2$, giving $\mathcal{A}_{\text{disc}}(\phi_A) = \left\|\mu_1 - \mu_0\right\|^2/(4(\sigma^2 + \lambda))$. Meanwhile, $\Sigma_B^{(B)} = 0$, so $\mathcal{A}_{\text{disc}}(\phi_B) = 0$ and $\mathcal{C}_{\text{RTA}}(\phi_B) = 0$.

By Theorem 5.5, the approximation error for ϕ_A is:

$$\Delta_{\phi_A} \leq \log 2 - \frac{1}{2} \log\left(1 + \frac{\left\|\mu_1 - \mu_0\right\|^2}{4(\sigma^2 + \lambda)(1 + \lambda)}\right).$$

This is tight: for the Gaussian model with a linear classifier, the Bayes error is exactly $\Phi(-\left\|\mu_1 - \mu_0\right\|/(2\sigma))$ where Φ is the standard normal CDF, and the bound correctly predicts that ϕ_A achieves near-Bayes-optimal performance while ϕ_B is useless.

Example 9.7 (Three-Model Comparison)**.** Consider a 4-class task in $\mathbb{R}^{10}$ with three source models:

ϕ_1: High $\mathcal{A}_{\text{disc}} = 0.8$, low $\hat{\mathcal{A}}_{\text{struct}} = 0.1$ (separates classes well but wastes dimensions on noise).

ϕ_2: Low $\mathcal{A}_{\text{disc}} = 0.1$, high $\hat{\mathcal{A}}_{\text{struct}} = 0.8$ (spectrally concentrated but poor at separating classes).

ϕ_3: Moderate $\mathcal{A}_{\text{disc}} = 0.5$, moderate $\hat{\mathcal{A}}_{\text{struct}} = 0.5$ (balanced).

The RTA scores are: $\mathcal{C}_{\text{RTA}}(\phi_1) = \sqrt{0.8 \times 0.1} = 0.283$, $\mathcal{C}_{\text{RTA}}(\phi_2) = \sqrt{0.1 \times 0.8} = 0.283$, and $\mathcal{C}_{\text{RTA}}(\phi_3) = \sqrt{0.5 \times 0.5} = 0.5$. Note that ϕ_1 and ϕ_2 receive *identical* RTA scores despite having opposite strengths: one excels in discrimination, the other in structural efficiency. The balanced ϕ_3 scores highest.

Now consider how each single-component metric would rank these models:

H-Score (which measures only $\mathcal{A}_{\text{disc}}$) ranks $\phi_1 \succ \phi_3 \succ \phi_2$, selecting ϕ_1. This ignores the fact that ϕ_1 wastes 90% of its variance on task-irrelevant directions.

CKA alone (which measures only $\hat{\mathcal{A}}_{\text{struct}}$) ranks $\phi_2 \succ \phi_3 \succ \phi_1$, selecting ϕ_2. This ignores the fact that ϕ_2 barely separates the classes.

RTA ranks $\phi_3 \succ \phi_1 = \phi_2$, selecting ϕ_3. The geometric mean correctly identifies the balanced model as the best overall choice, because it penalizes both extremes equally: wasting variance (ϕ_1) is just as costly as failing to separate classes (ϕ_2).

10 Discussion

The Axiomatic Contribution. We view the axioms (Section 4.1) as the most broadly applicable contribution of this paper. While our specific metric (RTA) and risk bounds are restricted to the linear-probing setting, the axioms are setting-independent: they prescribe what any transferability metric should satisfy regardless of the downstream protocol. We hope this perspective will guide the design and evaluation of future metrics, much as Shannon's axioms guided information measures and Arrow's axioms guided social choice theory.

Scope and Limitations. Our transfer risk bound (Theorem 5.12) applies exclusively to the linear probing setting with frozen features, a deliberate methodological choice enabling rigorous, self-contained proofs. This is the standard protocol for representation evaluation (Alain & Bengio, 2017; Kornblith et al., 2019b), but practical deployment often involves full fine-tuning. Extending the analysis to non-convex fine-tuning via neural tangent kernel (NTK) theory (Jacot et al., 2018) is an important open problem. The sub-Gaussian class-conditional assumption (Assumption 5.2) is satisfied by all bounded-output neural networks but excludes heavy-tailed feature distributions. All formal results are proven for multi-class classification; extending to regression and structured prediction requires different kernel constructions for the structural alignment component.

The Formal Role of Structural Alignment. We acknowledge a deliberate asymmetry in the theoretical treatment of the two RTA components. The discriminative alignment $\mathcal{A}_{\text{disc}}$ enters the excess-risk bound (Theorem 5.12) directly and provably: higher $\mathcal{A}_{\text{disc}}$ implies a smaller approximation error, full stop. The structural alignment $\hat{\mathcal{A}}_{\text{struct}}$ plays a different, complementary role. The formal estimation error bound (Proposition 5.10) depends on $C-1$ rather than on $\hat{\mathcal{A}}_{\text{struct}}$, because we achieve the dimension reduction by projecting onto the discriminative subspace analytically. However, $\hat{\mathcal{A}}_{\text{struct}}$ captures an important practical property (Proposition 5.11): how efficiently the representation allocates its variance budget to task-relevant directions. A representation with low $\hat{\mathcal{A}}_{\text{struct}}$ wastes capacity on noise, leading to practical instability even when the formal bound holds. We view this honest separation (one component with a direct formal bound, one with a well-characterized practical role) as preferable to overclaiming a formal connection that cannot be rigorously defended.

The Role of the Geometric Mean. The impossibility result (Theorem 9.2) shows that no single combination of $\mathcal{A}_{\text{disc}}$ and $\hat{\mathcal{A}}_{\text{struct}}$ is universally optimal across all loss functions. The geometric mean $\mathcal{C}_{\text{RTA}} = \sqrt{\mathcal{A}_{\text{disc}} \cdot \hat{\mathcal{A}}_{\text{struct}}}$ is one principled choice; other combinations $\mathcal{A}_{\text{disc}}^{\alpha} \cdot \hat{\mathcal{A}}_{\text{struct}}^{1-\alpha}$ may be preferred for specific settings. Investigating the optimal α for particular loss families is a natural extension.

Multi-Source and Dynamic Extensions. APKR selects a single source model. Extending to convex combinations $\phi_{\text{ens}} = \sum_k \alpha_k \phi_k$ with provable guarantees, or to dynamic routing where the source can be switched during training, are natural generalizations that introduce optimization-over-compositions challenges beyond the scope of this paper.

Empirical Validation. This paper is intentionally positioned as a purely theoretical contribution. Empirical validation of the RTA score against existing metrics on large-scale model repositories is an important

direction for future work. The theoretical results predict that RTA should achieve strong rank correlation with actual transfer performance, particularly on tasks where both discriminative and structural alignment matter (e.g., fine-grained recognition).

11 Conclusion

We have presented an axiomatic theory of transferability metrics for source-model selection in transfer learning. Our contributions are: (1) four axioms formalizing the requirements for principled compatibility metrics; (2) the RTA score, which provably satisfies all four axioms; (3) a transfer risk bound decomposing excess risk into approximation error (controlled by discriminative alignment) and estimation error (controlled by structural alignment); (4) a unification showing LEEP, LogME, and H-Score as incomplete instances of the framework; and (5) an impossibility result establishing fundamental limits on universal optimality.

We believe the axiomatic perspective, asking *"What should a metric satisfy?"* before *"Which metric is best?"*, offers lasting value beyond the specific results of this paper. The axioms and the risk decomposition provide a principled language for evaluating future transferability metrics and a formal target for future theoretical work on automated model selection.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *ICCV*, 2019.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *ICIP*, 2019.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *NeurIPS*, 2007.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021.

Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition, 1990.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 129:1789–1819, 2021.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 2005.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

Arthur Jacot, Franck Gabriel, and Hongler Clément. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

Jon Kleinberg. An impossibility theorem for clustering. In *NeurIPS*, 2003.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019a.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better ImageNet models transfer better? In *CVPR*, 2019b.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *ICLR*, 2019.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cédric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *ICML*, 2020.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2009.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

Robert J Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, 2009.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

Yang Tan, Yang Li, and Shao-Lun Huang. OTCE: A transferability metric for cross-domain cross-task representations. In *CVPR*, 2021.

Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *ICCV*, 2019.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *ICML*, 2021.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. In *ICLR*, 2017.