ATTRIBUTION AMBIGUITY CAN MAKE TARGET BAYES RULES NON-IDENTIFIABLE IN EEG TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable unsupervised domain adaptation (UDA) for electroencephalography (EEG) decoding depends on information constraining how source and target domains may differ. We ask how far such knowledge can take us: if the laws governing how latent neural components transform across domains were fully specified, would that be enough to determine the optimal target classifier? We examine this question in an idealized Oracle setting in which the full source distribution of observations and labels, the target observation marginal, and the corresponding component-specific operators are known exactly, while other common sources of uncertainty are removed. Yet a simple Gaussian construction yields a negative answer. We trace this failure to *attribution ambiguity*: scalp EEG observations reflect aggregate neural activity but not how predictive activity is partitioned across component types. Although each type's cross-domain operator is known, allocations indistinguishable in the source domain can remain consistent with different target observation–label relationships that require conflicting optimal decisions. To characterize the resulting learning barrier, we derive an information-theoretic lower bound showing that it cannot be overcome by algorithmic choice alone. With additional structure, a complementary upper bound shows that the target decision can be learned without full latent-attribution recovery. Finally, EEG simulations instantiate the theoretical mechanism, while real-EEG experiments provide complementary evidence for its learner-side predictive implications.

1 INTRODUCTION

Unsupervised domain adaptation (UDA) asks which predictive relationships learned from labeled source data remain valid in a target domain where labels are unavailable and the data-generating process may have changed. Without information constraining how that process may change, adaptation has no principled basis for determining which source-predictive relationships should survive the shift (Ben-David et al., 2010; Gulrajani & Hashimoto, 2022). Reliable adaptation therefore depends on knowledge of admissible cross-domain changes. This raises a sharper question: how far can such knowledge take us? At the information limit, we ask whether exact knowledge of the cross-domain transformation structure suffices to determine the target-optimal prediction rule. The answer separates an estimation problem from an information barrier. If that structure were decision-sufficient, accurate estimation of it would in principle suffice. If not, the target decision can remain underdetermined even when the transformations are known exactly—an ambiguity that no algorithm operating on the same evidence can resolve. This distinction matters concretely in EEG decoding, where adaptation without target labels still lags behind approaches using limited target supervision, while target-specific calibration remains difficult to eliminate (Liu et al., 2025; Xiong et al., 2026; Kastrati et al., 2025; Lee et al., 2026).

The standard EEG forward model (Hallez et al., 2007) gives the information-sufficiency question a concrete physical form. Scalp recordings are linear mixtures of latent neural activity, $\boldsymbol{X} = \boldsymbol{G}\boldsymbol{J}$, with cross-domain variation that can arise from changes in the lead-field matrix $\boldsymbol{G}$, reflecting anatomy and recording conditions, and from changes in the latent neural source activity $\boldsymbol{J}$. Crucially, cross-domain changes in the latent components of $\boldsymbol{J}$ are modeled by component-specific transformations, while their contributions are observed only through the aggregate scalp signal. Under this structure, we consider an idealized unsupervised domain adaptation setting. The learner has exact knowledge of the lead-field mappings in both domains and the fixed, label-independent component-specific operators

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.

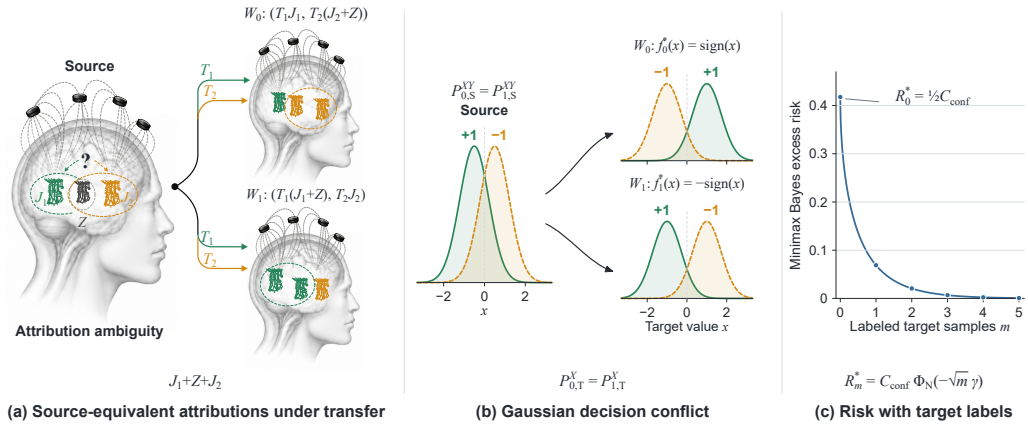

Figure 1: Attribution ambiguity and target decisions. (a) Different latent attributions preserve source observations but can alter target observations. (b) Identical Oracle information permits opposite target Bayes rules. (c) Labeled target samples reduce minimax excess risk when both candidate models are fully known. Panels (b,c) use $\Delta = 1$, $\sigma = 0.72$. Here, $\gamma = \Delta/\sigma$, C_{conf} is the pairwise Bayes conflict, and Φ_N is the standard normal CDF. See Sections 3.2 and 4.

that map source-domain component activity to target observations. All other information that could bear on the question is also given exactly, thereby ruling out all sources of alternative explanation, including estimation error, shifts in class priors or label semantics, and any remaining observational uncertainty. We call this collection the *Oracle information* (formalized in Section 3). With these explanations excluded, a single question remains: *does the Oracle information uniquely determine the target Bayes rule?* We show that it need not.

We construct two generative models compatible with the same Oracle information. Even a binary ambiguity between them suffices to produce opposite target Bayes rules. We term this obstruction *attribution ambiguity*: scalp observations reflect aggregate neural activity but do not reveal how predictive activity is partitioned across latent components. For prediction, any learner effectively induces a particular attribution of this predictive activity across components, yet the available information need not identify that attribution as the one relevant for target prediction. When latent components transform differently across domains, source-indistinguishable attributions can lead to different target-optimal decisions. Although existing UDA theory shows that target rules can remain non-identifiable without restrictive assumptions (Ben-David et al., 2010; Zhao et al., 2019; Gulrajani & Hashimoto, 2022; Dong et al., 2026), those results leave degrees of freedom that our Oracle setting eliminates (Section 2). We show that non-identifiability can persist even at this information extreme. An information-theoretic lower bound characterizes when attribution ambiguity gives rise to unavoidable excess risk for any algorithm (Section 4). Yet this barrier is not absolute. Under additional structure, the target rule is learnable without recovering the full latent attribution (Section 5). Finally, we instantiate this mechanism in EEG simulations under forward-model constraints and evaluate component-specific learner responses under matched target interventions on real EEG (Section 6).

2 Related Work

Whether available source and target information uniquely identifies the target classification rule is a basic question in UDA (Ben-David et al., 2010; Ben-David & Urner, 2012). Prior work locates cross-domain non-identifiability in label relationships left unconstrained by feature alignment (Zhao et al., 2019; Johansson et al., 2019; Mehra et al., 2021), target distributions not fully identified despite priors over changing modules (Zhang et al., 2020), and target label probabilities left undetermined by graphical knowledge of causal shifts (Jalaldoust & Bareinboim, 2024). Most closely related, observed distributions may not identify a label-preserving cross-domain map even when the candidate family contains one (Gulrajani & Hashimoto, 2022). Target-label uncertainty can also remain after observing labeled source and unlabeled target samples (Dong et al., 2026). The former leaves the cross-domain map unknown, whereas the latter does not fix shared component-specific transformations across

compatible instances. Whether non-identifiability survives after these decision-relevant degrees of freedom are eliminated, pushing the problem to an information extreme, remains open (see Appendix D for a detailed comparison).

UDA is a central setting in EEG decoding. One line of work studies the theoretical and statistical sources of transfer difficulty by quantifying marginal and conditional shifts or tracing performance degradation to inter-subject variability and shortcut learning (Albuquerque et al., 2022; Wang et al., 2024b). Another develops EEG-specific transfer methods that use covariance geometry or generative models to compensate for changes in neural activity and measurement (Zanini et al., 2018; Kobler et al., 2022; Mellot et al., 2024; Li et al., 2025). More recently, EEG transfer models have expanded from conventional approaches (Wu et al., 2022) to foundation models pretrained on large datasets (Kostas et al., 2021; Jiang et al., 2024; Wang et al., 2024a; 2025). Despite these advances, large-scale pretraining enables strong zero-shot transfer in vision and speech, whereas EEG decoding remains markedly dependent on target-specific calibration. The information requirements underlying this persistent gap remain insufficiently understood.

Classical blind source separation and independent component analysis concern the recovery of latent sources or mixing mechanisms from observed mixtures (Comon, 1994; Hyvärinen & Pajunen, 1999; Hyvärinen & Oja, 2000). In this paper, we instead study attribution ambiguity, which concerns whether source-predictive variation is uniquely attributable across latent components.

3 Attribution Ambiguity and Target Bayes-Rule Non-Identifiability

Problem setup. We consider transfer of the same EEG classification task from a source domain $\mathcal{S}$ to a target domain $\mathcal{T}$. Let $\boldsymbol{X}^{(e)} \in \mathcal{X}^{(e)}$ and $Y \in \mathcal{Y}$ denote the EEG observation and task label in domain $e \in \{\mathcal{S}, \mathcal{T}\}$, where $\mathcal{Y}$ is finite. The learner knows a class $\mathfrak{W}$ of candidate generative models. Each $W \in \mathfrak{W}$ specifies the joint distribution of latent variables and labels and the mechanisms generating observations in the two domains. For a measurable classifier $f : \mathcal{X}^{(\mathcal{T})} \to \mathcal{Y}$, let $R_W^{(\mathcal{T})}(f) := P_{W,\mathcal{T}}^{XY}[f(\boldsymbol{X}^{(\mathcal{T})}) \neq Y]$. Any classifier minimizing this risk over all measurable classifiers is called a target Bayes classification rule under W and is denoted by f_W^*.

Oracle information. Within $\mathfrak{W}$, the source- and target-domain lead-field mappings are fixed and known. Under model W, the learner also knows the full source observation–label law $P_{W,\mathcal{S}}^{XY}$, the target observation marginal $P_{W,\mathcal{T}}^{X}$, a shared class prior $P_{W,\mathcal{S}}^{Y} = P_{W,\mathcal{T}}^{Y} =: P_W^Y$ with fixed label semantics, and the fixed, label-independent component maps, collectively denoted by L and defined below. We call $\Phi_{\mathrm{orc}}(W) := (P_{W,\mathcal{S}}^{XY}, P_{W,\mathcal{T}}^{X}, P_W^Y, L)$ the Oracle information under W. By augmenting observed distributions with exact cross-domain structure that is not normally accessible, it provides a deliberately strong information-sufficiency test of whether $\Phi_{\mathrm{orc}}(W)$ suffices to determine f_W^*.

3.1 Source-Equivalent Attributions under Component-Specific Transformations

Source-equivalent attribution. Under the linear EEG forward model considered here, total source activity is mapped to scalp observations through the lead-field matrix. Let $\mathcal{J}^{(\mathcal{S})}$ denote the space of neural source activities in the source domain and $\boldsymbol{G}^{(\mathcal{S})}$ the source-domain lead-field matrix. Write the total source activity as $\boldsymbol{J}_1 + \boldsymbol{J}_2$, where $\boldsymbol{J}_1$ represents the neural processes of interest to the task and $\boldsymbol{J}_2$ represents the remaining neural processes included in the model. Both may carry label-predictive information. The source-domain observation is $\boldsymbol{X}^{(\mathcal{S})} = \boldsymbol{G}^{(\mathcal{S})}(\boldsymbol{J}_1 + \boldsymbol{J}_2)$. Let $(\boldsymbol{J}_1, \boldsymbol{J}_2, \boldsymbol{Z}, Y)$ be jointly defined on a common probability space. For the same total source activity, consider two candidate models W_0, W_Z with latent attributions

$$(\boldsymbol{J}_{1,W_0}, \boldsymbol{J}_{2,W_0}) = (\boldsymbol{J}_1, \boldsymbol{J}_2), \qquad (\boldsymbol{J}_{1,W_Z}, \boldsymbol{J}_{2,W_Z}) = (\boldsymbol{J}_1 + \boldsymbol{Z}, \boldsymbol{J}_2 - \boldsymbol{Z}).$$

Here, $\boldsymbol{Z} \in \mathcal{J}^{(\mathcal{S})}$ denotes activity attributed to different component types under the two candidate models. The two decompositions have identical total source activity and therefore satisfy $\boldsymbol{X}_{W_0}^{(\mathcal{S})} = \boldsymbol{X}_{W_Z}^{(\mathcal{S})}$ sample-wise; when paired with the same label Y, they also satisfy $P_{W_0,\mathcal{S}}^{XY} = P_{W_Z,\mathcal{S}}^{XY}$. The equal source joint laws make the models source-equivalent. If the joint distributions of $(\boldsymbol{J}_1, \boldsymbol{J}_2, Y)$

and $(\boldsymbol{J}_1 + \boldsymbol{Z}, \boldsymbol{J}_2 - \boldsymbol{Z}, Y)$ differ, we refer to this source-domain equivalence with different latent-component attributions as attribution ambiguity. The term $\boldsymbol{Z}$ need not be a deterministic function of Y or correspond to a single anatomical source or functional network. It may aggregate distributed task-related activity and carry label-predictive information.

Component-specific transformations. Whether source-equivalent attribution differences affect target observations depends on how the component activities are transformed across domains. Let $\mathcal{J}^{(\mathcal{T})}$ denote the space of neural source activities in the target domain and $\boldsymbol{G}^{(\mathcal{T})}$ the target-domain lead-field matrix. The lead-field matrices $\boldsymbol{G}^{(\mathcal{S})}$ and $\boldsymbol{G}^{(\mathcal{T})}$ are fixed across candidate models. The following assumption specifies the transformations.

Assumption 1 (Component-specific cross-domain transformations). *There exist linear operators $T_1, T_2 : \mathcal{J}^{(\mathcal{S})} \to \mathcal{J}^{(\mathcal{T})}$ that are fixed across candidate models and do not depend on the label value, with T_1 and T_2 specifying how the two types of neural source activity are transformed into the target domain, respectively. The meanings of the two component types and their correspondence to these operators are also held fixed across candidate models.*

Under Assumption 1, the target observation under model W is

$$\boldsymbol{X}_W^{(\mathcal{T})} = \boldsymbol{G}^{(\mathcal{T})}\big(T_1(\boldsymbol{J}_{1,W}) + T_2(\boldsymbol{J}_{2,W})\big) = L_1(\boldsymbol{J}_{1,W}) + L_2(\boldsymbol{J}_{2,W}),$$

Here, $L_k(\boldsymbol{J}) := \boldsymbol{G}^{(\mathcal{T})} T_k(\boldsymbol{J})$ defines the source-to-target observation map, and we write $L := (L_1, L_2)$. Let $\mathfrak{W}_{\mathrm{add}} \subseteq \mathfrak{W}$ denote the set of candidate models satisfying this additive generative structure and Assumption 1.

Target-domain manifestation. For the two source-equivalent models above, W_Z has $\boldsymbol{J}_1 + \boldsymbol{Z}$ and $\boldsymbol{J}_2 - \boldsymbol{Z}$ relative to W_0. Define the operator difference $D := L_1 - L_2$, so that $D(\boldsymbol{J}) = \boldsymbol{G}^{(\mathcal{T})}[T_1(\boldsymbol{J}) - T_2(\boldsymbol{J})]$. Then

$$\boldsymbol{X}_{W_Z}^{(\mathcal{S})} = \boldsymbol{X}_{W_0}^{(\mathcal{S})}, \qquad \boldsymbol{X}_{W_Z}^{(\mathcal{T})} - \boldsymbol{X}_{W_0}^{(\mathcal{T})} = D(\boldsymbol{Z}).$$

Attribution differences therefore do not necessarily affect target observations. Under the common coupling induced by the reattribution construction, the two target observations coincide almost surely when $D(\boldsymbol{Z}) = 0$ almost surely, whereas $D(\boldsymbol{Z}) \neq 0$ with positive probability implies that they differ with positive probability. This statement concerns the coupled target realizations; their target marginal distributions may still coincide.

Learner-side attribution. For each subject, the recorded scalp EEG is generated by a single realized biophysical process. The alternative latent descriptions above represent different attributions that are indistinguishable from the learner's evidence: the same source-domain observation remains compatible with multiple partitions of predictive activity across latent components. A learner nevertheless effectively induces a particular attribution of predictive activity across components through its predictor. Yet the available evidence need not identify that attribution as the one relevant for target prediction. When latent components undergo divergent cross-domain transformations, this unresolved attribution alters the target Bayes rule, inducing unavoidable excess risk.

3.2 Target Bayes-Rule Non-Identifiability under the Oracle Setting

The additive generative structure and component-specific transformation condition above still do not make the target Bayes classification rule identifiable from the Oracle information.

Proposition 1 (Oracle non-identifiability of target Bayes rules). *There exists a fixed binary additive generative class satisfying the conditions in Section 3.1 and two candidate models W_0, W_Z within this class such that $\Phi_{\mathrm{orc}}(W_0) = \Phi_{\mathrm{orc}}(W_Z)$, while their target Bayes classification rules are each unique almost everywhere and are opposite almost everywhere under the shared target observation marginal distribution.*

Gaussian construction. Let all neural source-activity and observation spaces be $\mathbb{R}$, and set $\boldsymbol{G}^{(\mathcal{S})} = \boldsymbol{G}^{(\mathcal{T})} = 1$. Let $Y \in \{-1, +1\}$ take either value with equal probability, and let $\xi \sim \mathcal{N}(0, \sigma^2)$ be independent of Y, where $\Delta, \sigma > 0$. Fix $T_1 = 2I, T_2 = I$, so that $L_1 = 2I, L_2 = I$. On the same probability space, define

$$(\boldsymbol{J}_{1,W_0}, \boldsymbol{J}_{2,W_0}) = \left(\frac{3\Delta}{2} Y, -2\Delta Y + \xi\right), \qquad (\boldsymbol{J}_{1,W_Z}, \boldsymbol{J}_{2,W_Z}) = \left(-\frac{\Delta}{2} Y, \xi\right).$$

The two decompositions differ by the reallocated activity $\boldsymbol{Z} = -2\Delta Y$, yet produce the same source-domain observation $\boldsymbol{X}^{(\mathcal{S})} = -\frac{\Delta}{2}Y + \xi$, and therefore induce the same source observation–label joint distribution. Applying the fixed component-specific operators L_1, L_2 gives the target observations $\boldsymbol{X}^{(\mathcal{T})}_{W_0} = \Delta Y + \xi$ and $\boldsymbol{X}^{(\mathcal{T})}_{W_Z} = -\Delta Y + \xi$, respectively. By the symmetry of Y, the two models have the same target observation marginal, $\frac{1}{2}\mathcal{N}(-\Delta, \sigma^2) + \frac{1}{2}\mathcal{N}(\Delta, \sigma^2)$, and share the same class priors, label semantics, and $L = (L_1, L_2)$. Hence, $\Phi_{\mathrm{orc}}(W_0) = \Phi_{\mathrm{orc}}(W_Z)$. However, the conditional relationship between the target observation and the common label Y is reversed across the two models. For $\boldsymbol{x} \neq 0$, their Bayes classification rules are

$$f^*_{W_0}(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{x}), \qquad f^*_{W_Z}(\boldsymbol{x}) = -\operatorname{sign}(\boldsymbol{x}).$$

Proposition 1 thus exposes a structural limit beyond statistical uncertainty. Even under Oracle knowledge, many-to-one aggregation leaves attribution unresolved, while component-specific transformations render the Bayes rule non-identifiable. A single binary ambiguity already suffices for decision reversal. This reversal reflects conflicting target observation–label relationships rather than label semantics (Appendix A.2), motivating the general characterization of decision conflict.

4 Bayes Conflict and Protocol-Relative Excess Risk

Attribution ambiguity need not be decision-relevant. Different latent attributions need not induce different target risk functions, and different target risk functions need not require different Bayes-optimal decisions. Appendix A.7.1 gives strict separation constructions for both distinctions among models sharing the same Oracle information. We call two models W, W' Oracle-compatible when $\Phi_{\mathrm{orc}}(W) = \Phi_{\mathrm{orc}}(W')$. For such models, the relevant decision-level question is whether a common measurable classifier can attain both Bayes risks. For any model W, let $r^{\mathrm{Bayes}}_W := \inf_f R^{(\mathcal{T})}_W(f)$ denote the target Bayes risk, where the infimum ranges over all measurable classifiers, and let $\ell^{\mathrm{Bayes}}_W(f) := R^{(\mathcal{T})}_W(f) - r^{\mathrm{Bayes}}_W$ denote the Bayes excess risk. Pairwise Bayes conflict quantifies the minimum summed Bayes excess risk when a common classifier must be used for two Oracle-compatible models. For models with a finite label space and a shared target marginal, its positivity admits an exact pointwise characterization.

Definition 1 (Pairwise Bayes conflict). *For two Oracle-compatible models W, W', define $C_{\mathrm{conf}}(W, W') := \inf_f [\ell^{\mathrm{Bayes}}_W(f) + \ell^{\mathrm{Bayes}}_{W'}(f)]$. If $C_{\mathrm{conf}}(W, W') > 0$, then W, W' have conflicting Bayes-optimal decisions.*

Lemma 1 (Bayes conflict under a shared target marginal). *Let W, W' be two Oracle-compatible models with a finite label space, and let $P^X_{\mathcal{T}}$ denote their shared target observation marginal. For each $V \in \{W, W'\}$, let $A^*_V(\boldsymbol{x})$ denote the set of Bayes-optimal classes under model V at input $\boldsymbol{x}$, and define $\mathcal{X}_{\mathrm{conf}}(W, W') := \{\boldsymbol{x} : A^*_W(\boldsymbol{x}) \cap A^*_{W'}(\boldsymbol{x}) = \varnothing\}$. Then $C_{\mathrm{conf}}(W, W') > 0$ if and only if $P^X_{\mathcal{T}}(\mathcal{X}_{\mathrm{conf}}(W, W')) > 0$. Allowing the classifier to randomize over predicted classes at each input $\boldsymbol{x}$ changes neither the minimum total Bayes excess risk nor this positivity condition.*

Bayes conflict captures the decision-relevant manifestation of attribution ambiguity. For an Oracle-compatible model pair, attribution ambiguity induces an irreducible decision conflict exactly when $C_{\mathrm{conf}}(W, W') > 0$ (Appendix A.3 extends this characterization beyond shared target marginals). Crucially, Bayes conflict is not by itself a learner-level barrier. Whether it yields a positive excess-risk lower bound further depends on how well the transfer protocol distinguishes the candidate models.

Protocol-relative excess risk. Standard transfer regimes (e.g., UDA, FSDA, DG) differ in the data available during training. Under the stronger shared Oracle setting, we abstract these differences through a transfer protocol $\mathcal{R}$, which specifies the additional random observations or data available beyond the shared Oracle information. Let $\mathbb{Q}_{\mathcal{R},W}$ denote the joint distribution of all protocol-visible information under model W. A possibly randomized learning rule A outputs a classifier $\widehat{f}_{\mathsf{A}}$. With $\mathrm{TV}(Q, Q') := \sup_B |Q(B) - Q'(B)|$, the following result relates achievable excess risk to decision conflict and the remaining overlap between protocol observation laws.

Theorem 1 (Protocol-relative Bayes excess-risk lower bound). *For any two Oracle-compatible models W, W' and any protocol $\mathcal{R}$,*

$$\inf_{\mathsf{A}} \max_{V \in \{W, W'\}} \mathbb{E}_{V,\mathsf{A}}\left[\ell^{\mathrm{Bayes}}_V(\widehat{f}_{\mathsf{A}})\right] \geq \frac{1}{2} C_{\mathrm{conf}}(W, W') \left[1 - \mathrm{TV}(\mathbb{Q}_{\mathcal{R},W}, \mathbb{Q}_{\mathcal{R},W'})\right].$$

The bound separates two distinct factors. The conflict C_{conf} measures the cost of using a common target decision across the candidate models, while $1 - \mathrm{TV}(\mathbb{Q}_{\mathcal{R},W}, \mathbb{Q}_{\mathcal{R},W'})$ measures the overlap between their protocol observation laws. When both are positive, every learning rule based on that protocol incurs positive worst-case expected Bayes excess risk. With only the shared Oracle information, the observation laws coincide, and the lower bound becomes $C_{\mathrm{conf}}(W, W')/2$ (Appendix A.4).

The Gaussian pair $\{W_0, W_Z\}$ from Section 3.2 gives an exact instance of this mechanism. Let $\gamma := \Delta/\sigma$, and let Φ_{N} denote the standard normal cumulative distribution function. Both models have Bayes risk $\Phi_{\mathrm{N}}(-\gamma)$, while their complementary target posteriors and common target marginal give $C_{\mathrm{conf}}(W_0, W_Z) = 2\Phi_{\mathrm{N}}(\gamma) - 1$. With only their shared Oracle information, the exact minimax Bayes excess risk is $C_{\mathrm{conf}}(W_0, W_Z)/2 = \Phi_{\mathrm{N}}(\gamma) - 1/2$. Intuitively, assigning equal prior mass to the two candidate worlds leaves exactly one bit of model uncertainty. We next quantify how labeled target samples progressively resolve this binary ambiguity and show that the lower bound is exact in this setting.

Proposition 2. *Suppose the learner knows that the candidate set is exactly $\{W_0, W_Z\}$, knows the complete definitions and shared parameters of both models, and receives m independent and identically distributed labeled target samples in addition to the shared Oracle information. Then, for every integer $m \geq 0$, the exact minimax expected Bayes excess risk is*

$$R^\star_m = \inf_{\mathsf{A}} \max_{V \in \{W_0, W_Z\}} \mathbb{E}_{V,\mathsf{A}}\left[\ell_V^{\mathrm{Bayes}}(\widehat{f}_{\mathsf{A}})\right] = [2\Phi_{\mathrm{N}}(\gamma) - 1]\, \Phi_{\mathrm{N}}(-\sqrt{m}\,\gamma).$$

For every finite m, $R^\star_m > 0$, and $R^\star_m \to 0$ as $m \to \infty$ (Fig. 1(c)). Appendix A.5.2 derives the matching lower bound and gives an attaining rule. In this construction, the component-specific operators are already known exactly through the Oracle information. Labeled target samples progressively distinguish Oracle-compatible models that require different target-optimal rules, thereby reducing the remaining decision ambiguity.

5 A Finite-Sample Guarantee under Additional Structure

Sections 3–4 establish that additive models can contain Oracle-compatible candidates with conflicting target Bayes rules, preventing arbitrarily large labeled-source and unlabeled-target samples from guaranteeing uniformly vanishing excess risk. Appendix A.7 shows that target-decision recovery does not require latent-attribution identification. Focusing on a Gaussian event-related potential (ERP) feature subclass, we now ask what additional structure makes the target decision recoverable in UDA.

An ERP is a scalp-voltage response time-locked to a task event (Luck, 2014). In the ERP feature subclass studied here, class-conditional mean differences provide the primary discriminative signal. We consider fixed ERP features formed by projecting each single-trial observation $\boldsymbol{X}^{(e)} \in \mathbb{R}^{d \times T}$, for $e \in \{\mathcal{S}, \mathcal{T}\}$, onto a prespecified unit-norm temporal template $\boldsymbol{h} \in \mathbb{R}^T$, yielding $\boldsymbol{U}^{(e)} = \Psi(\boldsymbol{X}^{(e)}) = \boldsymbol{X}^{(e)}\boldsymbol{h} \in \mathbb{R}^d$. Such features are commonly modeled as class-conditionally Gaussian with shared within-class covariance (Blankertz et al., 2011; Lotte et al., 2018; Sosulski et al., 2021). Specifically, we focus on equal-prior binary models satisfying $\boldsymbol{U}^{(e)} \mid Y = y \sim \mathcal{N}(\boldsymbol{\mu}_e + y\boldsymbol{\delta}_e/2, \boldsymbol{\Sigma}_e)$, parameterized by class-mean contrast $\boldsymbol{\delta}_e$, common within-class covariance $\boldsymbol{\Sigma}_e$, and marginal mean $\boldsymbol{\mu}_e$ (Appendix A.6.1). The same Gaussian ERP feature family is instantiated in Experiment 1A (Section 6.1), with $\boldsymbol{h}_D$ as the fixed temporal template. Target performance is measured by excess risk $\ell_W^{\mathrm{ERP}}(f)$ relative to the optimal classifier based on these features; its relation to the full-EEG Bayes risk is given in Appendix A.8.6. The subclass realizes the same latent-reallocation geometry and contains a nonempty forward-model family (Appendices A.8.1 and A.8.4, respectively).

To isolate the information required for decision recovery, consider the population geometry of the Gaussian Bayes rule. The rule depends on the target center $\boldsymbol{\mu}_\mathcal{T}$, within-class covariance $\boldsymbol{\Sigma}_\mathcal{T}$, and class-mean contrast $\boldsymbol{\delta}_\mathcal{T}$ oriented to the task labels. The target marginal supplies the center $\boldsymbol{\mu}_\mathcal{T}$ and covariance $\boldsymbol{K}_\mathcal{T} = \boldsymbol{\Sigma}_\mathcal{T} + \boldsymbol{\delta}_\mathcal{T}\boldsymbol{\delta}_\mathcal{T}^\top/4$. Extracting the Bayes rule requires resolving two algebraic dependencies: (1) isolating the contrast outer product from the within-class covariance, and (2) selecting the label-consistent sign along $\mathrm{span}(\boldsymbol{\delta}_\mathcal{T})$. Accordingly, our recovery procedure uses a covariance prediction and an oriented reference to resolve these two dependencies.

We supply these two cues through nominal feature-space maps $\boldsymbol{A}_{\mathrm{bg}}$ and $\boldsymbol{A}_{\mathrm{sig}}$ derived from the known component operators under sufficient conditions. The background map predicts $\boldsymbol{\Sigma}_\mathcal{T}$ from

$\boldsymbol{\Sigma}_{\mathcal{S}}$, while the signal map predicts a target-space reference $\boldsymbol{b}_\star$ from $\boldsymbol{\delta}_{\mathcal{S}}$. Under exact predictions, subtracting $\boldsymbol{A}_{\mathrm{bg}}\boldsymbol{\Sigma}_{\mathcal{S}}\boldsymbol{A}_{\mathrm{bg}}^\top$ from $\boldsymbol{K}_{\mathcal{T}}$ recovers the contrast outer product, and $\boldsymbol{b}_\star = \boldsymbol{A}_{\mathrm{sig}}\boldsymbol{\delta}_{\mathcal{S}}$ fixes its class-consistent sign. Oracle-compatible models satisfying these predictions and positive alignment share the target ERP-feature joint distribution and Bayes rule, while a nonempty forward-model family retains distinct latent allocations (Appendices A.8.3–A.8.4). Decision recovery therefore does not require attribution identification. Appendix A.8.2 gives conditions under which component operators yield exact predictions and explains why the operators alone need not suffice.

Extending population recovery to finite samples requires robustness to imperfect nominal transfer and sampling noise. The signal-reference error is $\varepsilon_{\mathrm{sig}} = \|\boldsymbol{A}_{\mathrm{sig}}\boldsymbol{\delta}_{\mathcal{S}} - \boldsymbol{b}_\star\|_2$, and the relative covariance-prediction error is $\rho_{\mathrm{bg}} = \|\boldsymbol{\Sigma}_{\mathcal{T}}^{-1/2}(\boldsymbol{\Sigma}_{\mathcal{T}} - \boldsymbol{A}_{\mathrm{bg}}\boldsymbol{\Sigma}_{\mathcal{S}}\boldsymbol{A}_{\mathrm{bg}}^\top)\boldsymbol{\Sigma}_{\mathcal{T}}^{-1/2}\|_{\mathrm{op}}$. These measures govern the stability of the recovery, and Theorem 2 bounds their allowable magnitudes. Stable recovery across the candidate class further requires the following uniform condition.

Assumption 2. *Throughout the subclass, the eigenvalues of $\boldsymbol{\Sigma}_{\mathcal{S}}$ and $\boldsymbol{\Sigma}_{\mathcal{T}}$, together with $\|\boldsymbol{\delta}_{\mathcal{T}}\|_2$, lie between fixed positive lower and upper bounds; $\|\boldsymbol{b}_\star\|_2$ is bounded away from zero; and $\|\boldsymbol{A}_{\mathrm{sig}}\|_{\mathrm{op}}$ is bounded above. For some fixed $\kappa_0 > 0$, every model also satisfies $\boldsymbol{b}_\star^\top\boldsymbol{\delta}_{\mathcal{T}}/(\|\boldsymbol{b}_\star\|_2\|\boldsymbol{\delta}_{\mathcal{T}}\|_2) \geq \kappa_0$.*

Assumption 2 stabilizes covariance inversion, spectral recovery, and sign selection. Its alignment condition fixes orientation relative to the task labels and excludes the sign-reversed witness from Sections 3–4. This orientation cannot be verified from unlabeled target data alone. Appendices A.6.1 and A.8.3 give the quantitative conditions and relation to the earlier witness.

The finite-sample learner realizes this recovery pipeline using empirical moments from independent sets of n i.i.d. labeled source samples and N i.i.d. unlabeled target samples. Let $\widehat{\boldsymbol{\delta}}_{\mathcal{S}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}}$ estimate the source contrast and covariance, let $\widehat{\boldsymbol{\mu}}_{\mathcal{T}}, \widehat{\boldsymbol{K}}_{\mathcal{T}}$ estimate the target center and marginal covariance, and let $\boldsymbol{P}_1(\boldsymbol{M})$ denote the leading spectral projector of $\boldsymbol{M}$. The procedure then follows three steps (Appendix A.6.2): (1) nominal transport constructs the covariance estimate $\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}} = \boldsymbol{A}_{\mathrm{bg}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}}\boldsymbol{A}_{\mathrm{bg}}^\top$ and reference $\widehat{\boldsymbol{b}} = \boldsymbol{A}_{\mathrm{sig}}\widehat{\boldsymbol{\delta}}_{\mathcal{S}}$; (2) spectral projection of $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{K}}_{\mathcal{T}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{T}}$ extracts the empirical contrast subspace and orients it as $\widehat{\boldsymbol{v}} = \boldsymbol{P}_1(\widehat{\boldsymbol{M}})\widehat{\boldsymbol{b}}/\|\boldsymbol{P}_1(\widehat{\boldsymbol{M}})\widehat{\boldsymbol{b}}\|_2$; and (3) covariance inversion converts $\widehat{\boldsymbol{v}}$ into the discriminant normal, yielding $\widehat{f}(\boldsymbol{X}) = \mathrm{sign}[\widehat{\boldsymbol{v}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}}^{-1}(\Psi(\boldsymbol{X}) - \widehat{\boldsymbol{\mu}}_{\mathcal{T}})]$. Theorem 2 bounds its finite-sample excess risk under sampling and moment-prediction errors.

Theorem 2 (Finite-sample excess risk under moment-prediction errors). *Under the Gaussian ERP-feature subclass above and Assumption 2, there exist fixed positive thresholds $\varepsilon_\star, \rho_\star$ and constants $C_0, C_1 > 0$, depending only on the uniform bounds and reference-alignment margin in Assumption 2, such that the following holds. For any model W satisfying $\varepsilon_{\mathrm{sig}} \leq \varepsilon_\star$ and $\rho_{\mathrm{bg}} \leq \rho_\star$, and any $\alpha \in (0,1)$, if $n, N \geq C_0[d + \log(16/\alpha)]$, then, with probability at least $1-\alpha$ over the training samples, $\ell_W^{\mathrm{ERP}}(\widehat{f}) \leq C_1[(d + \log(16/\alpha))/n + (d + \log(16/\alpha))/N + \rho_{\mathrm{bg}}^2]$.*

The two transfer errors enter differently. Reference error affects the discrete sign choice, whereas covariance mismatch continuously perturbs the discriminant and contributes ρ_{bg}^2; under exact predictions, the rate is $O((d + \log(1/\alpha))/n + (d + \log(1/\alpha))/N)$, with proofs and constants in Appendices A.6.3–A.6.5. Beyond the rate, the analysis clarifies what information is needed to close the residual ambiguity in this ERP subclass. The maps $\boldsymbol{A}_{\mathrm{bg}}$ and $\boldsymbol{A}_{\mathrm{sig}}$ make these requirements explicit: $\boldsymbol{A}_{\mathrm{bg}}$ predicts $\boldsymbol{\Sigma}_{\mathcal{T}}$ from $\boldsymbol{\Sigma}_{\mathcal{S}}$, while $\boldsymbol{A}_{\mathrm{sig}}\boldsymbol{\delta}_{\mathcal{S}}$ provides the reference $\boldsymbol{b}_\star$, whose alignment with $\boldsymbol{\delta}_{\mathcal{T}}$ fixes the label-consistent orientation. Appendix A.8.2 gives one sufficient structural route by which the known component transformations can induce these predictions, while also showing that the transformations alone need not do so in general. Under this information structure, the target decision is learnable from finite labeled-source and unlabeled-target samples with an explicit excess-risk guarantee.

6 Experiments

6.1 Controlled EEG Simulation of Attribution Ambiguity

Setup. The preceding theory establishes attribution ambiguity in an abstract additive model. Experiment 1 asks whether the same attribution mechanism can arise under linear EEG forward-model constraints and be reflected in non-interchangeable target-trained classifiers. The experiment uses

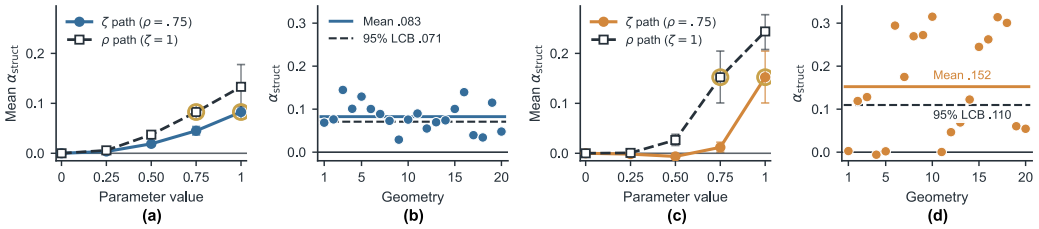

Figure 2: Attribution and classifier exchange loss under EEG forward-model constraints. (a,c) Mean α_{struct} along the reallocation (ζ) and operator-difference (ρ) paths for ERP means (a) and spatial bandpower (c). Error bars show two-sided 95% configuration-bootstrap intervals across 20 fixed source configurations. Gold markers show the primary condition $(\zeta, \rho) = (1, 0.75)$. (b,d) Configuration-wise α_{struct} at the primary condition, with the mean and one-sided 95% lower confidence bound.

paired candidate models that produce identical source observations trial by trial and share the same pair of component-specific target operators. They differ only in the latent allocation of predictive activity across two components. This construction is instantiated for ERP class means (Experiment 1A) and spatial bandpower under zero class-conditional means (Experiment 1B) across 20 fixed spatial source configurations sharing a common lead field. The parameter ζ controls the reallocation of predictive activity, whereas ρ controls the operator difference $D(\rho) = L_1(\rho) - L_2(\rho)$. For each candidate, a shrinkage-LDA reference classifier (Blankertz et al., 2011) is fitted to labeled target training data solely for the exchange comparison. Both classifiers are evaluated on target test data from both candidates. The bidirectional exchange loss α_{struct} is the smaller of the two balanced-error increases produced by exchanging the classifiers between candidates. Thus, $\alpha_{\text{struct}} > 0$ only when both exchanges increase error relative to using each candidate's own reference classifier. The primary condition is $(\zeta, \rho) = (1, 0.75)$. Full construction and analysis details are given in Appendices B.2–B.3.

Results. Under the primary condition, bidirectional exchange losses emerge for both ERP means and spatial bandpower under EEG forward-model constraints. Figures 2(a) and 2(c) show that α_{struct} returns to zero or near zero when either reallocation ζ or the operator difference $D(\rho)$ is removed. The positive bidirectional exchange loss in this construction therefore depends on both factors. Mean α_{struct} is 0.0826 for ERP means and 0.1523 for spatial bandpower, with one-sided 95% lower confidence bounds of 0.0708 and 0.1097, respectively (Figures 2(b) and 2(d)). Both lower bounds exceed zero, establishing positive mean bidirectional exchange loss across the source configurations for both signal patterns (Appendix B.4.1). In further controls, differences in the full target EEG are retained, while the classifier features are identical trial-wise for ERP and identically distributed within each class for spatial bandpower. In these controls, α_{struct} returns to zero or near zero. Thus, differences in the full target EEG alone are insufficient to produce the observed exchange losses (Appendix B.4.4).

6.2 Learner-Side Attribution under Real EEG Transfer

Setup. Section 3 shows that unresolved attribution of predictive activity across EEG components can become target-visible under component-specific transformations. Experiment 2 asks whether this attribution phenomenon has an observable learner-side realization in real EEG. Among predictors with similar source behavior, do larger differences in how predictive activity is organized across components accompany larger target prediction differences? Learner-side attribution is difficult to observe and compare across independently trained networks. The additive model $\boldsymbol{X}^{(\mathcal{S})} = \boldsymbol{G}^{(\mathcal{S})}\boldsymbol{J}_1 + \boldsymbol{G}^{(\mathcal{S})}\boldsymbol{J}_2$ suggests a direct probe. If these component-wise observation contributions were accessible, removing them individually and measuring the resulting prediction changes would reveal how differently learners depend on the components for prediction. For this purpose, we fix four ERP component coordinates from source-domain data using xDAWN (Rivet et al., 2009) and reconstruct their sensor–time contributions $\boldsymbol{X}_k$ through the Haufe forward mapping (Haufe et al., 2014), yielding the exact additive decomposition $\boldsymbol{X} = \sum_{k=1}^{4} \boldsymbol{X}_k + \boldsymbol{R}$ (Appendix C.1.3). This gives all learners the same component coordinates, in the same additive observation-space form as the theoretical model,

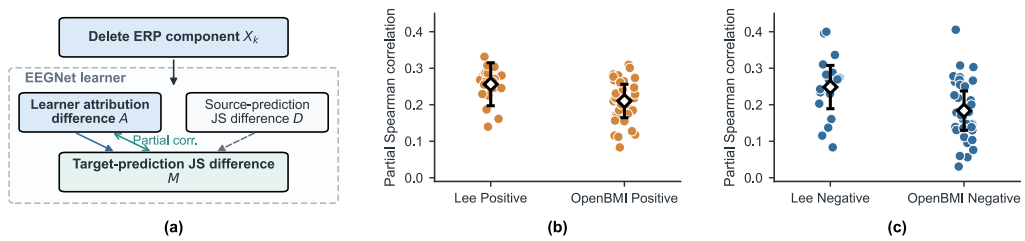

Figure 3: Learner attribution differences and target prediction differences. (a) Analysis schematic. (b,c) Partial Spearman correlations. Dots indicate target subjects; diamonds and error bars show the mean and its 95% model jackknife confidence interval.

and makes this otherwise difficult comparison operationally possible. We use a DG protocol on the dataset of Lee et al. (2024) and OpenBMI (Lee et al., 2019), training 100 EEGNets (Lawhern et al., 2018) per dataset with the same source data and procedure while varying only the random seed (Appendices C.1.1–C.1.2). We delete each reconstructed component $\boldsymbol{X}_k$ one at a time and record the resulting change in logit margin (Appendix C.1.4). For each learner pair and each ERP class (positive or negative), A measures the difference in their source component-deletion responses, M their predictive divergence on a fixed target subject, and D their source predictive divergence. We test the target-wise partial Spearman association between A and M after adjusting for D, separately by class (Appendix C.1.5).

Results. Larger source learner-side attribution differences were consistently associated with larger target predictive divergences after adjusting for same-class source predictive divergence (Fig. 3). The mean target-wise partial Spearman correlations were 0.256 (95% model-jackknife CI $[0.197, 0.315]$) and 0.248 ($[0.189, 0.308]$) for the positive and negative classes on Lee, and 0.210 ($[0.165, 0.256]$) and 0.184 ($[0.130, 0.238]$) on OpenBMI. All target-wise estimates were strictly positive in both classes, for all $20/20$ target subjects on Lee and $36/36$ on OpenBMI (Appendix C.2). These results establish a consistent association, but two alternative explanations remain to be examined. Could this adjusted association nevertheless be concentrated among learner pairs with relatively large source predictive divergence? More fundamentally, could A reflect generic learner dissimilarity rather than component-specific attribution? Together, two supplementary analyses further strengthen empirical support for the main finding (Appendices C.3.1–C.3.2).

7 Discussion and Conclusion

A natural expectation in transfer learning is that more complete knowledge of admissible cross-domain changes should make source-predictive information more reliably transferable. Within the additive data-generating structure studied here, our results identify a structural limitation to this expectation, which we term attribution ambiguity. A minimal Gaussian witness shows that this limitation can persist even under deliberately strong Oracle information, beyond finite-sample estimation uncertainty. With equal prior mass on the two Oracle-compatible worlds, this witness leaves exactly one bit of model uncertainty, yet their target Bayes rules are opposite. Earlier identifiability studies leave the cross-domain map or latent structural roles unresolved and use additional assumptions to identify them (Gulrajani & Hashimoto, 2022; Kong et al., 2022; Ng et al., 2025). Our analysis instead fixes the component-specific transformation laws and their correspondence to component types (see Appendix D). The additive EEG model then shows how many-to-one aggregation can leave predictive activity unresolved across these known transformations.

This structural limitation has implications for transfer under aggregated observations. Reliable transfer may require information that distinguishes decision-conflicting attributions, or structure that makes the remaining alternatives share the same optimal decision. Target supervision provides one such source of information. The same issue can arise more broadly in additive-mixture transfer systems with heterogeneous component-specific transformations. Attribution ambiguity thus marks a boundary between knowing how a system changes and knowing the predictive consequences of that change. For such systems, transfer methods may therefore benefit from treating information sufficiency, rather than shift characterization alone, as a central design consideration.

AI USE DISCLOSURE

In this work, we used ChatGPT to assist with translation and language polishing, code development, consistency checks for mathematical notation, and cross-checking theoretical derivations. We also used it to support literature searches and to identify areas where the coverage of related work could be extended. All AI-assisted outputs were manually reviewed and verified by the authors before inclusion. References identified through AI-assisted searches were checked against the original publications. The authors made all final scientific and editorial decisions and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

The assumptions and definitions for the theoretical results are presented in Sections 3–5; complete proofs and supporting constructions are provided in Appendix A. Appendix B documents the source-equivalent constructions, sampling procedures, evaluation, and mechanism controls for Experiment 1. Appendix C documents the datasets and splits, preprocessing, learner training, component decomposition, statistical analysis, and supplementary controls for Experiment 2. The computational environment shared by both experiments is reported immediately before Appendix B.

REFERENCES

Isabela Albuquerque, João Monteiro, Olivier Rosanne, and Tiago H. Falk. Estimating distribution shifts for predicting cross-subject generalization in electroencephalography-based mental workload assessment. *Frontiers in Artificial Intelligence*, 5:992732, 2022. doi: 10.3389/frai.2022.992732. URL https://doi.org/10.3389/frai.2022.992732.

Shai Ben-David and Ruth Urner. On the Hardness of Domain Adaptation and the Utility of Unlabeled Target Samples. In *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pp. 139–153. Springer, 2012. doi: 10.1007/978-3-642-34106-9_14. URL https://doi.org/10.1007/978-3-642-34106-9_14.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 129–136. PMLR, 2010. URL https://proceedings.mlr.press/v9/david10a.html.

Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage*, 56(2):814–825, 2011. doi: 10.1016/j.neuroimage.2010.06.048.

Pierre Comon. Independent Component Analysis, A New Concept? *Signal Processing*, 36(3):287–314, 1994. doi: 10.1016/0165-1684(94)90029-9. URL https://www.sciencedirect.com/science/article/pii/0165168494900299.

Zhiyi Dong, Zixuan Liu, and Yongyi Mao. On the Hardness of Unsupervised Domain Adaptation: Optimal Learners and Information-Theoretic Perspective. In Sarath Chandar, Razvan Pascanu, Eric Eaton, Bing Liu, Rupam Mahmood, and Amal Rannen-Triki (eds.), *Proceedings of The 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pp. 89–111. PMLR, 2026. URL https://proceedings.mlr.press/v330/dong26a.html.

Bradley Efron. Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics*, 9(2):139–158, 1981. doi: 10.2307/3314608. URL https://doi.org/10.2307/3314608.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3): 586–596, 1981. doi: 10.1214/aos/1176345462. URL https://doi.org/10.1214/aos/1176345462.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267, 2013. doi: 10.3389/fnins.2013.00267.

Ishaan Gulrajani and Tatsunori Hashimoto. Identifiability Conditions for Domain Adaptation. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7982–7997. PMLR, 2022. URL https://proceedings.mlr.press/v162/gulrajani22a.html.

Hans Hallez, Bart Vanrumste, Roberta Grech, Joseph Muscat, Wim De Clercq, Anneleen Vergult, Yves D'Asseler, Kenneth P Camilleri, Simon G Fabri, Sabine Van Huffel, and Ignace Lemahieu. Review on solving the forward problem in EEG source analysis. *Journal of NeuroEngineering and Rehabilitation*, 4:46, 2007. doi: 10.1186/1743-0003-4-46. URL https://link.springer.com/article/10.1186/1743-0003-4-46.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014. doi: 10.1016/j.neuroimage.2013.10.067. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811913010914.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Aapo Hyvärinen and Erkki Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(4–5):411–430, 2000. doi: 10.1016/S0893-6080(00)00026-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear Independent Component Analysis: Existence and Uniqueness Results. *Neural Networks*, 12(3):429–439, 1999. doi: 10.1016/S0893-6080(98)00140-3. URL https://www.sciencedirect.com/science/article/pii/S0893608098001403.

Kasra Jalaldoust and Elias Bareinboim. Transportable Representations for Domain Generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12790–12800, 2024. doi: 10.1609/aaai.v38i11.29175. URL https://ojs.aaai.org/index.php/AAAI/article/view/29175.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *International Conference on Learning Representations*, pp. 16405–16426, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47393e8594c82ce8fd83adc672cf9872-Abstract-Conference.html.

Fredrik D. Johansson, David Sontag, and Rajesh Ranganath. Support and Invertibility in Domain-Invariant Representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 527–536. PMLR, 2019. URL https://proceedings.mlr.press/v89/johansson19a.html.

Ard Kastrati, Josua Bürki, Jonas Lauer, Cheng Xuan, Raffaele Iaquinto, and Roger Wattenhofer. EEG-Bench: A benchmark for EEG foundation models in clinical applications. arXiv preprint, 2025. URL https://arxiv.org/abs/2512.08959. BrainBodyFM@NeurIPS 2025 workshop paper; arXiv:2512.08959v1.

Reinmar J. Kobler, Jun-ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6219–6235, 2022. doi: 10.52202/068431-0450. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/28ef7ee7cd3e03093acc39e1272411b7-Abstract-Conference.html.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial Disentanglement for Domain Adaptation. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11455–11472. PMLR, 2022. URL https://proceedings.mlr.press/v162/kong22a.html.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659. URL https://doi.org/10.3389/fnhum.2021.653659.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c. URL https://iopscience.iop.org/article/10.1088/1741-2552/aace8c.

Gabriel Jason Lee, Jathurshan Pradeepkumar, and Jimeng Sun. Test-Time Adaptation for EEG Foundation Models: A Systematic Study under Real-World Distribution Shifts. In *Proceedings of the 11th Machine Learning for Healthcare Conference*, volume 340 of *Proceedings of Machine Learning Research*, pp. 948–984. PMLR, 2026. URL https://proceedings.mlr.press/v340/lee26a.html.

Jongmin Lee, Minju Kim, Dojin Heo, Jongsu Kim, Min-Ki Kim, Taejun Lee, Jongwoo Park, HyunYoung Kim, Minho Hwang, Laehyun Kim, and Sung-Phil Kim. A comprehensive dataset for home appliance control using ERP-based BCIs with the application of inter-subject transfer learning. *Frontiers in Human Neuroscience*, 18:1320457, 2024. doi: 10.3389/fnhum.2024.1320457. URL https://www.frontiersin.org/articles/10.3389/fnhum.2024.1320457/full.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *GigaScience*, 8(5):giz002, 2019. doi: 10.1093/gigascience/giz002. URL https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giz002/5304369.

Shanglin Li, Motoaki Kawanabe, and Reinmar Kobler. SPDIM: Source-Free Unsupervised Conditional and Label Shift Adaptation in EEG. In *International Conference on Learning Representations*, pp. 101907–101931, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fc6cd575b8340be6ee70beaed9ca4015-Paper-Conference.pdf.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115. URL https://ieeexplore.ieee.org/document/61115/.

Haiqi Liu, C. L. Philip Chen, and Tong Zhang. FACE: Few-shot adapter with cross-view fusion for cross-subject EEG emotion recognition. arXiv preprint, 2025. URL https://arxiv.org/abs/2503.18998. arXiv:2503.18998v1.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. doi: 10.1088/1741-2552/aab2f2.

Steven J. Luck. *An Introduction to the Event-Related Potential Technique*. The MIT Press, Cambridge, MA, 2nd edition, 2014. ISBN 9780262525855. URL https://mitpress.mit.edu/9780262525855/an-introduction-to-the-event-related-potential-technique/.

Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, and Jihun Hamm. Understanding the Limits of Unsupervised Domain Adaptation via Data Poisoning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17347–17359, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/90cc440b1b8caa520c562ac4e4bbcb51-Abstract.html.

Apolline Mellot, Antoine Collas, Sylvain Chevallier, Alexandre Gramfort, and Denis A. Engemann. Geodesic optimization for predictive shift adaptation on EEG data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 32828–32855, 2024. doi: 10.52202/079017-1033. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/399888844ffc8fa95c2db196f3f1b60b-Abstract-Conference.html.

Ignavier Ng, Yan Li, Zijian Li, Yujia Zheng, Guangyi Chen, and Kun Zhang. A General Representation-Based Approach to Multi-Source Domain Adaptation. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45911–45933. PMLR, 2025. URL https://proceedings.mlr.press/v267/ng25a.html.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Bertrand Rivet, Antoine Souloumiac, Virginie Attina, and Guillaume Gibert. xDAWN algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Transactions on Biomedical Engineering*, 56(8):2035–2043, 2009. doi: 10.1109/TBME.2009.2012869.

Jan Sosulski, Jan-Philipp Kemmer, and Michael Tangermann. Improving covariance matrices derived from tiny training datasets for the classification of event-related potentials with linear discriminant analysis. *Neuroinformatics*, 19(3):461–476, 2021. doi: 10.1007/s12021-020-09501-8.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009. doi: 10.1007/b13794.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok (eds.), *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006. URL https://www.cambridge.org/core/books/abs/compressed-sensing/introduction-to-the-nonasymptotic-analysis-of-random-matrices/EC438B0A9D9ED5E010D76918C8E9E9A1.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280, 2024a. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In *International Conference on Learning Representations*, pp. 75310–75346, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbbd6d915cb90be21c1254a82d45cedd-Abstract-Conference.html.

Yihe Wang, Taida Li, Yujun Yan, Wenzhan Song, and Xiang Zhang. What causes performance degradation in cross-subject EEG classification? arXiv preprint, 2024b. URL https://arxiv.org/abs/2410.03057. arXiv:2410.03057v3, revised 2026; scheduled for the IEEE SMC 2026 Brain-Machine Interface Workshop.

Dongrui Wu, Yifan Xu, and Bao-Liang Lu. Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2022. doi: 10.1109/TCDS.2020.3007453. URL https://doi.org/10.1109/TCDS.2020.3007453.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A Comprehensive Benchmark for the Systematic Evaluation and Diagnostic Analyses of EEG Foundation Models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=vGeNaFHdET.

Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968. doi: 10.1214/aoms/1177698520.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008.

Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2018. doi: 10.1109/TBME.2017.2742541.

Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, Qingsong Liu, and Clark Glymour. Domain Adaptation as a Problem of Inference on Graphical Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4965–4976, 2020. URL https://proceedings.neurips.cc/paper/2020/file/3430095c577593aad3c39c701712bcfe-Paper.pdf.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On Learning Invariant Representations for Domain Adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532. PMLR, 2019. URL https://proceedings.mlr.press/v97/zhao19a.html.

Appendix: Proofs and Supplementary Experiments

A Proofs and Supplementary Theoretical Results

This appendix proves Proposition 1, Proposition 2, Lemma 1, Theorem 1, and Theorem 2 from the main text. Section A.2 constructs two Gaussian models with identical Oracle information and opposite target Bayes rules. Section A.3 characterizes when a general compatible pair admits no common Bayes rule. Section A.4 relates the cost of a common decision to the overlap of the protocol observation laws, and A.5 proves that the resulting bound is attained in the Gaussian example. Section A.6 establishes a finite-sample guarantee under additional structure, first proving an explicit stability theorem and then deriving Theorem 2 of the main text.

The supplementary results in A.7 strictly separate latent structure, risk, and optimal decisions and give related consequences. Section A.8 covers generative constructions, population recovery, and the scope of the ERP subclass. Section A.9 contains the moment concentration proof used in A.6 and technical details on the measure-theoretic versions used in A.3. Sections A.2–A.6 can be read without first reading A.7–A.8.

A.1 Notation and Conventions

All finite-dimensional Euclidean spaces carry their Borel σ-algebras, and the label space is finite. Let $\mathcal{F}_{\mathrm{all}}$ be the class of all measurable classifiers from the target observation space to the label space. For a candidate generative model W, define

$$\begin{aligned} R_W^{(\mathcal{T})}(f) &= P_{W,\mathcal{T}}^{XY}[f(\boldsymbol{X}^{(\mathcal{T})}) \neq Y], \\ r_W^{\mathrm{Bayes}} &= \inf_{f \in \mathcal{F}_{\mathrm{all}}} R_W^{(\mathcal{T})}(f), \\ \ell_W^{\mathrm{Bayes}}(f) &= R_W^{(\mathcal{T})}(f) - r_W^{\mathrm{Bayes}}. \end{aligned} \tag{1}$$

A measurable pointwise optimal choice attains the Bayes risk in a finite label space. Target risk always refers to a test observation independent of the training data; expectations for randomized learners also include their internal randomness. Sections A.2–A.5 use the Bayes risk based on the full target observation. Section A.6 instead uses the optimal risk based on fixed ERP features, defined there.

We retain the additive generative class $\mathfrak{W}_{\mathrm{add}}$, fixed lead fields, and fixed, label-independent component-specific operators defined by $L_k(\boldsymbol{J}) = \boldsymbol{G}^{(\mathcal{T})} T_k(\boldsymbol{J})$ from the main text. Write $L = (L_1, L_2)$ and $D = L_1 - L_2$. Two models are Oracle-compatible if and only if they have the same

$$\Phi_{\mathrm{orc}}(W) = \big(P_{W,\mathcal{S}}^{XY}, P_{W,\mathcal{T}}^{X}, P_W^Y, L\big). \tag{2}$$

Here Φ_{N} denotes the standard normal cumulative distribution function, and $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm. Binary constructions use labels $\{-1, +1\}$ with the convention $\mathrm{sign}(0) = +1$; tie-breaking on null sets does not affect risk. For probability measures on the same measurable space, we use $\mathrm{TV}(Q, Q') = \sup_B |Q(B) - Q'(B)|$.

A.2 Proof of Proposition 1

Proposition 1 (Oracle non-identifiability of target Bayes rules, restated). There exists a fixed binary model class satisfying the additive generative structure and Assumption 1 of the main text, containing two models W_0, W_Z with identical Φ_{orc} whose target Bayes classification rules are each unique almost everywhere and opposite almost everywhere under their shared target marginal.

Proof. The reallocation cancels in the source domain but reverses the target class-conditional means. Equal class priors ensure that this reversal preserves the target marginal.

Fixed operators and source equivalence. Take both the neural source-activity and observation spaces to be $\mathbb{R}$, and fix

$$\boldsymbol{G}^{(\mathcal{S})} = \boldsymbol{G}^{(\mathcal{T})} = 1, \qquad T_1 = L_1 = 2I, \qquad T_2 = L_2 = I.$$

Let Y take the values ± 1 with equal probability, and let $\xi \sim \mathcal{N}(0, \sigma^2)$ be independent of Y, where $\Delta, \sigma > 0$. On a common probability space, define

$$\begin{aligned}(\boldsymbol{J}_{1,W_0}, \boldsymbol{J}_{2,W_0}) &= (3\Delta Y/2, -2\Delta Y + \xi),\\ (\boldsymbol{J}_{1,W_Z}, \boldsymbol{J}_{2,W_Z}) &= (-\Delta Y/2, \xi).\end{aligned} \tag{3}$$

The two models differ by the reallocation $\boldsymbol{Z} = -2\Delta Y$ and have different latent joint distributions, but both produce

$$\boldsymbol{X}^{(\mathcal{S})} = -\Delta Y/2 + \xi.$$

The common coupling also uses the same Y, so the source observation–label joint distributions coincide.

Oracle compatibility. Applying the fixed target operators gives

$$\boldsymbol{X}^{(\mathcal{T})}_{W_0} = \Delta Y + \xi, \qquad \boldsymbol{X}^{(\mathcal{T})}_{W_Z} = -\Delta Y + \xi. \tag{4}$$

By symmetry of the equal class priors, the shared target marginal is

$$P_X = \tfrac{1}{2}\mathcal{N}(-\Delta, \sigma^2) + \tfrac{1}{2}\mathcal{N}(\Delta, \sigma^2). \tag{5}$$

The models also share label semantics, class priors in both domains, and L. Thus $\Phi_{\mathrm{orc}}(W_0) = \Phi_{\mathrm{orc}}(W_Z)$.

Opposite target Bayes rules. Let φ_σ denote the zero-mean Gaussian density with standard deviation σ. Equation (4) gives the equal-prior log-likelihood ratios

$$\begin{aligned}\log \frac{p_{W_0}(\boldsymbol{x} \mid Y = +1)}{p_{W_0}(\boldsymbol{x} \mid Y = -1)} &= \log \frac{\varphi_\sigma(\boldsymbol{x} - \Delta)}{\varphi_\sigma(\boldsymbol{x} + \Delta)} = \frac{2\Delta \boldsymbol{x}}{\sigma^2},\\ \log \frac{p_{W_Z}(\boldsymbol{x} \mid Y = +1)}{p_{W_Z}(\boldsymbol{x} \mid Y = -1)} &= -\frac{2\Delta \boldsymbol{x}}{\sigma^2}.\end{aligned} \tag{6}$$

The two posteriors therefore satisfy

$$\eta_{W_0}(+1 \mid \boldsymbol{x}) = \left[1 + \exp\left(-\frac{2\Delta \boldsymbol{x}}{\sigma^2}\right)\right]^{-1}, \qquad \eta_{W_Z}(+1 \mid \boldsymbol{x}) = 1 - \eta_{W_0}(+1 \mid \boldsymbol{x}). \tag{7}$$

For $\boldsymbol{x} \neq 0$, the Bayes rules are $f^*_{W_0}(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{x})$ and $f^*_{W_Z}(\boldsymbol{x}) = -\operatorname{sign}(\boldsymbol{x})$, respectively. Both select $+1$ at $\boldsymbol{x} = 0$. The posteriors tie only at this point, and $P_X(\{0\}) = 0$. Each rule is therefore unique almost everywhere, and the two are opposite almost everywhere. □

Remark (Coupled observations and marginal distributions). Under the common coupling, a general reallocation $(\boldsymbol{J}_1, \boldsymbol{J}_2) \mapsto (\boldsymbol{J}_1 + \boldsymbol{Z}, \boldsymbol{J}_2 - \boldsymbol{Z})$ satisfies

$$\begin{aligned}\boldsymbol{X}^{(\mathcal{S})}_{W_Z} - \boldsymbol{X}^{(\mathcal{S})}_{W_0} &= \boldsymbol{G}^{(\mathcal{S})}(\boldsymbol{Z} - \boldsymbol{Z}) = 0,\\ \boldsymbol{X}^{(\mathcal{T})}_{W_Z} - \boldsymbol{X}^{(\mathcal{T})}_{W_0} &= (L_1 - L_2)(\boldsymbol{Z}) = D(\boldsymbol{Z}).\end{aligned} \tag{8}$$

The target observations coincide almost surely if and only if $D(\boldsymbol{Z}) = 0$ almost surely. A positive probability of nonzero $D(\boldsymbol{Z})$ implies only that the coupled target observations differ with positive probability; it does not imply different marginals or conflicting actions. In this construction, $D(\boldsymbol{Z}) = -2\Delta Y$ is nonzero almost surely, yet the target marginals coincide. In particular, $L_1 = L_2$ makes every reallocation invisible to target observations; even when $L_1 \neq L_2$, a nonzero reallocation can lie in $\ker D$. The correspondence between operators and component types remains fixed, whereas $\boldsymbol{Z}$ may depend on Y. The reversal concerns the conditional relationship between target observations and labels with fixed semantics, not the label semantics themselves.

A.3 General characterization of Bayes conflict

Proposition 1 gives a specific pair with no common Bayes rule. We now characterize when this conflict occurs for an arbitrary Oracle-compatible pair, using the pairwise Bayes conflict from the main text:

$$C_{\mathrm{conf}}(W, W') = \inf_{f \in \mathcal{F}_{\mathrm{all}}} \left[\ell^{\mathrm{Bayes}}_W(f) + \ell^{\mathrm{Bayes}}_{W'}(f)\right]. \tag{9}$$

Lemma 1 (General characterization of positive pairwise Bayes conflict). Let the label space be $[K]$, and let $\mu = P^X_{W,\mathcal{T}} + P^X_{W',\mathcal{T}}$ and $p_V = dP^X_{V,\mathcal{T}}/d\mu$, where $V \in \{W, W'\}$. Denote the set of Bayes-optimal classes at input $\boldsymbol{x}$ by $A^*_V(\boldsymbol{x})$, and define

$$\mathcal{X}_{\mathrm{conf}}(W, W') = \{\boldsymbol{x} : p_W(\boldsymbol{x})p_{W'}(\boldsymbol{x}) > 0,\ A^*_W(\boldsymbol{x}) \cap A^*_{W'}(\boldsymbol{x}) = \varnothing\}. \tag{10}$$

Then $C_{\mathrm{conf}}(W, W') > 0$ if and only if $\mu(\mathcal{X}_{\mathrm{conf}}(W, W')) > 0$. Allowing randomization over classes at each test input changes neither the minimum total Bayes excess risk nor its positivity condition.

Proof. We express optimization over a common classifier as pointwise minimization and characterize when the pointwise minimum is strictly positive. Fix measurable regular conditional versions of the posteriors and nonnegative measurable versions of the densities. All pointwise relations below hold almost everywhere under the relevant dominating measure.

Integral representation of the conflict. For $a \in [K]$, write

$$\begin{aligned}
\eta_V(a \mid \boldsymbol{x}) &= P^{XY}_{V,\mathcal{T}}(Y = a \mid \boldsymbol{X}^{(\mathcal{T})} = \boldsymbol{x}),\\
c_V(\boldsymbol{x}, a) &= \max_{b \in [K]} \eta_V(b \mid \boldsymbol{x}) - \eta_V(a \mid \boldsymbol{x}),\\
A^*_V(\boldsymbol{x}) &= \arg\max_{a \in [K]} \eta_V(a \mid \boldsymbol{x}).
\end{aligned}$$

The quantity $c_V(\boldsymbol{x}, a)$ is the pointwise Bayes excess cost of choosing class a. Decomposing conditional risk gives

$$\begin{aligned}
R^{(\mathcal{T})}_V(f) &= \int [1 - \eta_V(f(\boldsymbol{x}) \mid \boldsymbol{x})]\, P^X_{V,\mathcal{T}}(d\boldsymbol{x}),\\
r^{\mathrm{Bayes}}_V &= \int [1 - \max_a \eta_V(a \mid \boldsymbol{x})]\, P^X_{V,\mathcal{T}}(d\boldsymbol{x}),\\
\ell^{\mathrm{Bayes}}_V(f) &= \int c_V(\boldsymbol{x}, f(\boldsymbol{x}))\, P^X_{V,\mathcal{T}}(d\boldsymbol{x}).
\end{aligned} \tag{11}$$

Under the common dominating measure, the total excess risk across the two models is

$$\int \big[p_W(\boldsymbol{x})c_W(\boldsymbol{x}, f(\boldsymbol{x})) + p_{W'}(\boldsymbol{x})c_{W'}(\boldsymbol{x}, f(\boldsymbol{x}))\big] d\mu(\boldsymbol{x}).$$

Let

$$g(\boldsymbol{x}) = \min_{a \in [K]} \big[p_W(\boldsymbol{x})c_W(\boldsymbol{x}, a) + p_{W'}(\boldsymbol{x})c_{W'}(\boldsymbol{x}, a)\big].$$

Every common classifier has total excess risk at least $\int g\, d\mu$. Conversely, let $f^\dagger(\boldsymbol{x})$ select the smallest-index class attaining the minimum above. The finite label space makes this choice measurable, and it attains the lower bound pointwise. Hence

$$C_{\mathrm{conf}}(W, W') = \int \min_{a \in [K]} \big[p_W(\boldsymbol{x})c_W(\boldsymbol{x}, a) + p_{W'}(\boldsymbol{x})c_{W'}(\boldsymbol{x}, a)\big] d\mu(\boldsymbol{x}). \tag{12}$$

Positivity of the conflict. If $p_W(\boldsymbol{x}) = 0$, any choice $a \in A^*_{W'}(\boldsymbol{x})$ yields $g(\boldsymbol{x}) = 0$; the case $p_{W'}(\boldsymbol{x}) = 0$ is analogous. On the region $p_W p_{W'} > 0$, the sum of the two nonnegative costs is zero if and only if some class makes both costs zero. Thus

$$\begin{aligned}
g(\boldsymbol{x}) = 0 \iff A^*_W(\boldsymbol{x}) \cap A^*_{W'}(\boldsymbol{x}) \neq \varnothing \quad \text{on } \{p_W p_{W'} > 0\},\\
\{\boldsymbol{x} : g(\boldsymbol{x}) > 0\} = \mathcal{X}_{\mathrm{conf}}(W, W') \quad \mu\text{-almost everywhere.}
\end{aligned}$$

Finiteness of the label set is essential here: if the optimal class sets are disjoint, every class has strictly positive weighted total cost, and so does their minimum. Together with $g \geq 0$ and equation (12), this proves the positivity condition. Using sets of optimal classes also accommodates posterior ties. Two arbitrarily selected Bayes rules can differ even when a common optimal class exists.

Test-time class randomization. For a measurable randomized classifier $\pi : \mathcal{X}^{(\mathcal{T})} \to \Delta_K$, the weighted total excess cost at input $\boldsymbol{x}$ is

$$\sum_{a=1}^{K} \pi_a(\boldsymbol{x}) \big[p_W(\boldsymbol{x})c_W(\boldsymbol{x}, a) + p_{W'}(\boldsymbol{x})c_{W'}(\boldsymbol{x}, a)\big] \geq g(\boldsymbol{x}).$$

This expression is linear in $\pi(\boldsymbol{x})$, so a vertex of the simplex attains its minimum, which remains $g(\boldsymbol{x})$. Randomization therefore changes neither the minimum total excess risk nor its positivity condition. □

Remark (Simplification under a shared marginal). For Oracle-compatible models, $\mu = 2P_X$ and $p_W = p_{W'} = 1/2$ almost everywhere. Equation (12) therefore simplifies to

$$C_{\mathrm{conf}}(W, W') = \int \min_{a \in [K]} [c_W(\boldsymbol{x}, a) + c_{W'}(\boldsymbol{x}, a)] \, P_X(d\boldsymbol{x}). \tag{13}$$

The weighted form matches the common-support notation in the lemma and also applies to finite-label models with different target marginals. Section A.9.2 verifies invariance to the chosen versions; A.7.2 gives consequences for common Bayes classifiers and restricted classifier classes.

A.4 Proof of Theorem 1

The cost of a common decision is not itself the learning risk: the protocol may provide information that distinguishes the models. The remaining overlap of the observation laws converts this decision cost into a risk lower bound (Tsybakov, 2009).

Theorem 1 (Protocol-relative Bayes excess-risk lower bound, restated). Fix two Oracle-compatible models W, W'. Let $\mathbb{Q}_{\mathcal{R},V}$ be the joint distribution of all information visible under the protocol in model V, including the shared Oracle information and any permitted additional observations. Then

$$\inf_{\mathsf{A}} \max_{V \in \{W, W'\}} \mathbb{E}_{V,\mathsf{A}}\big[\ell_V^{\mathrm{Bayes}}(\widehat{f}_{\mathsf{A}})\big] \geq \frac{C_{\mathrm{conf}}(W, W')}{2} \left[1 - \mathrm{TV}\big(\mathbb{Q}_{\mathcal{R},W}, \mathbb{Q}_{\mathcal{R},W'}\big)\right]. \tag{14}$$

The infimum ranges over all learning rules, including randomized rules. The expectation includes both protocol data and internal learner randomness.

Proof. On the overlapping part of the observation laws, the learner must use the same conditional output law. Over this common mass, the total excess risk across the two models is at least C_{conf}.

Let $\mathsf{O}_{\mathcal{R}}$ be the protocol observation space, and abbreviate $Q = \mathbb{Q}_{\mathcal{R},W}$ and $Q' = \mathbb{Q}_{\mathcal{R},W'}$. Take $\nu_0 = Q + Q'$, let $h = dQ/d\nu_0$ and $h' = dQ'/d\nu_0$, and define the common submeasure ν by

$$\frac{d\nu}{d\nu_0} = \min\{h, h'\}.$$

Then $\nu \leq Q, Q'$, and

$$\begin{aligned} \nu(\mathsf{O}_{\mathcal{R}}) &= \int \min\{h, h'\} \, d\nu_0 \\ &= 1 - \frac{1}{2} \int |h - h'| \, d\nu_0 = 1 - \mathrm{TV}(Q, Q'). \end{aligned} \tag{15}$$

Fix an arbitrary learning rule A. Equip its allowed classifier output space with a σ-algebra making each map $f \mapsto \ell_V^{\mathrm{Bayes}}(f)$ measurable, and represent its conditional output law by the Markov kernel $\kappa_{\mathsf{A}}(df \mid o)$. The same observation o must induce the same output kernel, without additional dependence on the unknown true model. If the output permits test-time class randomization, A.3 ensures that the lower bound on total conflict remains C_{conf}.

For brevity, write $\ell_V = \ell_V^{\mathrm{Bayes}}$. Nonnegativity, $\nu \leq Q, Q'$, and equation (9) give

$$\begin{aligned} &\mathbb{E}_{W,\mathsf{A}} \ell_W(\widehat{f}_{\mathsf{A}}) + \mathbb{E}_{W',\mathsf{A}} \ell_{W'}(\widehat{f}_{\mathsf{A}}) \\ &\quad \geq \int_{\mathsf{O}_{\mathcal{R}}} \int [\ell_W(f) + \ell_{W'}(f)] \, \kappa_{\mathsf{A}}(df \mid o) \, \nu(do) \\ &\quad \geq C_{\mathrm{conf}}(W, W') \, \nu(\mathsf{O}_{\mathcal{R}}) \\ &\quad = C_{\mathrm{conf}}(W, W') \, [1 - \mathrm{TV}(Q, Q')]. \end{aligned} \tag{16}$$

The maximum is at least the average of the two terms, yielding the right-hand side of equation (14) for this arbitrary A. Taking the infimum over A proves the result. □

Remark A.1 (Joint protocol observation laws). With only the shared Oracle information, $Q = Q'$, and the bound is $C_{\mathrm{conf}}/2$. Adding labeled source data and unlabeled target data sampled independently under the usual sampling scheme still does not distinguish the models, since the Oracle information already determines their distributions. The same bound therefore applies to weaker information access through these samples. Independence cannot be omitted: additional pairing can change the joint observation law even when the individual marginals coincide. The main text lists UDA, FSDA, and DG only to illustrate differences in information access. The strengthened Oracle setting here includes the full target marginal and is not the usual DG protocol, which does not access target data during training.

A.5 Gaussian example and proof of Proposition 2

We return to the Gaussian pair in A.2 and show that the general bound in A.4 is attained exactly in this two-model problem. Keep the same parameters and write $\gamma = \Delta/\sigma > 0$. We compute the conflict and the total variation between protocol observation laws, then give a likelihood-ratio learning rule attaining the bound.

A.5.1 Bayes Risk, Pairwise Bayes Conflict, and the Oracle-Only Minimax Value

By equation (4), $Y\boldsymbol{X}^{(\mathcal{T})}$ has distribution $\mathcal{N}(\Delta, \sigma^2)$ under W_0 and $\mathcal{N}(-\Delta, \sigma^2)$ under W_Z. The corresponding Bayes error events are $Y\boldsymbol{X}^{(\mathcal{T})} < 0$ and $Y\boldsymbol{X}^{(\mathcal{T})} > 0$. Thus

$$r_{W_0}^{\mathrm{Bayes}} = r_{W_Z}^{\mathrm{Bayes}} = \Phi_{\mathrm{N}}(-\gamma). \tag{17}$$

The target posteriors are complementary and the input marginals coincide. For any common classifier, including one with test-time randomization, the two conditional error probabilities therefore sum pointwise to 1, giving

$$\begin{aligned} R_{W_0}^{(\mathcal{T})}(f) + R_{W_Z}^{(\mathcal{T})}(f) &= 1, \\ C_{\mathrm{conf}}(W_0, W_Z) &= 1 - 2\Phi_{\mathrm{N}}(-\gamma) = 2\Phi_{\mathrm{N}}(\gamma) - 1. \end{aligned} \tag{18}$$

With only the shared Oracle information, Theorem 1 gives the lower bound $\Phi_{\mathrm{N}}(\gamma) - 1/2$. Choosing the two opposite Bayes rules with equal probability attains it: under either model, the correct rule has risk $\Phi_{\mathrm{N}}(-\gamma)$, and the reversed rule has risk $1 - \Phi_{\mathrm{N}}(-\gamma)$. Hence

$$R_0^\star = \tfrac{1}{2} C_{\mathrm{conf}}(W_0, W_Z) = \Phi_{\mathrm{N}}(\gamma) - \tfrac{1}{2}. \tag{19}$$

A.5.2 Proposition 2: The Lower Bound and an Attaining Rule

Proposition 2 (Exact minimax excess risk with labeled target samples, restated). The learner knows that the candidate set is exactly $\{W_0, W_Z\}$, together with both models' full definitions, shared parameters Δ, σ, and Bayes rules. In addition to the shared Oracle information, it receives m i.i.d. labeled target samples. For every integer $m \geq 0$,

$$\begin{aligned} R_m^\star &:= \inf_{\mathsf{A}} \max_{V \in \{W_0, W_Z\}} \mathbb{E}_{V,\mathsf{A}}\big[\ell_V^{\mathrm{Bayes}}(\widehat{f}_{\mathsf{A}})\big] \\ &= \big[2\Phi_{\mathrm{N}}(\gamma) - 1\big]\Phi_{\mathrm{N}}(-\sqrt{m}\,\gamma). \end{aligned} \tag{20}$$

Proof. Section A.5.1 gives the optimal value for $m = 0$. We prove the lower bound and achievability for $m \geq 1$, then check the unified formula.

Lower bound. Let $U_i = Y_i \boldsymbol{X}_i^{(\mathcal{T})}$. Under the two models,

$$U_i \mid W_0 \sim \mathcal{N}(\Delta, \sigma^2), \qquad U_i \mid W_Z \sim \mathcal{N}(-\Delta, \sigma^2).$$

Within each model, U_i and Y_i are independent. The transformation $(\boldsymbol{X}_i^{(\mathcal{T})}, Y_i) \mapsto (U_i, Y_i)$ is invertible, and the distribution of Y_i is the same across models. The Oracle information is also shared. The log-likelihood ratio for all protocol observations is therefore

$$\log \frac{d\mathbb{Q}_{\mathcal{R},W_0}}{d\mathbb{Q}_{\mathcal{R},W_Z}} = \frac{2\Delta}{\sigma^2} \sum_{i=1}^{m} U_i. \tag{21}$$

Write $S_m = \sum_{i=1}^{m} U_i$. Then

$$S_m \mid W_0 \sim \mathcal{N}(m\Delta, m\sigma^2), \qquad S_m \mid W_Z \sim \mathcal{N}(-m\Delta, m\sigma^2).$$

The likelihood ratio exceeds 1 on $\{S_m > 0\}$, so the total variation is

$$\begin{aligned} \mathrm{TV}(\mathbb{Q}_{\mathcal{R},W_0}, \mathbb{Q}_{\mathcal{R},W_Z}) &= \mathbb{Q}_{\mathcal{R},W_0}(S_m > 0) - \mathbb{Q}_{\mathcal{R},W_Z}(S_m > 0) \\ &= 2\Phi_{\mathrm{N}}(\sqrt{m}\,\gamma) - 1. \end{aligned} \tag{22}$$

Substituting $1 - \mathrm{TV} = 2\Phi_{\mathrm{N}}(-\sqrt{m}\,\gamma)$ and equation (18) into Theorem 1 gives the right-hand side of equation (20) as a lower bound on $R_m^\star$.

Achievability. Define the learning rule

$$\widehat{f}_m = \begin{cases} f^*_{W_0}, & S_m \geq 0, \\ f^*_{W_Z}, & S_m < 0. \end{cases} \tag{23}$$

The two Gaussian laws of S_m give the same probability of choosing the wrong rule:

$$\mathbb{Q}_{\mathcal{R},W_0}(S_m < 0) = \mathbb{Q}_{\mathcal{R},W_Z}(S_m \geq 0) = \Phi_{\mathrm{N}}(-\sqrt{m}\,\gamma).$$

Choosing the correct rule incurs zero excess risk. Choosing the wrong rule incurs

$$\ell^{\mathrm{Bayes}}_{W_0}(f^*_{W_Z}) = \ell^{\mathrm{Bayes}}_{W_Z}(f^*_{W_0}) = 1 - 2\Phi_{\mathrm{N}}(-\gamma) = 2\Phi_{\mathrm{N}}(\gamma) - 1.$$

Thus, under every $V \in \{W_0, W_Z\}$,

$$\mathbb{E}_V \ell^{\mathrm{Bayes}}_V(\widehat{f}_m) = \big[2\Phi_{\mathrm{N}}(\gamma) - 1\big]\Phi_{\mathrm{N}}(-\sqrt{m}\,\gamma).$$

The same learning rule attains equal risks under both models and matches the lower bound, proving the minimax identity.

The case $m = 0$**.** Equation (19) and $\Phi_{\mathrm{N}}(0) = 1/2$ show that equation (20) also holds for $m = 0$. This case uses the independent fair random choice in A.5.1; substituting $S_0 = 0$ into the deterministic tie-breaking rule in equation (23) would not give that choice. For $m \geq 1$, the event $S_m = 0$ has probability zero. □

Remark A.2 (Scope of the exact value). For every finite m, $R_m^\star > 0$, and the value tends to zero as $m \to \infty$. This proposition concerns a two-model problem with fully known candidates and parameters, rather than a general FSDA sample-complexity result for unknown candidates or parameters.

A.6 Proof of Theorem 2

This proof uses only the class-conditional distributions of the fixed ERP features, moment-prediction conditions, and independent sampling. It does not require latent-attribution recovery. Lemma A.1 provides a joint moment concentration event. On this event, we successively control the target covariance, the mean-contrast direction up to sign, the choice of sign, and the discriminant normal, then convert parameter errors into a quadratic excess-risk bound. Theorem A.1 retains explicit error terms, which A.6.5 combines into the rate in Theorem 2 of the main text. The generative constructions and population identifiability results in A.8 are not additional premises of this proof.

A.6.1 Feature Model, Risk Benchmark, and Uniform Conditions

Fix a prespecified unit-norm temporal template $\boldsymbol{h} \in \mathbb{R}^T$, and let $\boldsymbol{U}^{(e)} = \Psi(\boldsymbol{X}^{(e)}) = \boldsymbol{X}^{(e)}\boldsymbol{h} \in \mathbb{R}^d$, where $e \in \{\mathcal{S}, \mathcal{T}\}$. Within the additive generative class of the main text, consider the binary ERP subclass with equal class priors satisfying

$$\boldsymbol{U}^{(e)} \mid Y = y \sim \mathcal{N}(\boldsymbol{\mu}_e + y\boldsymbol{\delta}_e/2, \boldsymbol{\Sigma}_e), \qquad P(Y = +1) = P(Y = -1) = \tfrac{1}{2}. \tag{24}$$

Here $\boldsymbol{\delta}_e$ is the full class-mean contrast, class $+1$ minus class -1; $\boldsymbol{\Sigma}_e$ is the common within-class covariance; and $\boldsymbol{\mu}_e$ is also the marginal mean. The target marginal covariance therefore satisfies

$$\boldsymbol{K}_{\mathcal{T}} := \mathrm{Cov}(\boldsymbol{U}^{(\mathcal{T})}) = \boldsymbol{\Sigma}_{\mathcal{T}} + \tfrac{1}{4}\boldsymbol{\delta}_{\mathcal{T}}\boldsymbol{\delta}_{\mathcal{T}}^\top. \tag{25}$$

Let $\boldsymbol{b}_\star \in \mathbb{R}^d$ be the model-defined reference vector from the main text used to orient the class-mean contrast. Its generative expression and distinction from a component contribution are given in A.8.1. The protocol supplies nominal maps $\boldsymbol{A}_{\mathrm{bg}}, \boldsymbol{A}_{\mathrm{sig}} \in \mathbb{R}^{d\times d}$ fixed across candidate models; they cannot be chosen using the unknown true model. Unless implied by the known forward maps and fixed source-activity space, they constitute additional side information. Write

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathrm{nom}} &= \boldsymbol{A}_{\mathrm{bg}}\boldsymbol{\Sigma}_{\mathcal{S}}\boldsymbol{A}_{\mathrm{bg}}^{\top},\\ \rho_{\mathrm{bg}} &= \left\|\boldsymbol{\Sigma}_{\mathcal{T}}^{-1/2}(\boldsymbol{\Sigma}_{\mathcal{T}}-\boldsymbol{\Sigma}_{\mathrm{nom}})\boldsymbol{\Sigma}_{\mathcal{T}}^{-1/2}\right\|_{\mathrm{op}},\\ \varepsilon_{\mathrm{sig}} &= \|\boldsymbol{A}_{\mathrm{sig}}\boldsymbol{\delta}_{\mathcal{S}}-\boldsymbol{b}_\star\|_2.\end{aligned} \tag{26}$$

Assumption 2. There exist constants, fixed across candidate models,

$$0<\lambda_-\le\lambda_+<\infty,\quad 0<\Delta_0\le\Delta_+<\infty,\quad b_0>0,\quad L_{\mathrm{sig}}<\infty,\quad 0<\kappa_0\le 1,$$

such that every candidate model satisfies

$$\begin{aligned}&\lambda_-\boldsymbol{I}_d\preceq\boldsymbol{\Sigma}_e\preceq\lambda_+\boldsymbol{I}_d \quad (e\in\{\mathcal{S},\mathcal{T}\}),\\ &\Delta_0\le\|\boldsymbol{\delta}_{\mathcal{T}}\|_2\le\Delta_+, \qquad \|\boldsymbol{b}_\star\|_2\ge b_0,\\ &\|\boldsymbol{A}_{\mathrm{sig}}\|_{\mathrm{op}}\le L_{\mathrm{sig}}, \qquad \frac{\boldsymbol{b}_\star^{\top}\boldsymbol{\delta}_{\mathcal{T}}}{\|\boldsymbol{b}_\star\|_2\|\boldsymbol{\delta}_{\mathcal{T}}\|_2}\ge\kappa_0.\end{aligned} \tag{27}$$

Here d is the dimension of the fixed nonredundant observation coordinates. If the observations satisfy deterministic linear constraints, redundant coordinates must first be removed before positive definiteness is imposed on the covariances. The covariance spectral bounds control moment-estimation and inversion stability; Δ_0 provides a rank-one eigengap; and $b_0, L_{\mathrm{sig}}, \kappa_0$ control errors in choosing the sign. The alignment margin relative to the class labels cannot be verified from unlabeled target data.

Denote this subclass from the main text by $\mathfrak{W}_{\mathrm{ERP}}$, and its subclass with both moment-prediction errors equal to zero by $\mathfrak{W}^0_{\mathrm{ERP}}$. This section uses

$$\begin{aligned}\mathcal{F}_{\mathrm{ERP}} &= \{g\circ\Psi : g:\mathbb{R}^d\to\{-1,+1\}\text{ measurable}\},\\ r_W^{\mathrm{ERP}} &= \inf_{f\in\mathcal{F}_{\mathrm{ERP}}} R_W^{(\mathcal{T})}(f),\\ \ell_W^{\mathrm{ERP}}(f) &= R_W^{(\mathcal{T})}(f)-r_W^{\mathrm{ERP}}.\end{aligned} \tag{28}$$

This benchmark permits only the fixed features $\Psi(\boldsymbol{X})$. Section A.8.6 separately accounts for the approximation error relative to the full-EEG Bayes risk.

Theorem 2 (Finite-sample excess risk under moment-prediction errors, restated). Under the Gaussian ERP subclass above and Assumption 2, there exist fixed positive thresholds $\varepsilon_\star, \rho_\star$ and constants $C_0, C_1 > 0$, depending only on the uniform bounds and alignment margin in equation (27), such that the following holds. For any model satisfying $\varepsilon_{\mathrm{sig}} \le \varepsilon_\star$ and $\rho_{\mathrm{bg}} \le \rho_\star$, if

$$n, N \ge C_0[d+\log(16/\alpha)], \qquad \alpha\in(0,1),$$

then the learner defined in A.6.2 satisfies the excess-risk bound, with probability at least $1-\alpha$,

$$\ell_W^{\mathrm{ERP}}(\widehat{f}) \le C_1\left[\frac{d+\log(16/\alpha)}{n}+\frac{d+\log(16/\alpha)}{N}+\rho_{\mathrm{bg}}^2\right]. \tag{29}$$

Section A.6.5 proves this result using the explicit stability theorem below. The reference error enters through the sign-selection condition; the covariance error affects both spectral recovery and the continuous risk error.

A.6.2 Finite-Sample Learning Rule

The learner receives labeled source samples $\{(\boldsymbol{U}_i^{(\mathcal{S})}, Y_i)\}_{i=1}^n$, unlabeled target samples $\{\boldsymbol{U}_j^{(\mathcal{T})}\}_{j=1}^N$, the fixed feature map, and the two nominal maps. The two sample sets are independent, with i.i.d. observations within each set. The learner has no access to population distributions or true moments.

At the population level, equation (25) shows that, given the target within-class covariance, $\boldsymbol{K}_{\mathcal{T}} - \boldsymbol{\Sigma}_{\mathcal{T}} = \boldsymbol{\delta}_{\mathcal{T}}\boldsymbol{\delta}_{\mathcal{T}}^{\top}/4$ determines the one-dimensional subspace of the target mean contrast. The reference

condition and the prediction of $\boldsymbol{b}_\star$ by $\boldsymbol{A}_{\mathrm{sig}}\boldsymbol{\delta}_{\mathcal{S}}$ select the sign. The Gaussian Bayes discriminant normal is then $\boldsymbol{\Sigma}_{\mathcal{T}}^{-1}\boldsymbol{\delta}_{\mathcal{T}}$, whose positive scale does not affect the sign classifier. The learner below is the empirical counterpart of these three steps.

Empirical moments. Let $n_y = \#\{i : Y_i = y\}$. When $n_{+1}, n_{-1} \geq 2$ and $N \geq 1$, define

$$\overline{\boldsymbol{U}}_{\mathcal{S},y} = \frac{1}{n_y}\sum_{i:Y_i=y}\boldsymbol{U}_i^{(\mathcal{S})}, \quad \widehat{\boldsymbol{\delta}}_{\mathcal{S}} = \overline{\boldsymbol{U}}_{\mathcal{S},+1} - \overline{\boldsymbol{U}}_{\mathcal{S},-1},$$
$$\widehat{\boldsymbol{\mu}}_{\mathcal{T}} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{U}_j^{(\mathcal{T})}.$$

Estimate the covariances by

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}} &= \frac{1}{n-2}\sum_{y=\pm 1}\sum_{i:Y_i=y}(\boldsymbol{U}_i^{(\mathcal{S})} - \overline{\boldsymbol{U}}_{\mathcal{S},y})(\boldsymbol{U}_i^{(\mathcal{S})} - \overline{\boldsymbol{U}}_{\mathcal{S},y})^\top,\\ \widehat{\boldsymbol{K}}_{\mathcal{T}} &= \frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{U}_j^{(\mathcal{T})} - \widehat{\boldsymbol{\mu}}_{\mathcal{T}})(\boldsymbol{U}_j^{(\mathcal{T})} - \widehat{\boldsymbol{\mu}}_{\mathcal{T}})^\top.\end{aligned} \tag{30}$$

Spectral recovery and sign selection. Compute

$$\widehat{\boldsymbol{b}} = \boldsymbol{A}_{\mathrm{sig}}\widehat{\boldsymbol{\delta}}_{\mathcal{S}}, \qquad \widehat{\boldsymbol{\Sigma}}_{\mathcal{T}} = \boldsymbol{A}_{\mathrm{bg}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}}\boldsymbol{A}_{\mathrm{bg}}^\top, \qquad \widehat{\boldsymbol{M}} = \widehat{\boldsymbol{K}}_{\mathcal{T}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{T}}. \tag{31}$$

If $\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}} \succ 0$, the largest algebraic eigenvalue of $\widehat{\boldsymbol{M}}$ is positive and simple, and its spectral projector $\widehat{\boldsymbol{P}}_1$ satisfies $\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{b}} \neq 0$, set

$$\widehat{\boldsymbol{v}} = \frac{\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{b}}}{\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{b}}\|_2}, \qquad \widehat{f}(\boldsymbol{X}) = \mathrm{sign}\left[\widehat{\boldsymbol{v}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}}^{-1}(\Psi(\boldsymbol{X}) - \widehat{\boldsymbol{\mu}}_{\mathcal{T}})\right]. \tag{32}$$

If any condition fails, including insufficient class counts, output the constant classifier $+1$. The rule does not require $\widehat{\boldsymbol{K}}_{\mathcal{T}}$ to be positive definite. For $d = 1$, simplicity of the eigenvalue is automatic.

Measurability. The spectral projector of a simple eigenvalue, positive-definite matrix inversion, and nonzero vector normalization are continuous on their respective domains, and the branch conditions are measurable. The output is therefore jointly measurable in the training data and test input, and its risk is measurable. The spectral-projector formulation also avoids an arbitrary eigenvector sign choice.

A.6.3 Explicit Stability Bound and Joint Concentration Event

Fix $\alpha \in (0, 1)$ and let $t = \log(16/\alpha)$. For integers $\nu \geq 1$, define

$$\begin{aligned}a_\nu &= \frac{\sqrt{d} + \sqrt{2t}}{\sqrt{\nu}},\\ c_\nu &= 4\left[\sqrt{\frac{d\log 9 + t}{\nu}} + \frac{d\log 9 + t}{\nu}\right].\end{aligned} \tag{33}$$

The four sampling-error bounds correspond to the source mean contrast, source within-class covariance, target mean, and target marginal covariance:

$$\begin{aligned}\epsilon_\delta &= \sqrt{8\lambda_+}\,a_n, & \eta_n &= c_{n-2},\\ \epsilon_\mu &= \sqrt{\lambda_+}\,a_N + \frac{\Delta_+}{2}\sqrt{\frac{2t}{N}}, & \epsilon_K &= \lambda_+ c_N + \Delta_+\sqrt{\lambda_+}\,a_N + \epsilon_\mu^2.\end{aligned} \tag{34}$$

The four quantities in equation (34) bound sampling errors of the empirical moments. To track their propagation through the deterministic stability argument, define the following derived error bounds; they are neither new estimators nor additional assumptions:

$$\begin{aligned}e_\Sigma &= \rho_{\mathrm{bg}} + (1 + \rho_{\mathrm{bg}})\eta_n, & e_M &= \epsilon_K + \lambda_+ e_\Sigma,\\ e_v &= \frac{8e_M}{\Delta_0^2}, & e_{\mathrm{ref}} &= \frac{2(L_{\mathrm{sig}}\epsilon_\delta + \varepsilon_{\mathrm{sig}})}{b_0}.\end{aligned} \tag{35}$$

Also write $\chi = \sqrt{\lambda_+/\lambda_-}$ and $g_+ = \Delta_+/\sqrt{\lambda_-}$.

Theorem A.1 (Explicit finite-sample stability). For any $W \in \mathfrak{W}_{\mathrm{ERP}}$, if

$$n \geq \max\{8, 8t\}, \qquad N \geq 2, \qquad e_\Sigma \leq \tfrac{1}{2}, \qquad e_{\mathrm{ref}} + e_v \leq \tfrac{\kappa_0}{2}, \tag{36}$$

then, with probability at least $1-\alpha$, the learner in A.6.2 does not enter the fallback branch, and

$$\ell_W^{\mathrm{ERP}}(\widehat{f}) \leq \frac{g_+}{\sqrt{2\pi}} \left(\frac{\chi e_v + e_\Sigma}{1 - e_\Sigma} \right)^2 + \frac{\epsilon_\mu^2}{2\lambda_- \sqrt{2\pi \mathrm{e}}}. \tag{37}$$

These conditions establish stability; the algorithm need not know the true errors or model-class constants. The condition $e_v \leq 1/2$ already implies $e_M \leq \Delta_0^2/16$, so no separate spectral-separation assumption is needed.

The following lemma controls all sampling errors on one event. Its proof is in A.9.1; Theorem A.1 uses only its conclusion.

Lemma A.1 (Joint concentration of empirical moments). Under the distributional and sampling conditions of A.6.1–A.6.2, if $n \geq \max\{8, 8t\}$ and $N \geq 2$, there exists an event $\mathcal{E}$ with $P_W(\mathcal{E}) \geq 1-\alpha$ such that, on $\mathcal{E}$,

$$\begin{aligned} \min(n_{+1}, n_{-1}) &\geq n/4, \\ \|\widehat{\boldsymbol{\delta}}_{\mathcal{S}} - \boldsymbol{\delta}_{\mathcal{S}}\|_2 &\leq \epsilon_\delta, \\ -\eta_n \boldsymbol{\Sigma}_{\mathcal{S}} \preceq \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}} - \boldsymbol{\Sigma}_{\mathcal{S}} &\preceq \eta_n \boldsymbol{\Sigma}_{\mathcal{S}}, \\ \|\widehat{\boldsymbol{\mu}}_{\mathcal{T}} - \boldsymbol{\mu}_{\mathcal{T}}\|_2 \leq \epsilon_\mu, \qquad \|\widehat{\boldsymbol{K}}_{\mathcal{T}} - \boldsymbol{K}_{\mathcal{T}}\|_{\mathrm{op}} &\leq \epsilon_K. \end{aligned} \tag{38}$$

These errors need not be mutually independent. Target labels appear only as generative variables in the proof; the learner does not access them.

A.6.4 Proof of Theorem A.1

Proof. Fix the event $\mathcal{E}$ from Lemma A.1. All subsequent arguments are deterministic on this event.

Step 1: Propagating source covariance error to the target domain. Applying $\boldsymbol{A}_{\mathrm{bg}}(\cdot)\boldsymbol{A}_{\mathrm{bg}}^\top$ to the matrix-order bounds in equation (38) gives

$$-\eta_n \boldsymbol{\Sigma}_{\mathrm{nom}} \preceq \widehat{\boldsymbol{\Sigma}}_{\mathcal{T}} - \boldsymbol{\Sigma}_{\mathrm{nom}} \preceq \eta_n \boldsymbol{\Sigma}_{\mathrm{nom}}.$$

Equation (26) also gives

$$\|\boldsymbol{\Sigma}_{\mathcal{T}}^{-1/2} \boldsymbol{\Sigma}_{\mathrm{nom}} \boldsymbol{\Sigma}_{\mathcal{T}}^{-1/2}\|_{\mathrm{op}} \leq 1 + \rho_{\mathrm{bg}}.$$

Adding prediction mismatch and sampling error yields

$$\begin{aligned} &\left\| \boldsymbol{\Sigma}_{\mathcal{T}}^{-1/2} (\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}} - \boldsymbol{\Sigma}_{\mathcal{T}}) \boldsymbol{\Sigma}_{\mathcal{T}}^{-1/2} \right\|_{\mathrm{op}} \\ &\qquad \leq \rho_{\mathrm{bg}} + (1 + \rho_{\mathrm{bg}})\eta_n = e_\Sigma. \end{aligned} \tag{39}$$

Since $e_\Sigma \leq 1/2$, we have $\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}} \succ 0$. Moreover,

$$\begin{aligned} \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}} - \boldsymbol{\Sigma}_{\mathcal{T}}\|_{\mathrm{op}} &\leq \lambda_+ e_\Sigma, \\ \|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_{\mathrm{op}} \leq \epsilon_K + \lambda_+ e_\Sigma = e_M, \qquad \boldsymbol{M} &= \tfrac{1}{4} \boldsymbol{\delta}_{\mathcal{T}} \boldsymbol{\delta}_{\mathcal{T}}^\top. \end{aligned} \tag{40}$$

Propagation through the matrix-order bounds controls the relative error directly, without an additional factor of $\|\boldsymbol{A}_{\mathrm{bg}}\|_{\mathrm{op}}$.

Step 2: Recovering the mean-contrast direction up to sign. Let

$$\boldsymbol{v} = \frac{\boldsymbol{\delta}_{\mathcal{T}}}{\|\boldsymbol{\delta}_{\mathcal{T}}\|_2}, \qquad \lambda_\star = \frac{\|\boldsymbol{\delta}_{\mathcal{T}}\|_2^2}{4}, \qquad \boldsymbol{M} = \lambda_\star \boldsymbol{v}\boldsymbol{v}^\top.$$

The theorem's conditions ensure $e_M \leq \Delta_0^2/16 \leq \lambda_\star/4$. The variational principle gives $\widehat{\lambda}_1 \geq \lambda_\star - e_M > 0$. When $d \geq 2$, it also gives $\widehat{\lambda}_2 \leq e_M$, so the largest eigenvalue is simple.

For an associated unit eigenvector $\widetilde{\boldsymbol{v}}$, project the eigenvalue equation onto $\boldsymbol{v}^{\perp}$:

$$\widehat{\lambda}_1(\boldsymbol{I}-\boldsymbol{v}\boldsymbol{v}^\top)\widetilde{\boldsymbol{v}} = (\boldsymbol{I}-\boldsymbol{v}\boldsymbol{v}^\top)(\widehat{\boldsymbol{M}}-\boldsymbol{M})\widetilde{\boldsymbol{v}}.$$

The orthogonal component therefore has norm at most $e_M/(\lambda_\star - e_M)$. For the analysis, choose the sign so that $\boldsymbol{v}^\top\overline{\boldsymbol{v}} \geq 0$. For two unit vectors with nonnegative inner product, their distance is at most $\sqrt{2}$ times the norm of this orthogonal component. Using $e_M \leq \lambda_\star/4$ and $\lambda_\star \geq \Delta_0^2/4$ gives

$$\|\overline{\boldsymbol{v}}-\boldsymbol{v}\|_2 \leq \frac{\sqrt{2}e_M}{\lambda_\star - e_M} \leq \frac{2e_M}{\lambda_\star} \leq \frac{8e_M}{\Delta_0^2} = e_v. \tag{41}$$

This is a direct rank-one spectral perturbation argument; see Yu et al. (2015) for a general Davis–Kahan bound. So far, only the unsigned line $\mathrm{span}(\boldsymbol{\delta}_{\mathcal{T}})$ has been recovered: the sign of $\overline{\boldsymbol{v}}$ was chosen using the unknown $\boldsymbol{v}$. We next show that the learner chooses this same sign using the nominal reference provided by the protocol.

Step 3: Choosing the sign using the reference. The nominal reference and source mean error give

$$\|\widehat{\boldsymbol{b}}-\boldsymbol{b}_\star\|_2 \leq L_{\mathrm{sig}}\epsilon_\delta + \varepsilon_{\mathrm{sig}}.$$

For nonzero $\boldsymbol{x}, \boldsymbol{y}$, use

$$\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} \right\|_2 \leq \frac{2\|\boldsymbol{x}-\boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2}. \tag{42}$$

Let $\boldsymbol{s}_\star = \boldsymbol{b}_\star/\|\boldsymbol{b}_\star\|_2$ and $\widehat{\boldsymbol{s}} = \widehat{\boldsymbol{b}}/\|\widehat{\boldsymbol{b}}\|_2$. Then $\|\widehat{\boldsymbol{s}}-\boldsymbol{s}_\star\|_2 \leq e_{\mathrm{ref}}$. The normalization is well defined because $e_{\mathrm{ref}} \leq 1/2$ ensures $\|\widehat{\boldsymbol{b}}-\boldsymbol{b}_\star\|_2 \leq b_0/4 < \|\boldsymbol{b}_\star\|_2$. Combining equation (41) with the alignment margin gives

$$\widehat{\boldsymbol{s}}^\top\overline{\boldsymbol{v}} \geq \boldsymbol{s}_\star^\top\boldsymbol{v} - e_{\mathrm{ref}} - e_v \geq \kappa_0 - e_{\mathrm{ref}} - e_v \geq \kappa_0/2 > 0. \tag{43}$$

Thus $\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{b}} = \overline{\boldsymbol{v}}(\overline{\boldsymbol{v}}^\top\widehat{\boldsymbol{b}}) \neq 0$, with a positive coefficient. The algorithm selects $\widehat{\boldsymbol{v}} = \overline{\boldsymbol{v}}$, and hence $\|\widehat{\boldsymbol{v}}-\boldsymbol{v}\|_2 \leq e_v$. Thus none of the fallback conditions is triggered on $\mathcal{E}$. The reference error ensures the correct sign; it does not subsequently enter the risk as a continuous perturbation term.

Step 4: Controlling the whitened discriminant normal and center. The first three steps control the target mean-contrast direction. The Gaussian classifier uses the discriminant normal obtained after applying the inverse covariance. To compare the true and estimated normals in coordinates whitened by the true covariance, abbreviate $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathcal{T}}$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathcal{T}}$, and write

$$\begin{aligned} \boldsymbol{E}_\Sigma &= \boldsymbol{\Sigma}^{-1/2}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}, \\ \boldsymbol{a} &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}, & \widehat{\boldsymbol{a}} &= \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}}^{-1}\widehat{\boldsymbol{v}}, \\ \boldsymbol{u} &= \boldsymbol{a}/\|\boldsymbol{a}\|_2, & \widehat{\boldsymbol{u}} &= \widehat{\boldsymbol{a}}/\|\widehat{\boldsymbol{a}}\|_2. \end{aligned}$$

The identity

$$\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{I}+\boldsymbol{E}_\Sigma)\boldsymbol{\Sigma}^{1/2},$$

gives the exact relations

$$\begin{aligned} \widehat{\boldsymbol{a}} &= (\boldsymbol{I}+\boldsymbol{E}_\Sigma)^{-1}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}, \\ \widehat{\boldsymbol{a}}-\boldsymbol{a} &= (\boldsymbol{I}+\boldsymbol{E}_\Sigma)^{-1}\big[\boldsymbol{\Sigma}^{-1/2}(\widehat{\boldsymbol{v}}-\boldsymbol{v}) - \boldsymbol{E}_\Sigma\boldsymbol{a}\big]. \end{aligned} \tag{44}$$

Using $\|\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}} \leq \lambda_-^{-1/2}$, $\|\boldsymbol{a}\|_2 \geq \lambda_+^{-1/2}$, and $\|(\boldsymbol{I}+\boldsymbol{E}_\Sigma)^{-1}\|_{\mathrm{op}} \leq (1-e_\Sigma)^{-1}$, we obtain

$$\frac{\|\widehat{\boldsymbol{a}}-\boldsymbol{a}\|_2}{\|\boldsymbol{a}\|_2} \leq \frac{\chi\|\widehat{\boldsymbol{v}}-\boldsymbol{v}\|_2 + e_\Sigma}{1-e_\Sigma} \leq \frac{\chi e_v + e_\Sigma}{1-e_\Sigma}. \tag{45}$$

This argument does not require the matrices to commute. Applying equation (42) again and defining the whitened center error $z = \widehat{\boldsymbol{u}}^\top\boldsymbol{\Sigma}^{-1/2}(\widehat{\boldsymbol{\mu}}_{\mathcal{T}} - \boldsymbol{\mu})$ gives

$$\|\widehat{\boldsymbol{u}}-\boldsymbol{u}\|_2 \leq \frac{2(\chi e_v + e_\Sigma)}{1-e_\Sigma}, \qquad |z| \leq \frac{\epsilon_\mu}{\sqrt{\lambda_-}}. \tag{46}$$

Step 5: Converting perturbations into quadratic Gaussian excess risk. The first four steps control the classifier's direction and center. In the symmetric Gaussian model, both errors contribute

quadratically to excess risk. Let $g = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}_{\mathcal{T}}\|_2$. Since $\boldsymbol{u} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}_{\mathcal{T}}/g$, the whitened feature $\boldsymbol{V} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{U}^{(\mathcal{T})} - \boldsymbol{\mu})$ satisfies

$$\boldsymbol{V} \mid Y = y \sim \mathcal{N}(ygu/2, \boldsymbol{I}_d), \qquad r_W^{\mathrm{ERP}} = \Phi_{\mathrm{N}}(-g/2).$$

The estimated classifier in whitened coordinates is $\mathrm{sign}(\widehat{\boldsymbol{u}}^\top \boldsymbol{V} - z)$. Conditional on the training data, let $k = (g/2)\widehat{\boldsymbol{u}}^\top \boldsymbol{u}$. Its risk is exactly

$$R_W^{(\mathcal{T})}(\widehat{f}) = \tfrac{1}{2}[\Phi_{\mathrm{N}}(-k+z) + \Phi_{\mathrm{N}}(-k-z)]. \tag{47}$$

We bound the effects of direction and center separately. For the direction error,

$$g/2 - k = \frac{g}{4}\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2^2.$$

This identity, together with $\sup_x \Phi'_{\mathrm{N}}(x) = 1/\sqrt{2\pi}$, gives

$$\Phi_{\mathrm{N}}(-k) - \Phi_{\mathrm{N}}(-g/2) \le \frac{g}{4\sqrt{2\pi}}\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2^2.$$

For the center, let $F_k(z) = [\Phi_{\mathrm{N}}(-k+z) + \Phi_{\mathrm{N}}(-k-z)]/2$. Equal priors and symmetric means imply $F_k'(0) = 0$, and $\sup_z |F_k''(z)| \le 1/\sqrt{2\pi\mathrm{e}}$. The Taylor remainder bound therefore gives

$$\begin{aligned} \ell_W^{\mathrm{ERP}}(\widehat{f}) &= F_k(z) - F_k(0) + \Phi_{\mathrm{N}}(-k) - \Phi_{\mathrm{N}}(-g/2) \\ &\le \frac{z^2}{2\sqrt{2\pi\mathrm{e}}} + \frac{g}{4\sqrt{2\pi}}\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2^2. \end{aligned} \tag{48}$$

Substituting equation (46) and using $g \le g_+$ yields equation (37). The second-derivative bound is global and requires no additional higher-order remainder control. All arguments hold on $\mathcal{E}$, and $P_W(\mathcal{E}) \ge 1 - \alpha$. □

A.6.5 Derivation of Theorem 2

Proof of Theorem 2. We first derive the rate from Theorem A.1. Let $D_t = d + t$, $r_n = \sqrt{D_t/n}$, and $r_N = \sqrt{D_t/N}$. For the sample sizes and mismatch range chosen below, equations (34)–(35) give

$$\begin{aligned} \epsilon_\delta, \eta_n &= O(r_n), & \epsilon_\mu, \epsilon_K &= O(r_N), \\ e_\Sigma &= O(\rho_{\mathrm{bg}} + r_n), & e_v &= O(\rho_{\mathrm{bg}} + r_n + r_N). \end{aligned}$$

The constants depend only on the uniform bounds and alignment margin in equation (27). Once the stability conditions of Theorem A.1 hold uniformly, equation (37) yields excess risk of order $O(\rho_{\mathrm{bg}}^2 + r_n^2 + r_N^2)$. It remains to choose mismatch thresholds and a sample-size constant valid throughout the subclass. The following explicit choice verifies this uniformity.

Step 1: Reserving alignment margin for mismatch and sampling. Use the joint sufficient conditions

$$0 \le \rho_{\mathrm{bg}} \le \tfrac{1}{4}, \qquad \frac{2\varepsilon_{\mathrm{sig}}}{b_0} + \frac{8\lambda_+\rho_{\mathrm{bg}}}{\Delta_0^2} \le \frac{\kappa_0}{4}. \tag{49}$$

For example, the fixed thresholds

$$\varepsilon_\star = \frac{\kappa_0 b_0}{16}, \qquad \rho_\star = \min\left\{\frac{1}{4}, \frac{\kappa_0\Delta_0^2}{64\lambda_+}\right\} \tag{50}$$

allocate at most $\kappa_0/8$ to each mismatch term, leaving $\kappa_0/4$ for sampling errors.

Step 2: Choosing sample sizes uniformly. When $n, N \ge D_t$ and $n \ge 4$,

$$(\sqrt{d} + \sqrt{2t})^2 \le 3D_t, \qquad d\log 9 + t \le (\log 9)D_t, \qquad n - 2 \ge n/2.$$

Together with $r_n, r_N \le 1$, these inequalities bound equation (34) by a common constant $c \ge 1$:

$$\epsilon_\delta \le cr_n, \qquad \eta_n \le cr_n, \qquad \epsilon_\mu \le cr_N, \qquad \epsilon_K \le cr_N. \tag{51}$$

The constant c depends only on λ_+, Δ_+. The term ϵ_μ^2 can also be absorbed into order r_N, since $r_N^2 \le r_N$ and ϵ_μ/r_N is uniformly bounded.

Define

$$c_\Sigma = 5c/4, \qquad v_\rho = 8\lambda_+/\Delta_0^2,$$
$$c_v = \frac{8c}{\Delta_0^2}\max\{1, 5\lambda_+/4\}, \quad c_b = 2L_{\mathrm{sig}}c/b_0.$$

Equation (35) gives

$$\begin{aligned} e_\Sigma &\le \rho_{\mathrm{bg}} + c_\Sigma r_n, \\ e_v &\le v_\rho \rho_{\mathrm{bg}} + c_v(r_n + r_N), \\ e_{\mathrm{ref}} + e_v &\le \frac{2\varepsilon_{\mathrm{sig}}}{b_0} + v_\rho\rho_{\mathrm{bg}} + (c_b + c_v)r_n + c_v r_N. \end{aligned} \tag{52}$$

Choose

$$C_0 = \max\left\{8, (4c_\Sigma)^2, \left[\frac{4(c_b + 2c_v)}{\kappa_0}\right]^2\right\}. \tag{53}$$

If $n, N \ge C_0 D_t$, then $n \ge \max\{8, 8t\}$, $N \ge 2$, and

$$c_\Sigma r_n \le \tfrac{1}{4}, \qquad (c_b + c_v)r_n + c_v r_N \le \tfrac{\kappa_0}{4}.$$

Combining equations (49) and (52) gives $e_\Sigma \le 1/2$ and $e_{\mathrm{ref}} + e_v \le \kappa_0/2$. All conditions of Theorem A.1 hold.

Step 3: Collecting the quadratic risk terms. Let

$$q_\star = \max\{\chi v_\rho + 1, \chi c_v + c_\Sigma\}.$$

Then $\chi e_v + e_\Sigma \le q_\star(\rho_{\mathrm{bg}} + r_n + r_N)$. Using $(1 - e_\Sigma)^{-2} \le 4$ and the fact that the square of a sum of three terms is at most three times the sum of their squares, equation (37) gives

$$\begin{aligned} \ell_W^{\mathrm{ERP}}(\widehat{f}) &\le \frac{12 g_+ q_\star^2}{\sqrt{2\pi}}(\rho_{\mathrm{bg}}^2 + r_n^2 + r_N^2) + \frac{c^2 r_N^2}{2\lambda_-\sqrt{2\pi\mathrm{e}}} \\ &\le C_1\left[\rho_{\mathrm{bg}}^2 + \frac{D_t}{n} + \frac{D_t}{N}\right], \end{aligned}$$

where one may take

$$C_1 = \frac{12 g_+ q_\star^2}{\sqrt{2\pi}} + \frac{c^2}{2\lambda_-\sqrt{2\pi\mathrm{e}}}. \tag{54}$$

All thresholds and constants depend only on the uniform bounds and alignment margin in equation (27), not on the true model, dimension, or confidence level. Substituting $D_t = d + \log(16/\alpha)$ gives equation (29). Under exact moment predictions, the mismatch term vanishes. □

Remark (Interpretation of the constants). The thresholds in (50) reserve a fixed margin for sampling errors. The constants therefore do not depend on a model-specific remaining alignment margin that could tend to zero. These explicit constants establish achievability; they are not recommended practical sample sizes.

A.7 Supplementary Decision Theory: Strict Separation and Consequences of Pairwise Bayes Conflict

This section supplements the conceptual distinctions in Section 4 of the main text. Section A.7.1 separates structure, risk, and decisions within one fixed construction; A.7.2 relates common Bayes rules, maximum risk, and restricted classifier classes. These results are not premises of the main proofs in A.2–A.6.

A.7.1 Strict Separation of Latent Structure, Risk Functions, and Optimal Decisions

Let $\mathcal{C} \subseteq \mathfrak{W}_{\mathrm{add}}$ be a nonempty Oracle-compatible set: every pair of models in it has the same Φ_{orc}, but $\mathcal{C}$ need not contain all models with that Oracle information. For a fixed nonempty classifier class $\mathcal{F} \subseteq \mathcal{F}_{\mathrm{all}}$, write

$$r_W^{\mathcal{F}} = \inf_{f\in\mathcal{F}} R_W^{(\mathcal{T})}(f), \qquad \ell_W^{\mathcal{F}}(f) = R_W^{(\mathcal{T})}(f) - r_W^{\mathcal{F}}. \tag{55}$$

Structural identifiability $\mathrm{SI}(\mathcal{C})$ means that $\mathcal{L}_W(\boldsymbol{J}_{1,W}, \boldsymbol{J}_{2,W}, Y)$ is the same throughout $\mathcal{C}$. Risk identifiability $\mathrm{RI}_{\mathcal{F}}(\mathcal{C})$ means that every fixed $f \in \mathcal{F}$ has the same target risk throughout $\mathcal{C}$. Deterministic decision sufficiency $\mathrm{DS}^{\mathrm{det}}_{\varepsilon,\mathcal{F}}(\mathcal{C})$ means that a single $f \in \mathcal{F}$ satisfies $\ell^{\mathcal{F}}_W(f) \le \varepsilon$ for all $W \in \mathcal{C}$. We omit the subscript $\mathcal{F}$ when the classifier class is fixed.

Fixed generative maps send identical latent joint distributions to identical target joint distributions. A common risk function, in turn, permits a common near-optimal classifier. Thus

$$\mathrm{SI}(\mathcal{C}) \Longrightarrow \mathrm{RI}_{\mathcal{F}}(\mathcal{C}) \Longrightarrow \mathrm{DS}^{\mathrm{det}}_{\varepsilon,\mathcal{F}}(\mathcal{C}), \qquad \varepsilon > 0. \tag{56}$$

The second implication requires only an ε-optimal rule for the common infimum, not attainment of the minimum. If the minimum is attained, it also holds for $\varepsilon = 0$. The following single model family disproves both converse implications.

Proposition A.1 (Strict separation under identical Oracle information). Under binary 0-1 loss, there exist a fixed additive generative subclass $\mathfrak{W}_{\mathrm{fix}} \subseteq \mathfrak{W}_{\mathrm{add}}$, fixed component-specific operators, and a nonempty classifier class $\mathcal{F}$, with all models sharing identical Oracle information, and nonempty sets $\mathcal{C}_{\mathrm{str}} \subseteq \mathcal{C}_{\mathrm{dec}} = \mathfrak{W}_{\mathrm{fix}}$ satisfying

$$\begin{aligned} \mathrm{RI}(\mathcal{C}_{\mathrm{str}}) \ \wedge \ \neg\mathrm{SI}(\mathcal{C}_{\mathrm{str}}), \\ \mathrm{DS}^{\mathrm{det}}_{0}(\mathcal{C}_{\mathrm{dec}}) \ \wedge \ \neg\mathrm{RI}(\mathcal{C}_{\mathrm{dec}}). \end{aligned} \tag{57}$$

The same construction also works for $\mathcal{F} = \mathcal{F}_{\mathrm{all}}$.

Proof. The construction varies two parameters independently: r changes only a component allocation invisible to target observations, whereas q changes classification risk while preserving the optimal decision direction. Neither changes the full Oracle information.

Fixed construction and Oracle compatibility. Take $T = 1$, with neural source-activity and observation spaces $\mathbb{R}^2$ in both domains. Fix

$$\begin{aligned} \boldsymbol{G}^{(\mathcal{S})} = \boldsymbol{G}^{(\mathcal{T})} = \boldsymbol{I}_2, \qquad T_2(\boldsymbol{J}) = L_2(\boldsymbol{J}) = \boldsymbol{J}, \\ T_1(\boldsymbol{J}) = L_1(\boldsymbol{J}) = \mathrm{diag}(1,2)\boldsymbol{J}, \qquad D(\boldsymbol{J}) = \mathrm{diag}(0,1)\boldsymbol{J}. \end{aligned}$$

These operators are fixed across models and label-independent. First take the threshold classifier class $\mathcal{F} = \{f_t : t \in \mathbb{R}\}$, where $f_t(\boldsymbol{x}) = +1$ when $x_2 \ge t$, and -1 otherwise.

Fix $a, \sigma, \tau > 0$ and a nondegenerate interval $\mathcal{Q} = [q_-, q_+] \subset (1/2, 1)$. Let $Y \sim \mathrm{Unif}\{-1, +1\}$, $\Lambda \sim \mathrm{Unif}(0,1)$, $\xi \sim \mathcal{N}(0, \sigma^2)$, and $\omega \sim \mathcal{N}(0, \tau^2)$ be mutually independent. For $q \in \mathcal{Q}$, define

$$V_q = 2\mathbf{1}\{\Lambda \le q\} - 1, \qquad B_q = YV_q.$$

Then $P(B_q = Y) = q$ and $P(B_q = -Y) = 1 - q$. By symmetry of Y, the marginal of B_q assigns equal probability to ± 1 for every q.

For $r \in \{0, 1\}$, define the latent activities in model $W_{q,r}$ by

$$\boldsymbol{J}_{1,q,r} = \begin{pmatrix} r\omega \\ a(B_q - Y) \end{pmatrix}, \qquad \boldsymbol{J}_{2,q,r} = \begin{pmatrix} (1-r)\omega \\ a(2Y - B_q) + \xi \end{pmatrix}. \tag{58}$$

Let $\mathfrak{W}_{\mathrm{fix}} = \{W_{q,r} : q \in \mathcal{Q},\ r \in \{0,1\}\}$. The latent variables have finite second moments and satisfy the additive observation structure with fixed operators, so all models belong to $\mathfrak{W}_{\mathrm{add}}$. Direct calculation gives

$$\boldsymbol{X}^{(\mathcal{S})}_{q,r} = \begin{pmatrix} \omega \\ aY + \xi \end{pmatrix}, \qquad \boldsymbol{X}^{(\mathcal{T})}_{q,r} = \begin{pmatrix} \omega \\ aB_q + \xi \end{pmatrix}. \tag{59}$$

The source joint distribution does not depend on q, r. The target marginal is

$$\mathcal{N}(0, \tau^2) \otimes \left[\tfrac{1}{2}\mathcal{N}(-a, \sigma^2) + \tfrac{1}{2}\mathcal{N}(a, \sigma^2)\right],$$

which is also independent of q, r. Together with the fixed priors and operators, this gives the same Φ_{orc} for every model.

Different structures with identical risks. Fix $q_\star \in \mathcal{Q}$ and take $\mathcal{C}_{\mathrm{str}} = \{W_{q_\star,0}, W_{q_\star,1}\}$. By equation (59), changing r preserves the target observations. The coupling uses the same label, so the target joint distributions coincide. Every measurable classifier therefore has the same risk under both models; in particular, $\mathrm{RI}_{\mathcal{F}}(\mathcal{C}_{\mathrm{str}})$ holds.

However, the first latent coordinate changes from

$$([\boldsymbol{J}_{1,q_\star,0}]_1, [\boldsymbol{J}_{2,q_\star,0}]_1) = (0, \omega) \quad \text{to} \quad ([\boldsymbol{J}_{1,q_\star,1}]_1, [\boldsymbol{J}_{2,q_\star,1}]_1) = (\omega, 0).$$

Since ω is nondegenerate, the latent joint laws differ, giving $\neg\mathrm{SI}(\mathcal{C}_{\mathrm{str}})$.

Different risks with a common optimal rule. Take $\mathcal{C}_{\mathrm{dec}} = \mathfrak{W}_{\mathrm{fix}}$. The first coordinate is independent of Y and has the same density in both classes, so it suffices to analyze the second target coordinate z. Its class-conditional densities are

$$p_q(z \mid Y = +1) = q\varphi_\sigma(z - a) + (1 - q)\varphi_\sigma(z + a),$$
$$p_q(z \mid Y = -1) = q\varphi_\sigma(z + a) + (1 - q)\varphi_\sigma(z - a).$$

Their difference is $(2q - 1)[\varphi_\sigma(z - a) - \varphi_\sigma(z + a)]$. Since $q > 1/2$ and $\varphi_\sigma(z - a)/\varphi_\sigma(z + a) = \exp(2az/\sigma^2)$, its sign agrees with z. All models therefore share the Bayes rule $f_0(\boldsymbol{x}) = \mathrm{sign}(x_2)$. The second target coordinate is continuous, so ties do not affect risk. Since $f_0 \in \mathcal{F}$, this proves $\mathrm{DS}^{\mathrm{det}}_{0,\mathcal{F}}(\mathcal{C}_{\mathrm{dec}})$.

On the other hand, $Y[\boldsymbol{X}^{(\mathcal{T})}_{q,r}]_2 = aV_q + Y\xi$, where $Y\xi \sim \mathcal{N}(0, \sigma^2)$ is independent of V_q. Hence

$$\begin{aligned} R^{(\mathcal{T})}_{W_{q,r}}(f_0) &= q\Phi_{\mathrm{N}}(-a/\sigma) + (1 - q)\Phi_{\mathrm{N}}(a/\sigma), \\ R^{(\mathcal{T})}_{W_{q_-,0}}(f_0) - R^{(\mathcal{T})}_{W_{q_+,0}}(f_0) &= (q_+ - q_-)\,[2\Phi_{\mathrm{N}}(a/\sigma) - 1] > 0. \end{aligned} \tag{60}$$

The same rule has different risks under two compatible models, giving $\neg\mathrm{RI}_{\mathcal{F}}(\mathcal{C}_{\mathrm{dec}})$. Both strict separations use the same generative class, operators, loss, and classifier class, with $\mathcal{C}_{\mathrm{str}} \subseteq \mathcal{C}_{\mathrm{dec}}$.

The separation between structure and risk already holds for all measurable classifiers. In the separation between risk and decisions, f_0 is also a Bayes rule, not merely optimal within the threshold class. Both separations therefore remain valid with $\mathcal{F}_{\mathrm{all}}$. □

A.7.2 Common Bayes Rules and Restricted Classifier Classes

Remark A.3 (Zero conflict and maximum excess risk). The measurable optimal choice $f^\dagger$ from A.3 gives

$$C_{\mathrm{conf}}(W, W') = 0 \iff \text{a common measurable classifier attains both Bayes risks.}$$

When the conflict is zero, the two nonnegative excess risks of $f^\dagger$ sum to zero, so both vanish. The converse follows from the definition.

The minimum total excess risk must be distinguished from the minimum maximum excess risk. Let

$$\Gamma^{\mathrm{Bayes}}_{\max}(W, W') = \inf_{f \in \mathcal{F}_{\mathrm{all}}} \max\{\ell^{\mathrm{Bayes}}_W(f), \ell^{\mathrm{Bayes}}_{W'}(f)\}.$$

Nonnegativity, the maximum–average inequality, and attainment by $f^\dagger$ give

$$\tfrac{1}{2}C_{\mathrm{conf}}(W, W') \le \Gamma^{\mathrm{Bayes}}_{\max}(W, W') \le C_{\mathrm{conf}}(W, W'). \tag{61}$$

Neither inequality is an equality in general. Although test-time randomization cannot reduce total conflict, it can reduce maximum risk. The minimax value in A.5 equals $C_{\mathrm{conf}}/2$ because of the Gaussian construction's symmetry and the separately verified attaining rule.

Restricted classifier classes. For any nonempty $\mathcal{F} \subseteq \mathcal{F}_{\mathrm{all}}$, define the approximation error

$$a_V(\mathcal{F}) = r^{\mathcal{F}}_V - r^{\mathrm{Bayes}}_V \ge 0.$$

Equation (55) gives $\ell^{\mathcal{F}}_V(f) = \ell^{\mathrm{Bayes}}_V(f) - a_V(\mathcal{F})$ for $f \in \mathcal{F}$. The total conflict cost within the class therefore satisfies

$$\begin{aligned} \Gamma_\Sigma(W, W'; \mathcal{F}) &:= \inf_{f \in \mathcal{F}} [\ell^{\mathcal{F}}_W(f) + \ell^{\mathcal{F}}_{W'}(f)] \\ &= \inf_{f \in \mathcal{F}} [\ell^{\mathrm{Bayes}}_W(f) + \ell^{\mathrm{Bayes}}_{W'}(f)] - a_W(\mathcal{F}) - a_{W'}(\mathcal{F}) \\ &\ge \big[C_{\mathrm{conf}}(W, W') - a_W(\mathcal{F}) - a_{W'}(\mathcal{F})\big]_+. \end{aligned} \tag{62}$$

The last step also uses $\Gamma_\Sigma \ge 0$, where $[z]_+ = \max\{z, 0\}$. Thus the full-observation Bayes conflict requires an approximation-error correction before it can bound conflict within an arbitrary restricted classifier class. Section A.8.6 uses the same decomposition for ERP risk.

A.8 Generative Constructions, Population Recovery, and Scope of the ERP-Feature Subclass

Section A.6 proves a finite-sample guarantee under feature-distribution and moment conditions. Here we give a trial-level generative construction and sufficient conditions for moment prediction, prove population decision recovery under exact predictions, and construct a nonempty forward-model family retaining attribution ambiguity. We then distinguish population identification based on the full marginal from the finite-sample guarantee for the second-order estimator, and clarify the risk benchmark and mismatch term.

A.8.1 Trial-Level Generative Construction

On a fixed source-activity space $\mathcal{V} \subseteq \mathcal{J}^{(\mathcal{S})}$, write

$$F_{\mathcal{S}}(\boldsymbol{J}) = (\boldsymbol{G}^{(\mathcal{S})}\boldsymbol{J})\boldsymbol{h}, \qquad F_k(\boldsymbol{J}) = L_k(\boldsymbol{J})\boldsymbol{h}.$$

One trial-level construction of the class-mean contrasts in the main text is

$$\begin{aligned} \boldsymbol{J}_{1,W} &= \boldsymbol{C}_{1,W} + \boldsymbol{B}_{1,W} + \frac{Y}{2}(\boldsymbol{H}_W + \boldsymbol{Q}_W), \\ \boldsymbol{J}_{2,W} &= \boldsymbol{C}_{2,W} + \boldsymbol{B}_{2,W} - \frac{Y}{2}\boldsymbol{Q}_W. \end{aligned} \tag{63}$$

Here $\boldsymbol{C}_{1,W}, \boldsymbol{C}_{2,W}, \boldsymbol{H}_W, \boldsymbol{Q}_W \in \mathcal{V}$ are deterministic matrices, and $(\boldsymbol{B}_{1,W}, \boldsymbol{B}_{2,W})$ are zero-mean jointly Gaussian residuals independent of the equiprobable label Y. The latent Gaussian distribution may be degenerate, the two residual components may be correlated, and trials are independent. Varying $\boldsymbol{Q}_W$ while holding the other quantities fixed preserves total source activity, with reallocated activity $\boldsymbol{Z} = Y(\boldsymbol{Q}_{W'} - \boldsymbol{Q}_W)/2$.

Omitting the model subscript, equation (63) yields equation (24), with

$$\begin{aligned} \boldsymbol{\mu}_{\mathcal{S}} &= F_{\mathcal{S}}(\boldsymbol{C}_1 + \boldsymbol{C}_2), & \boldsymbol{\mu}_{\mathcal{T}} &= F_1(\boldsymbol{C}_1) + F_2(\boldsymbol{C}_2), \\ \boldsymbol{\delta}_{\mathcal{S}} &= F_{\mathcal{S}}(\boldsymbol{H}), & \boldsymbol{\delta}_{\mathcal{T}} &= \boldsymbol{b}_\star + (F_1 - F_2)(\boldsymbol{Q}), \\ \boldsymbol{b}_\star &= F_1(\boldsymbol{H}). \end{aligned}$$

The within-class covariances are

$$\boldsymbol{\Sigma}_{\mathcal{S}} = \mathrm{Cov}(F_{\mathcal{S}}(\boldsymbol{B}_1 + \boldsymbol{B}_2)), \qquad \boldsymbol{\Sigma}_{\mathcal{T}} = \mathrm{Cov}(F_1(\boldsymbol{B}_1) + F_2(\boldsymbol{B}_2)). \tag{64}$$

Remark A.4 (Generative construction and statistical assumptions). Equation (63) is one generative construction satisfying the main-text conditions, not an additional latent-Gaussian assumption required by Theorem 2. Section A.6 uses only the class-conditional Gaussian feature distributions in both domains, independent sampling, and moment conditions. The class-mean contrast of the total source activity is $\boldsymbol{H}$, whereas the actual contrast of the first component type is $\boldsymbol{H} + \boldsymbol{Q}$. Thus $\boldsymbol{b}_\star$ is defined by the generative model and cannot be interpreted as an identified contribution of the first component type.

A.8.2 Sufficient Conditions for Exact Covariance and Reference Predictions

Exact moment predictions require

$$\boldsymbol{A}_{\mathrm{sig}}\boldsymbol{\delta}_{\mathcal{S}} = \boldsymbol{b}_\star, \qquad \boldsymbol{A}_{\mathrm{bg}}\boldsymbol{\Sigma}_{\mathcal{S}}\boldsymbol{A}_{\mathrm{bg}}^\top = \boldsymbol{\Sigma}_{\mathcal{T}}.$$

By A.8.1, the following source-space relations suffice:

$$\begin{aligned} F_1(\boldsymbol{H}) &= \boldsymbol{A}_{\mathrm{sig}}F_{\mathcal{S}}(\boldsymbol{H}), \\ F_1(\boldsymbol{B}_1) + F_2(\boldsymbol{B}_2) &= \boldsymbol{A}_{\mathrm{bg}}F_{\mathcal{S}}(\boldsymbol{B}_1 + \boldsymbol{B}_2) \quad \text{almost surely.} \end{aligned} \tag{65}$$

Taking covariances in the second relation yields the desired covariance identity. This sample-wise condition is stronger than covariance equality and is not an assumption of Theorem 2.

To derive these relations from maps on the entire source-activity space, consider $F_k(\boldsymbol{J}) = \boldsymbol{A}_k F_{\mathcal{S}}(\boldsymbol{J})$ for every $\boldsymbol{J} \in \mathcal{V}$. The following null-space inclusion characterizes when this factorization exists.

Lemma A.2 (Feature-space factorization). There exists a matrix $\boldsymbol{A}_k$ satisfying $F_k(\boldsymbol{J}) = \boldsymbol{A}_k F_{\mathcal{S}}(\boldsymbol{J})$ for every $\boldsymbol{J} \in \mathcal{V}$ if and only if $\ker F_{\mathcal{S}} \subseteq \ker F_k$. Under the condition $\boldsymbol{\Sigma}_{\mathcal{S}} \succ 0$ used here, any such $\boldsymbol{A}_k$ is unique and satisfies $\boldsymbol{A}_k \boldsymbol{u} = (F_k \circ F_{\mathcal{S}}^{\dagger})(\boldsymbol{u})$ for every $\boldsymbol{u} \in \mathbb{R}^d$.

Proof. Necessity follows from $F_{\mathcal{S}}(\boldsymbol{J}) = 0 \Rightarrow F_k(\boldsymbol{J}) = 0$. Conversely, define on $\operatorname{range} F_{\mathcal{S}}$

$$\boldsymbol{A}_k F_{\mathcal{S}}(\boldsymbol{J}) = F_k(\boldsymbol{J}).$$

The null-space inclusion makes this definition independent of the chosen preimage; extend it linearly to the full feature space. Moreover, $\boldsymbol{\Sigma}_{\mathcal{S}} \succ 0$ implies that $F_{\mathcal{S}}$ is surjective, since otherwise a nonzero feature direction would have zero variance. The factorization map is therefore uniquely determined on the entire feature space and equals the composition $F_k \circ F_{\mathcal{S}}^{\dagger}$, represented by the matrix $\boldsymbol{A}_k$. □

For example, when $\boldsymbol{B}_1 = 0$, the factorizations $F_1(\boldsymbol{J}) = \boldsymbol{A}_{\mathrm{sig}} F_{\mathcal{S}}(\boldsymbol{J})$ and $F_2(\boldsymbol{J}) = \boldsymbol{A}_{\mathrm{bg}} F_{\mathcal{S}}(\boldsymbol{J})$ for every $\boldsymbol{J} \in \mathcal{V}$ imply equation (65). However, Theorem 2 requires only the corresponding moment-prediction conditions, not these full-space factorizations.

Remark A.5 (Limits of known component operators). For general correlated residuals $(\boldsymbol{B}_1, \boldsymbol{B}_2)$, even both full-space factorizations above do not imply the covariance prediction. The first residual component is transformed by $\boldsymbol{A}_{\mathrm{sig}}$, rather than $\boldsymbol{A}_{\mathrm{bg}}$. The target covariance also depends on how residual covariances and cross-covariances are allocated. Known component-specific operators therefore do not replace the moment-prediction conditions in equation (26).

Converting operator error into moment-prediction error also requires control of source-signal scale. If $E_1(\boldsymbol{J}) := F_1(\boldsymbol{J}) - \boldsymbol{A}_{\mathrm{sig}} F_{\mathcal{S}}(\boldsymbol{J})$, then $\varepsilon_{\mathrm{sig}} \le \|E_1\|_{\mathrm{op}} \|\boldsymbol{H}\|_{\mathrm{F}}$. Observation covariance bounds do not control all energy in the null space of the source observation map.

A.8.3 Population Recovery under Exact Covariance and Reference Predictions

Proposition A.2 (Population decision recovery with attribution ambiguity). Oracle-compatible models in $\mathfrak{W}_{\mathrm{ERP}}^0$ using the same features and side-information maps have identical target feature joint distributions $P_{\mathcal{T}}^{UY}$. Hence any nonempty compatible set satisfies RI and $\mathrm{DS}_0^{\mathrm{det}}$ on $\mathcal{F}_{\mathrm{ERP}}$. Structural identifiability can still fail if the generative construction permits a nonzero reallocation direction $\boldsymbol{K}$ satisfying $(L_1 - L_2)(\boldsymbol{K}) = 0$.

Proof. Identical source feature joint distributions and target feature marginals determine the same $\boldsymbol{\delta}_{\mathcal{S}}, \boldsymbol{\Sigma}_{\mathcal{S}}, \boldsymbol{\mu}_{\mathcal{T}}, \boldsymbol{K}_{\mathcal{T}}$. Exact moment predictions further determine $\boldsymbol{b}_\star$ and $\boldsymbol{\Sigma}_{\mathcal{T}}$. Let $\boldsymbol{M} = \boldsymbol{K}_{\mathcal{T}} - \boldsymbol{\Sigma}_{\mathcal{T}}$. Then

$$\boldsymbol{M} = \tfrac{1}{4} \boldsymbol{\delta}_{\mathcal{T}} \boldsymbol{\delta}_{\mathcal{T}}^{\top}, \qquad \boldsymbol{\delta}_{\mathcal{T}} = 2\sqrt{\lambda_1(\boldsymbol{M})} \frac{\boldsymbol{P}_1(\boldsymbol{M}) \boldsymbol{b}_\star}{\|\boldsymbol{P}_1(\boldsymbol{M}) \boldsymbol{b}_\star\|_2}. \tag{66}$$

Here $\lambda_1(\boldsymbol{M}) > 0$ is the unique positive eigenvalue of the nonzero rank-one matrix $\boldsymbol{M}$, and $\boldsymbol{P}_1(\boldsymbol{M})$ is its orthogonal projector. The alignment condition ensures $\boldsymbol{P}_1(\boldsymbol{M}) \boldsymbol{b}_\star \neq 0$ and selects the correct sign. Thus $\boldsymbol{\mu}_{\mathcal{T}}, \boldsymbol{\delta}_{\mathcal{T}}, \boldsymbol{\Sigma}_{\mathcal{T}}$ and $P_{\mathcal{T}}^{UY}$ are identical across models, with common optimal rule

$$f_{\mathrm{ERP}}^{\star}(\boldsymbol{X}) = \operatorname{sign}\left[\boldsymbol{\delta}_{\mathcal{T}}^{\top} \boldsymbol{\Sigma}_{\mathcal{T}}^{-1} (\Psi(\boldsymbol{X}) - \boldsymbol{\mu}_{\mathcal{T}})\right].$$

If equation (63) permits a nonzero $\boldsymbol{K}$ satisfying $(L_1 - L_2)(\boldsymbol{K}) = 0$, replace $\boldsymbol{Q}$ by $\boldsymbol{Q} + \boldsymbol{K}$. Then $\boldsymbol{J}_1' = \boldsymbol{J}_1 + Y\boldsymbol{K}/2$ and $\boldsymbol{J}_2' = \boldsymbol{J}_2 - Y\boldsymbol{K}/2$, preserving total source activity and full target observations sample-wise. The complete Oracle information and all moment conditions therefore remain unchanged. However,

$$\mathbb{E}[\boldsymbol{J}_1' \mid Y = +1] - \mathbb{E}[\boldsymbol{J}_1 \mid Y = +1] = \boldsymbol{K}/2 \neq 0,$$

so the latent joint distributions differ. Structural ambiguity can therefore remain even when risks and decisions are identifiable. Section A.8.4 gives a nonempty example satisfying these conditions. □

Remark A.6 (Relation to Proposition 1). Temporarily denote the positive scalar Δ in A.2 by a. Taking $\boldsymbol{H} = -a$, $\boldsymbol{B}_1 = 0$, $\boldsymbol{B}_2 = \xi$, and $\boldsymbol{C}_1 = \boldsymbol{C}_2 = 0$, with $\boldsymbol{Q}_{W_0} = 4a$ and $\boldsymbol{Q}_{W_Z} = 0$, equation (63) reproduces the construction of Proposition 1. Both models have

$$\boldsymbol{A}_{\mathrm{sig}} = 2, \qquad \boldsymbol{A}_{\mathrm{bg}} = 1, \qquad \boldsymbol{\delta}_{\mathcal{S}} = -a, \qquad \boldsymbol{b}_\star = -2a,$$

and both moment-prediction errors vanish. However, the target contrasts are $2a$ and $-2a$, with cosines -1 and 1 relative to the common reference. Assumption 2 therefore excludes W_0. The additional structural restriction excludes the conflicting pair; the shared observations do not identify the true model on their own.

A.8.4 A NONEMPTY FORWARD-MODEL FAMILY WITH REMAINING ATTRIBUTION AMBIGUITY

Example A.1 (Forward construction with structural ambiguity). We construct a model family satisfying the assumptions of Theorem 2. The parameter ζ changes the target decision and observable marginal, whereas r changes attribution along a null direction while preserving the full Oracle information.

Take $d \geq 2$, a full-row-rank matrix $\boldsymbol{G} \in \mathbb{R}^{d\times p}$, and a right inverse $\boldsymbol{R}_G$ satisfying $\boldsymbol{G}\boldsymbol{R}_G = \boldsymbol{I}_d$. Fix the shared lead field to $\boldsymbol{G}$, and let

$$\begin{aligned} \boldsymbol{A}_{\mathrm{bg}} = \boldsymbol{I}_d, \quad & \boldsymbol{A}_{\mathrm{sig}} = \mathrm{diag}(1,2,1,\ldots,1), \\ T_2(\boldsymbol{J}) = \boldsymbol{J}, \quad & T_1(\boldsymbol{J}) = \boldsymbol{J} + \boldsymbol{R}_G(\boldsymbol{A}_{\mathrm{sig}} - \boldsymbol{I}_d)\boldsymbol{G}\boldsymbol{J}. \end{aligned}$$

Then $\boldsymbol{G}T_1(\boldsymbol{J}) = \boldsymbol{A}_{\mathrm{sig}}\boldsymbol{G}\boldsymbol{J}$ and $\boldsymbol{G}T_2(\boldsymbol{J}) = \boldsymbol{G}\boldsymbol{J}$ for every $\boldsymbol{J}$, with fixed, label-independent component-specific operators. For a fixed unit-norm template $\boldsymbol{h}$, set the quantities in equation (63) as follows:

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{R}_G(\boldsymbol{e}_1 + \boldsymbol{e}_2)\boldsymbol{h}^\top, \\ \boldsymbol{Q}_{\zeta,r} &= \zeta\boldsymbol{R}_G\boldsymbol{e}_2\boldsymbol{h}^\top + r\boldsymbol{R}_G\boldsymbol{e}_1\boldsymbol{h}^\top, \quad \zeta \in [0,1],\ r \in \{0,1\}, \\ \boldsymbol{B}_1 &= \tfrac{1}{2}\boldsymbol{R}_G\boldsymbol{e}_1\xi_1\boldsymbol{h}^\top, \qquad \boldsymbol{B}_2 = \boldsymbol{R}_G\boldsymbol{\xi}\boldsymbol{h}^\top - \boldsymbol{B}_1, \\ \boldsymbol{\xi} &\sim \mathcal{N}(0, \boldsymbol{I}_d), \qquad \boldsymbol{C}_1 = \boldsymbol{C}_2 = 0. \end{aligned} \tag{67}$$

Here ξ_1 is the first coordinate of $\boldsymbol{\xi}$, and $\boldsymbol{\xi}$ is independent of Y. Both residual components are nonzero and correlated. Since $(\boldsymbol{A}_{\mathrm{sig}} - \boldsymbol{I}_d)\boldsymbol{e}_1 = 0$,

$$L_1(\boldsymbol{B}_1) + L_2(\boldsymbol{B}_2) = \boldsymbol{G}(\boldsymbol{B}_1 + \boldsymbol{B}_2) = \boldsymbol{\xi}\boldsymbol{h}^\top.$$

It follows that

$$\begin{aligned} \boldsymbol{\delta}_{\mathcal{S}} &= \boldsymbol{e}_1 + \boldsymbol{e}_2, & \boldsymbol{b}_\star &= \boldsymbol{e}_1 + 2\boldsymbol{e}_2, \\ \boldsymbol{\delta}_{\mathcal{T},\zeta} &= \boldsymbol{e}_1 + (2+\zeta)\boldsymbol{e}_2, & \boldsymbol{\Sigma}_{\mathcal{S}} &= \boldsymbol{\Sigma}_{\mathcal{T}} = \boldsymbol{I}_d. \end{aligned}$$

Both moment-prediction errors vanish. The alignment cosine satisfies

$$\frac{5+2\zeta}{\sqrt{5}\sqrt{1+(2+\zeta)^2}} \geq \frac{1}{\sqrt{2}},$$

so a uniform choice is

$$\lambda_- = \lambda_+ = 1, \quad \Delta_0 = \sqrt{5}, \quad \Delta_+ = \sqrt{10}, \quad b_0 = \sqrt{5}, \quad L_{\mathrm{sig}} = 2, \quad \kappa_0 = 1/\sqrt{2}.$$

Varying ζ preserves the source distribution but changes the target Bayes normal and $\boldsymbol{K}_{\mathcal{T},\zeta} = \boldsymbol{I}_d + \boldsymbol{\delta}_{\mathcal{T},\zeta}\boldsymbol{\delta}_{\mathcal{T},\zeta}^\top/4$; these models do not share the full Oracle information. Fixing ζ and varying r instead changes attribution along the nonzero direction $\boldsymbol{K} = \boldsymbol{R}_G\boldsymbol{e}_1\boldsymbol{h}^\top$, with $(L_1 - L_2)(\boldsymbol{K}) = 0$. The complete Oracle information and target joint distribution then remain unchanged, whereas the latent joint distribution changes. The same two-model set satisfies $\neg\mathrm{SI}$, $\mathrm{RI}_{\mathcal{F}_{\mathrm{ERP}}}$, and $\mathrm{DS}_0^{\mathrm{det}}$, and Theorem 2 applies to both models.

Remark (Nonemptiness and empirical scope). This construction establishes that the subclass is nonempty and permits structural ambiguity. It does not assert that general real EEG satisfies these conditions. The right inverse is used only to construct the example, not as input to the learner.

A.8.5 HIGHER-ORDER MOMENTS AND POPULATION IDENTIFIABILITY

Theorem 2 gives finite-sample sufficient conditions for a second-order estimator. Exact covariance prediction is not necessary to recover the mean contrast up to sign from the full target marginal. Fourth moments show this directly in the equal-prior Gaussian model with common within-class covariance in A.6.1.

For any $\boldsymbol{q} \in \mathbb{R}^d$, let $V_q = \boldsymbol{q}^\top(\boldsymbol{U}^{(\mathcal{T})} - \boldsymbol{\mu}_{\mathcal{T}})$, $s_q^2 = \boldsymbol{q}^\top\boldsymbol{\Sigma}_{\mathcal{T}}\boldsymbol{q}$, and $a_q = \boldsymbol{q}^\top\boldsymbol{\delta}_{\mathcal{T}}/2$. The symmetric Gaussian mixture gives

$$\mathbb{E}V_q^2 = s_q^2 + a_q^2, \qquad \mathbb{E}V_q^4 = 3s_q^4 + 6s_q^2a_q^2 + a_q^4.$$

These identities imply

$$\mathrm{cum}_4(V_q) = \mathbb{E}V_q^4 - 3(\mathbb{E}V_q^2)^2 = -\tfrac{1}{8}(\boldsymbol{q}^\top \boldsymbol{\delta}_{\mathcal{T}})^4. \tag{68}$$

The full target marginal determines every $(\boldsymbol{q}^\top \boldsymbol{\delta}_{\mathcal{T}})^2$, hence the quadratic form $\boldsymbol{\delta}_{\mathcal{T}}\boldsymbol{\delta}_{\mathcal{T}}^\top$, and then $\boldsymbol{\Sigma}_{\mathcal{T}}$ through equation (25). The remaining sign must still be matched to labels with fixed semantics. Under the reference-norm and alignment conditions of Assumption 2, if $\varepsilon_{\mathrm{sig}} < \kappa_0 b_0$, then

$$(\boldsymbol{A}_{\mathrm{sig}}\boldsymbol{\delta}_{\mathcal{S}})^\top \boldsymbol{\delta}_{\mathcal{T}} \geq (\kappa_0 \|\boldsymbol{b}_\star\|_2 - \varepsilon_{\mathrm{sig}})\|\boldsymbol{\delta}_{\mathcal{T}}\|_2 > 0,$$

so the nominal reference selects the correct sign. Equation (49) already ensures this condition.

This population result agrees with parameter identifiability of finite Gaussian mixtures (Yakowitz & Spragins, 1968). Thus ρ_{bg} measures prediction mismatch for the second-order spectral estimator in A.6; nonzero mismatch does not imply population non-identifiability in this subclass. The calculation above gives no finite-sample rate for a higher-order moment estimator.

A.8.6 Risk Benchmarks and Interpretation of the Finite-Sample Bound

Feature risk and full-observation risk. Let $a_{\mathrm{ERP}}(W) = r_W^{\mathrm{ERP}} - r_W^{\mathrm{Bayes}} \geq 0$. Every branch of the algorithm in A.6.2 outputs a classifier in $\mathcal{F}_{\mathrm{ERP}}$, so

$$\ell_W^{\mathrm{Bayes}}(\widehat{f}) = a_{\mathrm{ERP}}(W) + \ell_W^{\mathrm{ERP}}(\widehat{f}). \tag{69}$$

A bound on full-EEG Bayes excess risk must include this additional approximation error. If $Y \perp\!\!\!\perp \boldsymbol{X}^{(\mathcal{T})} \mid \Psi(\boldsymbol{X}^{(\mathcal{T})})$, the posterior depends only on the fixed features and $a_{\mathrm{ERP}}(W) = 0$. For example,

$$\boldsymbol{X}^{(\mathcal{T})} = \boldsymbol{U}^{(\mathcal{T})}\boldsymbol{h}^\top + \boldsymbol{R}, \qquad \boldsymbol{R}\boldsymbol{h} = 0, \qquad \boldsymbol{R} \perp\!\!\!\perp (\boldsymbol{U}^{(\mathcal{T})}, Y)$$

is sufficient. Requiring only that the class-mean contrast lie along $\boldsymbol{h}$ is insufficient, since the remaining observations may still carry label information.

High-probability guarantees and expected risk. Let $B_{n,N,\alpha}(W)$ denote the deterministic right-hand side of the high-probability bound. Since $0 \leq \ell_W^{\mathrm{ERP}} \leq 1$,

$$\mathbb{E}_W \ell_W^{\mathrm{ERP}}(\widehat{f}) \leq B_{n,N,\alpha}(W) + \alpha. \tag{70}$$

This connects to the expected-risk formulation in Section 4 of the main text while retaining the different model classes and information conditions of the two sections.

The mismatch term is not a risk floor. Take $\boldsymbol{\Sigma}_{\mathcal{T}} = \boldsymbol{I}_d$, $\boldsymbol{\delta}_{\mathcal{T}} = \boldsymbol{e}_1$, and $\boldsymbol{\Sigma}_{\mathrm{nom}} = (1-r)\boldsymbol{I}_d$, with sufficiently small $r > 0$ and a correctly oriented reference. In the infinite-sample limit,

$$\boldsymbol{K}_{\mathcal{T}} - \boldsymbol{\Sigma}_{\mathrm{nom}} = \tfrac{1}{4}\boldsymbol{e}_1\boldsymbol{e}_1^\top + r\boldsymbol{I}_d, \qquad \boldsymbol{\Sigma}_{\mathrm{nom}}^{-1}\boldsymbol{e}_1 = \boldsymbol{e}_1/(1-r).$$

The optimal direction and decision are recovered despite $\rho_{\mathrm{bg}} = r > 0$. For $T = 1$, this example is realized by shared identity lead fields, $T_1 = T_2 = I$, $\boldsymbol{H} = \boldsymbol{e}_1$, $\boldsymbol{Q} = 0$, $\boldsymbol{B}_1 = 0$, $\boldsymbol{B}_2 \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $\boldsymbol{A}_{\mathrm{sig}} = \boldsymbol{I}_d$, and $\boldsymbol{A}_{\mathrm{bg}} = \sqrt{1-r}\boldsymbol{I}_d$. The mismatch term in equation (29) is therefore a stability term in an upper bound, not an unavoidable risk floor.

A.9 Auxiliary Proofs

A.9.1 Proof of Lemma A.1

Proof. We control source class counts, the source mean contrast and within-class covariance, and the target mean and marginal covariance, then intersect their success events. All error bounds use t, a_ν, c_ν from A.6.3 and equation (34).

Step 1: Gaussian vector and covariance bounds. The χ^2 tail bounds of Laurent & Massart (2000) give, for $Q_\nu \sim \chi^2_\nu$ and $s > 0$,

$$P(Q_\nu - \nu > 2\sqrt{\nu s} + 2s) \leq e^{-s},$$
$$P(\nu - Q_\nu > 2\sqrt{\nu s}) \leq e^{-s}.$$

Consequently, for $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_d)$,

$$P(\|\boldsymbol{z}\|_2 > \sqrt{d} + \sqrt{2t}) \leq e^{-t}. \tag{71}$$

For independent standard Gaussian vectors $\boldsymbol{z}_1, \dots, \boldsymbol{z}_\nu$, take a spherical $1/4$-net $\mathcal{N}$ of size at most 9^d. Symmetric matrices satisfy $\|\boldsymbol{S}\|_{\mathrm{op}} \le 2\max_{q\in\mathcal{N}} |\boldsymbol{q}^\top \boldsymbol{S}\boldsymbol{q}|$. Apply the tail bounds at each net point, set $s = d\log 9 + t$, and take a union bound to obtain

$$P\left(\left\|\frac{1}{\nu}\sum_{i=1}^{\nu} \boldsymbol{z}_i\boldsymbol{z}_i^\top - \boldsymbol{I}_d\right\|_{\mathrm{op}} > c_\nu\right) \le 2e^{-t}. \tag{72}$$

See Vershynin (2012) for the spherical-net bound used here.

Step 2: Source class counts and mean contrast. For equiprobable source labels, Hoeffding's inequality (Hoeffding, 1963) gives

$$P\big(\min(n_{+1}, n_{-1}) < n/4\big) \le 2e^{-n/8}.$$

On the class-count success event, conditional on all source labels,

$$\widehat{\boldsymbol{\delta}}_{\mathcal{S}} - \boldsymbol{\delta}_{\mathcal{S}} \sim \mathcal{N}\big(0, (n_{+1}^{-1} + n_{-1}^{-1})\boldsymbol{\Sigma}_{\mathcal{S}}\big), \qquad (n_{+1}^{-1} + n_{-1}^{-1})\boldsymbol{\Sigma}_{\mathcal{S}} \preceq \frac{8}{n}\boldsymbol{\Sigma}_{\mathcal{S}}.$$

Equation (71) gives, with conditional failure probability at most e^{-t},

$$\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}} - \boldsymbol{\delta}_{\mathcal{S}}\|_2 \le \sqrt{8\lambda_+}\, a_n = \epsilon_\delta.$$

Step 3: Pooled source within-class covariance. Conditional on source labels, removing the two class means corresponds to an orthogonal projector P of rank $n-2$. Let $\boldsymbol{Z}_{\mathcal{S}}$ be the $n \times d$ whitened residual matrix with independent standard normal entries. Then

$$(n-2)\boldsymbol{\Sigma}_{\mathcal{S}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}}^{-1/2} = \boldsymbol{Z}_{\mathcal{S}}^\top \boldsymbol{P}\boldsymbol{Z}_{\mathcal{S}}. \tag{73}$$

Gaussian orthogonal invariance expresses the right-hand side as the sum of outer products of $n-2$ independent standard Gaussian vectors. Equation (72) gives, with conditional failure probability at most $2e^{-t}$,

$$-\eta_n\boldsymbol{\Sigma}_{\mathcal{S}} \preceq \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}} - \boldsymbol{\Sigma}_{\mathcal{S}} \preceq \eta_n\boldsymbol{\Sigma}_{\mathcal{S}}, \qquad \eta_n = c_{n-2}.$$

This argument uses only Gaussianity of the source feature residuals.

Step 4: Mean and covariance of the target Gaussian mixture. The target marginal is a Gaussian mixture, so a single-Gaussian covariance bound does not apply directly. For the analysis only, write

$$\boldsymbol{U}_j^{(\mathcal{T})} - \boldsymbol{\mu}_{\mathcal{T}} = \boldsymbol{\xi}_j + Y_j\boldsymbol{\delta}_{\mathcal{T}}/2, \qquad \boldsymbol{\xi}_j \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\mathcal{T}}),$$

where $\boldsymbol{\xi}_j$ is independent of the equiprobable label Y_j, and the pairs $(\boldsymbol{\xi}_j, Y_j)$ are i.i.d. Then

$$\widehat{\boldsymbol{\mu}}_{\mathcal{T}} - \boldsymbol{\mu}_{\mathcal{T}} = \overline{\boldsymbol{\xi}} + \overline{Y}\boldsymbol{\delta}_{\mathcal{T}}/2.$$

The Gaussian vector bound and Hoeffding's inequality (Hoeffding, 1963) give $\|\overline{\boldsymbol{\xi}}\|_2 \le \sqrt{\lambda_+}a_N$ and $|\overline{Y}| \le \sqrt{2t/N}$, respectively, with total failure probability at most $3e^{-t}$. Hence

$$\|\widehat{\boldsymbol{\mu}}_{\mathcal{T}} - \boldsymbol{\mu}_{\mathcal{T}}\|_2 \le \sqrt{\lambda_+}a_N + \frac{\Delta_+}{2}\sqrt{\frac{2t}{N}} = \epsilon_\mu.$$

Let $\boldsymbol{w}_N = N^{-1}\sum_j Y_j\boldsymbol{\xi}_j$. Conditional on target labels, $\boldsymbol{w}_N \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\mathcal{T}}/N)$, so $\|\boldsymbol{w}_N\|_2 \le \sqrt{\lambda_+}a_N$ fails with probability at most e^{-t}. Since $Y_j^2 = 1$, the centered covariance has the exact decomposition

$$\begin{aligned}\widehat{\boldsymbol{K}}_{\mathcal{T}} - \boldsymbol{K}_{\mathcal{T}} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\xi}_j\boldsymbol{\xi}_j^\top - \boldsymbol{\Sigma}_{\mathcal{T}} \\ + \tfrac{1}{2}(\boldsymbol{\delta}_{\mathcal{T}}\boldsymbol{w}_N^\top + \boldsymbol{w}_N\boldsymbol{\delta}_{\mathcal{T}}^\top) \\ - (\widehat{\boldsymbol{\mu}}_{\mathcal{T}} - \boldsymbol{\mu}_{\mathcal{T}})(\widehat{\boldsymbol{\mu}}_{\mathcal{T}} - \boldsymbol{\mu}_{\mathcal{T}})^\top.\end{aligned} \tag{74}$$

Equation (72) bounds the norm of the first term by $\lambda_+ c_N$, with an additional failure probability at most $2e^{-t}$. The second term has norm at most $\Delta_+\sqrt{\lambda_+}a_N$, and the third at most ϵ_μ^2. Adding them gives

$$\|\widehat{\boldsymbol{K}}_{\mathcal{T}} - \boldsymbol{K}_{\mathcal{T}}\|_{\mathrm{op}} \le \lambda_+ c_N + \Delta_+\sqrt{\lambda_+}a_N + \epsilon_\mu^2 = \epsilon_K.$$

These latent target labels are used only in the decomposition and probability analysis; the algorithm does not access them.

Step 5: A joint success event. Let $\mathcal{E}$ be the intersection of all success events above. The total failure probability is at most

$$\begin{aligned} P_W(\mathcal{E}^c) &\le 2e^{-n/8} + (1+2+3+1+2)e^{-t} \\ &\le 11e^{-t} < \alpha, \end{aligned} \tag{75}$$

where we used $n \ge 8t$ and $t = \log(16/\alpha)$. Since $n \ge 8$, the class-count success event also ensures at least two samples per class. All conclusions of equation (38) therefore hold simultaneously on $\mathcal{E}$. The union bound does not require the estimators to be mutually independent. □

A.9.2 Posterior and Density Versions and Measurable Selection

For the finite label space, fix measurable regular conditional versions of the target posteriors and nonnegative measurable versions of the Radon–Nikodym derivatives. Minima and maxima of finitely many measurable functions are measurable, as is the pointwise selection that breaks ties by the smallest class index. Thus $f^\dagger$ in A.3 is measurable and attains the pointwise minimum.

The conclusions of A.3 are invariant to these version choices. If two posterior versions differ only on a $P^X_{W,\mathcal{T}}$-null set N_W, then

$$0 = P^X_{W,\mathcal{T}}(N_W) = \int_{N_W} p_W \, d\mu,$$

so $p_W = 0$ holds μ-almost everywhere on N_W. Such a change leaves the weighted integral in equation (12) unchanged and can alter the conflict region, which requires $p_W p_{W'} > 0$, only on a μ-null set. The same holds for W'. Changing the Radon–Nikodym versions of the densities likewise leaves the conclusions unchanged.

Computational environment. All experiments were run on a workstation equipped with an Intel Core i9-12900K CPU and an NVIDIA GeForce RTX 3070 Ti GPU, using Python 3.10.

B Experiment 1: Controlled EEG Simulation of Attribution Ambiguity

B.1 Experimental Overview

Can identical source scalp EEG correspond to different latent-component attributions such that the two candidate models' target-trained classifiers cannot be exchanged without increasing error? Experiment 1 constructs this comparison under a shared linear EEG forward model. Within each condition, the candidates share the same cross-domain transformation operators and differ only in the latent allocation of class-related activity. Their source observations remain identical trial-wise.

Across conditions, ζ controls the attribution difference, whereas ρ controls the component-operator difference through the relative rotation angle. Experiment 1A places class structure in ERP class-conditional means; Experiment 1B places it in the spatial bandpower of class-conditionally zero-mean signals. Both use the same reallocation path and spatial source configurations (Figure 4).

For each condition, each candidate model generates a labeled target training set and an independent test set. The fitted classifiers measure rule exchange; they are not UDA learners. Applying both classifiers to both test sets compares the error of each model's own rule with that of the exchanged rule. The bidirectional exchange loss α_{struct} is the smaller balanced-error increase across the two exchange directions and is positive only if both exchanges increase error. The primary condition is $(\zeta, \rho) = (1, 0.75)$. The full parameter grid and controls examine parameter dependence, rotation allocation, classifier features, and the association between reallocated structure and labels.

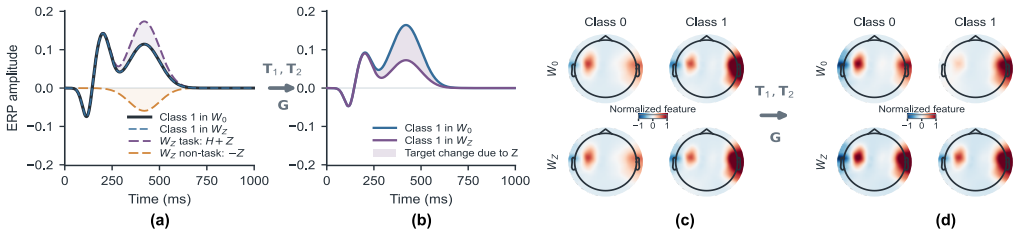

Figure 4: Two EEG statistical instances of the same attribution mechanism. (a,b) Reallocation in ERP class-conditional means changes the target class-contrast direction. (c,d) The same relation changes spatial bandpower allocation in the class-conditionally zero-mean instance. The panels show noise-free population quantities for one source configuration to illustrate the construction, rather than aggregate classification results. Both instances are evaluated on the same 20 configurations; their exact definitions are in B.2.

Source configurations, calibrated parameters, and the primary condition are determined before risk evaluation. Sections B.2–B.3 describe construction and sampling; B.4 reports results.

B.2 Source-Equivalent Construction and Target Transformations

B.2.1 Source Equivalence and Component-Specific Transformations

Reallocation enters the two component activities with opposite signs and cancels in their source-domain sum. Different component-specific target transformations can prevent this cancellation in the target domain.

Both domains use the same lead field $\boldsymbol{G} \in \mathbb{R}^{d \times p}$, so $\boldsymbol{G}^{(\mathcal{S})} = \boldsymbol{G}^{(\mathcal{T})} = \boldsymbol{G}$. Here p is the source dimension and d the number of sensors. Let $\boldsymbol{J}_1$ carry the base task structure, $\boldsymbol{J}_2$ the remaining modeled source activity and background variation, and $\boldsymbol{Z}$ the reallocated class-related activity. Along

Table 1: Experimental units and comparison levels in Experiment 1. Source configurations are the units of aggregation and resampling. Conditions, candidate models, and paired observations serve different comparisons within each configuration.

Level	Setting	Role
Signal instance	ERP means (1A); spatial bandpower (1B)	Estimated separately; both must meet the primary criterion
Source configuration	20 pairs of source regions and their projections under a shared lead field	Shared by both instances; units for aggregation and bootstrap resampling
Condition	52 per instance and configuration	Parameter grid and mechanism controls; results in Table 4
Candidate model	W_0 and $W_{\zeta Z}$	Share base random draws to compare target consequences of different attributions
Finite samples	Independent training/test draws and paired sampling	Fit and evaluate classifiers while controlling sampling variation in comparisons

$\zeta \in [0, 1]$, define

$$\begin{aligned}(\boldsymbol{J}_{1,W_0}, \boldsymbol{J}_{2,W_0}) &= (\boldsymbol{J}_1, \boldsymbol{J}_2),\\ (\boldsymbol{J}_{1,W_{\zeta Z}}, \boldsymbol{J}_{2,W_{\zeta Z}}) &= (\boldsymbol{J}_1 + \zeta \boldsymbol{Z}, \boldsymbol{J}_2 - \zeta \boldsymbol{Z}),\\ \boldsymbol{X}^{(\mathcal{S})}_{W_{\zeta Z}} &= \boldsymbol{G}(\boldsymbol{J}_1 + \zeta \boldsymbol{Z} + \boldsymbol{J}_2 - \zeta \boldsymbol{Z}) = \boldsymbol{X}^{(\mathcal{S})}_{W_0}.\end{aligned} \tag{76}$$

The value $\zeta = 0$ gives W_0, and $\zeta = 1$ gives W_Z. The two candidates use the same trial label and base random draw, preserving source observations and their label associations along the entire path.

The target transformations for the two component types are $T_1(\rho)$ and $T_2(\rho)$. Write $[L_k(\rho)](\boldsymbol{J}) = \boldsymbol{G}[T_k(\rho)](\boldsymbol{J})$ and $D(\rho) = L_1(\rho) - L_2(\rho)$. Each between-model pair shares measurement noise $\boldsymbol{\eta}^{(\mathcal{T})}_\rho$, giving

$$\begin{aligned}\boldsymbol{X}^{(\mathcal{T})}_{W_0} &= [L_1(\rho)](\boldsymbol{J}_1) + [L_2(\rho)](\boldsymbol{J}_2) + \boldsymbol{\eta}^{(\mathcal{T})}_\rho,\\ \boldsymbol{X}^{(\mathcal{T})}_{W_{\zeta Z}} &= [L_1(\rho)](\boldsymbol{J}_1 + \zeta \boldsymbol{Z}) + [L_2(\rho)](\boldsymbol{J}_2 - \zeta \boldsymbol{Z}) + \boldsymbol{\eta}^{(\mathcal{T})}_\rho,\\ \boldsymbol{X}^{(\mathcal{T})}_{W_{\zeta Z}} - \boldsymbol{X}^{(\mathcal{T})}_{W_0} &= \zeta [D(\rho)](\boldsymbol{Z}).\end{aligned} \tag{77}$$

Thus ζ controls the amount of reallocation, ρ controls the component-operator difference $D(\rho)$, and $\boldsymbol{Z}$ specifies the reallocated signal structure. Each controlled elementary component has the form $\boldsymbol{J} = \boldsymbol{u}\boldsymbol{h}^\top$, with spatial loading $\boldsymbol{u} \in \mathbb{R}^p$ and temporal process $\boldsymbol{h} \in \mathbb{R}^T$. A source configuration consists of two source regions and their lead-field projections, providing loading directions $\boldsymbol{u}_1, \boldsymbol{u}_2$. The base task structure and $\boldsymbol{Z}$ are constructed in their span.

Spatial operators. Equation (77) specifies the difference between the candidates. Opposing rotations in a shared sensor plane implement $D(\rho)$, leaving directions outside that plane unchanged. The following construction makes these rotations correspond exactly to source-space transformations under the shared lead field.

The first source region gives a loading $\boldsymbol{u}_1$, scaled so that $\boldsymbol{q}_1 := \boldsymbol{G}\boldsymbol{u}_1$ and $\|\boldsymbol{q}_1\|_2 = 1$. The second gives $\tilde{\boldsymbol{u}}_2$, whose lead-field projection $\tilde{\boldsymbol{q}}_2 := \boldsymbol{G}\tilde{\boldsymbol{u}}_2$ is also normalized first. Apply the same orthogonalization and scaling in source and sensor space:

$$\begin{aligned}\boldsymbol{r}_2 &= \tilde{\boldsymbol{q}}_2 - (\boldsymbol{q}_1^\top \tilde{\boldsymbol{q}}_2)\boldsymbol{q}_1,\\ \boldsymbol{u}_2 &= \frac{\tilde{\boldsymbol{u}}_2 - (\boldsymbol{q}_1^\top \tilde{\boldsymbol{q}}_2)\boldsymbol{u}_1}{\|\boldsymbol{r}_2\|_2}, \qquad \boldsymbol{q}_2 = \frac{\boldsymbol{r}_2}{\|\boldsymbol{r}_2\|_2} = \boldsymbol{G}\boldsymbol{u}_2.\end{aligned} \tag{78}$$

Direction signs follow a fixed rule. Then $\boldsymbol{G}\boldsymbol{u}_i = \boldsymbol{q}_i$, with $\boldsymbol{q}_1, \boldsymbol{q}_2$ orthonormal. Write $\boldsymbol{H} = [\boldsymbol{u}_1, \boldsymbol{u}_2]$ and $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2] = \boldsymbol{G}\boldsymbol{H}$. For rotation angle θ, define

$$\begin{aligned} \boldsymbol{R}_2(\theta) &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \\ \boldsymbol{R}(\theta) &= \boldsymbol{I}_d + \boldsymbol{Q}\big(\boldsymbol{R}_2(\theta) - \boldsymbol{I}_2\big)\boldsymbol{Q}^\top, \\ \mathbf{T}(\theta) &= \boldsymbol{I}_p + \boldsymbol{H}\big(\boldsymbol{R}_2(\theta) - \boldsymbol{I}_2\big)\boldsymbol{Q}^\top\boldsymbol{G}, \\ \boldsymbol{G}\mathbf{T}(\theta) &= \boldsymbol{R}(\theta)\boldsymbol{G}. \end{aligned} \tag{79}$$

The identity ensures that transforming source activity before applying the lead field is equivalent to rotating its sensor contribution directly. The orthogonal rotation $\boldsymbol{R}(\theta)$ preserves the sum of squares of the transformed sensor contribution.

For a maximum relative angle $\phi_{\max}$, define $\phi(\rho) = \rho\phi_{\max}$ and set

$$\begin{aligned} [T_1(\rho)](\boldsymbol{J}) &= \mathbf{T}\left(\frac{\phi(\rho)}{2}\right)\boldsymbol{J}, \qquad [T_2(\rho)](\boldsymbol{J}) = \mathbf{T}\left(-\frac{\phi(\rho)}{2}\right)\boldsymbol{J}, \\ [L_1(\rho)](\boldsymbol{J}) &= \boldsymbol{R}\left(\frac{\phi(\rho)}{2}\right)\boldsymbol{G}\boldsymbol{J}, \quad [L_2(\rho)](\boldsymbol{J}) = \boldsymbol{R}\left(-\frac{\phi(\rho)}{2}\right)\boldsymbol{G}\boldsymbol{J}. \end{aligned} \tag{80}$$

Let $\boldsymbol{M}_D(\rho) := [\boldsymbol{R}(\phi(\rho)/2) - \boldsymbol{R}(-\phi(\rho)/2)]\boldsymbol{G} \in \mathbb{R}^{d\times p}$, so that $[D(\rho)](\boldsymbol{J}) = \boldsymbol{M}_D(\rho)\boldsymbol{J}$. When $\rho = 0$, both component types undergo the same transformation, and the target observation difference in equation (77) vanishes. The simulation uses the lead field $\boldsymbol{G} \in \mathbb{R}^{64\times 2500}$. Section B.3.1 specifies the construction and calibration of $\phi_{\max}$.

B.2.2 ERP Class Means: Experiment 1A

Experiment 1A retains an early ERP shared by both classes and a base late discriminative component. Reallocation changes the spatial direction of the late class structure. The classifier reads only the late-template projection, so the shared early component does not contribute to its features.

Let $Y \in \{0, 1\}$ be the class label and $\boldsymbol{Z}_y$ the reallocated component in class y. The base activities and reallocated components satisfy

$$\begin{aligned} &\mathbb{E}[\boldsymbol{J}_1 \mid Y = 0] = \boldsymbol{C}, \quad \mathbb{E}[\boldsymbol{J}_1 \mid Y = 1] = \boldsymbol{C} + \boldsymbol{D}_{\mathrm{ERP}}, \\ &\mathbb{E}[\boldsymbol{J}_2 \mid Y = y] = 0, \qquad\qquad \boldsymbol{Z}_0 = 0, \qquad \boldsymbol{Z}_1 = \boldsymbol{Z}_{\mathrm{ERP}}. \end{aligned} \tag{81}$$

Let $\boldsymbol{u}_C = (\boldsymbol{u}_1 + \boldsymbol{u}_2)/\sqrt{2}$. Denote the early and late templates by $\boldsymbol{h}_C, \boldsymbol{h}_D$, respectively, and define

$$\boldsymbol{C} = 0.60\,\boldsymbol{u}_C\boldsymbol{h}_C^\top, \qquad \boldsymbol{D}_{\mathrm{ERP}} = \boldsymbol{u}_1\boldsymbol{h}_D^\top, \qquad \boldsymbol{Z}_{\mathrm{ERP}} = c_A\boldsymbol{u}_2\boldsymbol{h}_D^\top, \quad c_A = 0.50. \tag{82}$$

The parameter $c_A = 0.50$ sets the relative amplitude of the reallocated late component. In the source domain, the reallocated component's late-ERP class-contrast vector has 0.50 times the ℓ_2 norm of the base late component's vector. This structural scale is not selected using classification results. The components share the late temporal template but have orthogonal sensor-space directions.

On the shared time grid, with time in seconds, let $g_{\mu,\sigma}(t) = \exp[-(t-\mu)^2/(2\sigma^2)]$. Write $t_1, \ldots, t_T$ for the sampling times and $\boldsymbol{r}_C = (r_C(t_\ell))_{\ell=1}^T$, $\boldsymbol{r}_D = (r_D(t_\ell))_{\ell=1}^T$ for the sampled waveform vectors. Define the scalar waveforms and normalized templates by

$$\begin{aligned} r_C(t) &= -0.65\, g_{0.120,0.025}(t) + g_{0.200,0.040}(t), \\ r_D(t) &= g_{0.420,0.080}(t), \qquad \boldsymbol{h}_D = \boldsymbol{r}_D/\|\boldsymbol{r}_D\|_2, \\ \boldsymbol{h}_C &= \frac{\boldsymbol{r}_C - (\boldsymbol{r}_C^\top\boldsymbol{h}_D)\boldsymbol{h}_D}{\|\boldsymbol{r}_C - (\boldsymbol{r}_C^\top\boldsymbol{h}_D)\boldsymbol{h}_D\|_2}. \end{aligned} \tag{83}$$

Readout and target consequences. The templates are unit-norm and orthogonal. The classifier reads the fixed late projection $\Phi_A(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{h}_D$; template orthogonality removes the shared early ERP from this feature. Along the reallocation path, the full sensor-by-time class-1 minus class-0 mean difference is the following $d \times T$ matrix:

$$\Delta\boldsymbol{\mu}^{(\mathcal{T})}_{W_\zeta z} = [L_1(\rho)](\boldsymbol{D}_{\mathrm{ERP}}) + \zeta[D(\rho)](\boldsymbol{Z}_1 - \boldsymbol{Z}_0). \tag{84}$$

The first term is the base late class-mean contrast matrix, and the second is the change due to reallocation, which is linear in ζ. Multiplication by $\boldsymbol{h}_D$ gives the corresponding ERP-feature mean-contrast vector.

B.2.3 SPATIAL BANDPOWER: EXPERIMENT 1B

Experiment 1B encodes class differences in spatial energy allocation, without ERP class-conditional mean differences or a difference in mean total bandpower as the base class cue. It tests whether the same reallocation mechanism can have target consequences through second-order structure.

Use the frequency band $\mathcal{B} = \{8, 9, 10, 11, 12\}$ Hz. For the in-band component $\boldsymbol{V}_\mathcal{B}$ of a source process $\boldsymbol{V}$, define

$$\boldsymbol{\Sigma}^V_{y,\mathcal{B}} = \mathbb{E}[T^{-1}\boldsymbol{V}_\mathcal{B}\boldsymbol{V}_\mathcal{B}^\top \mid Y = y], \qquad \Delta\boldsymbol{\Sigma}^V_\mathcal{B} = \boldsymbol{\Sigma}^V_{1,\mathcal{B}} - \boldsymbol{\Sigma}^V_{0,\mathcal{B}}. \tag{85}$$

The construction enforces zero class-conditional means for the source processes. For each of $\boldsymbol{J}_1$ and $\boldsymbol{Z}$, the lead-field projection has equal mean total bandpower in the two classes but different class-conditional spatial second moments. Specifically,

$$\begin{aligned} &\mathbb{E}[\boldsymbol{J}_1 \mid Y = y] = \mathbb{E}[\boldsymbol{J}_2 \mid Y = y] = \mathbb{E}[\boldsymbol{Z} \mid Y = y] = 0, \\ &\operatorname{tr}(\boldsymbol{G}\boldsymbol{\Sigma}^{J_1}_{0,\mathcal{B}}\boldsymbol{G}^\top) = \operatorname{tr}(\boldsymbol{G}\boldsymbol{\Sigma}^{J_1}_{1,\mathcal{B}}\boldsymbol{G}^\top), \quad \boldsymbol{G}\Delta\boldsymbol{\Sigma}^{J_1}_\mathcal{B}\boldsymbol{G}^\top \neq 0, \\ &\operatorname{tr}(\boldsymbol{G}\boldsymbol{\Sigma}^{Z}_{0,\mathcal{B}}\boldsymbol{G}^\top) = \operatorname{tr}(\boldsymbol{G}\boldsymbol{\Sigma}^{Z}_{1,\mathcal{B}}\boldsymbol{G}^\top), \quad \boldsymbol{G}\Delta\boldsymbol{\Sigma}^{Z}_\mathcal{B}\boldsymbol{G}^\top \neq 0. \end{aligned} \tag{86}$$

The trace sums mean bandpower across sensors; equal traces do not require equal sensor-wise power allocations. Background activity $\boldsymbol{J}_2$ contains a zero-mean residual background and an isotropic narrowband carrier. The carrier consists of two independent, identically distributed zero-mean Gaussian temporal processes along $\boldsymbol{q}_1, \boldsymbol{q}_2$, with frequencies in $\mathcal{B}$ and no preferred direction in the sensor plane.

To construct the task process $\boldsymbol{J}_{1,y}$ and reallocated component $\boldsymbol{Z}_y$ in class y, each base trial generates four mutually orthogonal random temporal processes $\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{w}_1, \boldsymbol{w}_2$ with equal weights across the five frequencies. Let $p_H = 0.75$, $p_L = 0.25$, $\boldsymbol{v}_1 = (\boldsymbol{u}_1 + \boldsymbol{u}_2)/\sqrt{2}$, and $\boldsymbol{v}_2 = (\boldsymbol{u}_1 - \boldsymbol{u}_2)/\sqrt{2}$, and take

$$\begin{aligned} \boldsymbol{J}_{1,0} &= \sqrt{p_H}\boldsymbol{u}_1\boldsymbol{h}_1^\top + \sqrt{p_L}\boldsymbol{u}_2\boldsymbol{h}_2^\top, \\ \boldsymbol{J}_{1,1} &= \sqrt{p_L}\boldsymbol{u}_1\boldsymbol{h}_2^\top + \sqrt{p_H}\boldsymbol{u}_2\boldsymbol{h}_1^\top, \\ \boldsymbol{Z}_0 &= \gamma_B\big(\sqrt{p_H}\boldsymbol{v}_1\boldsymbol{w}_1^\top + \sqrt{p_L}\boldsymbol{v}_2\boldsymbol{w}_2^\top\big), \\ \boldsymbol{Z}_1 &= \gamma_B\big(\sqrt{p_L}\boldsymbol{v}_1\boldsymbol{w}_2^\top + \sqrt{p_H}\boldsymbol{v}_2\boldsymbol{w}_1^\top\big). \end{aligned} \tag{87}$$

The classes share these temporal processes and exchange only the spatial directions assigned high and low power. We measure the class contrasts of the reallocated and base task structures in the centered source sensor-bandpower features Φ_B, defined below. For each source configuration, we choose γ_B so that the vector for the reallocated structure has $c_B = 0.35$ times the ℓ_2 norm of the vector for the base task structure.

Cross-term control. The second-order expansion also includes cross-terms between background and reallocated activity. To isolate the added effect of the spatial structure of $\boldsymbol{Z}$, define

$$\boldsymbol{C}^{V,\widetilde{V}}_{y,\mathcal{B}} = \mathbb{E}[T^{-1}\boldsymbol{V}_\mathcal{B}\widetilde{\boldsymbol{V}}_\mathcal{B}^\top \mid Y = y], \qquad \boldsymbol{C}^{J_1,Z}_{y,\mathcal{B}} = \boldsymbol{C}^{J_2,Z}_{y,\mathcal{B}} = 0. \tag{88}$$

Temporal orthogonality removes the cross-moment between $\boldsymbol{J}_1$ and $\boldsymbol{Z}$. Within-model pairing that reverses the background while preserving $\boldsymbol{Z}$ cancels finite-sample cross-moments between background and $\boldsymbol{Z}$. Section B.3.2 gives the sampling procedure.

Readout and target consequences. The classifier reads centered spatial bandpower. It extracts sensor-wise bandpower and removes the common additive component of the power vector:

$$\begin{aligned} \boldsymbol{p}_\mathcal{B}(\boldsymbol{X}) &= \operatorname{diag}(T^{-1}\boldsymbol{X}_\mathcal{B}\boldsymbol{X}_\mathcal{B}^\top), \\ \boldsymbol{P}_\perp = \boldsymbol{I}_d - d^{-1}\mathbf{1}_d\mathbf{1}_d^\top, &\qquad \Phi_B(\boldsymbol{X}) = \boldsymbol{P}_\perp\boldsymbol{p}_\mathcal{B}(\boldsymbol{X}). \end{aligned} \tag{89}$$

The feature Φ_B expresses each sensor's power relative to the mean across sensors. This projection does not remove common multiplicative scaling. Equality of mean total bandpower across classes follows from the generative conditions, not feature centering. With the cross-moments removed, the target class-conditional second-moment difference satisfies

$$\Delta\boldsymbol{\Sigma}^{(\mathcal{T})}_{W_\zeta Z,\mathcal{B}} = \Delta\boldsymbol{\Sigma}^{(\mathcal{T})}_{W_0,\mathcal{B}} + \zeta^2\boldsymbol{M}_D(\rho)\Delta\boldsymbol{\Sigma}^Z_\mathcal{B}\boldsymbol{M}_D(\rho)^\top. \tag{90}$$

Unlike the mean change in 1A, the added second-order term scales with ζ^2. Both instances share a source-equivalent reallocation path and component-specific target transformations. They differ in class structure and readout: late ERP means in 1A, and spatial bandpower of class-conditionally zero-mean signals in 1B.

B.3 Calibration, Sampling, and Evaluation

B.3.1 Source-Configuration Selection and Calibration

We traverse the candidate sequence and retain configurations with valid spatial directions, second-order class differences that remain visible in the classifier features, and finite scaling factors. The first 20 qualifying configurations form the shared evaluation set for 1A and 1B. Selection uses neither classification accuracy, risk endpoints, nor control results.

Transformation and difficulty calibration. The maximum relative angle $\phi_{\max}$ and background signal-to-noise ratio are calibrated on 32 additional configurations disjoint from the evaluation set. Angle calibration compares the population class-contrast directions of the two candidates in the classifier features: late ERP projections for 1A and centered spatial bandpower for 1B. From a prespecified angle grid, we choose the smallest angle satisfying the directional-difference requirements for both instances, without using classification results on the evaluation set.

The signal-to-noise ratio is calibrated for classification difficulty. For each candidate value, the same classification procedure is applied to independent calibration training and test sets. We choose the value whose median balanced accuracy across configurations is closest to 0.775, breaking ties toward the lower value. The target 0.775 represents moderate classification difficulty. The EEG lead field is generated with MNE-Python (Gramfort et al., 2013). Table 2 lists the candidate rules and selected values.

Table 2: Forward model, calibration, and sampling settings. Calibration and risk evaluation use disjoint configuration sets.

Setting	Value or rule
EEG lead field	MNE-Python 1.11.0; `standard_1020` electrode montage; 64 sensors
Head model and source space	Four-layer spherical head model; 20 mm volumetric source grid; 250 scalar source coordinates
Configurations	20 evaluation and 32 calibration configurations, with no overlap
Trial	1 s; 250 Hz; $T = 250$
Residual neural background	First-order autoregressive processes in 128 source regions; coefficients from 0.90 to 0.98
Candidate relative angles	30° to 180° in increments of 15°
Angle eligibility	1A: directional angle at least 15°. 1B: all 32 angles finite and nonzero, median at least 15°, and 10th percentile at least 5°. Select the smallest eligible candidate.
Maximum relative angle	$\phi_{\max} = 135^\circ$
Candidate and selected signal-to-noise ratios	Candidates $-6, -10, -14$ dB; -14 dB for 1A and -6 dB for 1B
Measurement-noise power	0.10 times the residual-background power after lead-field propagation
Training samples	1000 base trials per class, yielding 2000 paired observations per class
Test samples	2500 base trials per class, yielding 5000 paired observations per class
Unlabeled scale estimation	10000 independent samples per evaluation configuration, excluded from classifier fitting
Difficulty-calibration samples	256 training and 512 test base trials per class and candidate value
Difficulty target	Median balanced accuracy across configurations closest to 0.775
Reference classifier	scikit-learn shrinkage-LDA

B.3.2 Background Generation and Paired Sampling

We specify the residual neural background in $\boldsymbol{J}_2$, measurement noise, and sampling pairs. In 1B, $\boldsymbol{J}_2$ also includes the narrowband carrier described in B.2.3; measurement noise is the separate additive sensor term in equation (77). For each background source region $r = 1, \ldots, 128$, generate the

temporally correlated process

$$\varphi_r = 0.90 + 0.08\frac{r-1}{127}, \qquad h_r(0) \sim \mathcal{N}(0,1),$$
$$h_r(t) = \varphi_r h_r(t-1) + \sqrt{1-\varphi_r^2}\,\varepsilon_r(t), \quad \varepsilon_r(t) \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{91}$$

The values $h_r(t)$ are scalar; their sampled sequence forms the temporal vector $\boldsymbol{h}_r$. Let $\boldsymbol{v}_{N,r}$ be the region's sensor topography under the shared lead field, and $\boldsymbol{N}^0_{\text{fwd}}$ the unscaled sensor representation of the residual neural background. Measurement noise is also generated as an unscaled matrix $\boldsymbol{\eta}$. The unscaled background and noise are

$$\boldsymbol{N}^0_{\text{fwd}} = \frac{1}{\sqrt{128}}\sum_{r=1}^{128} \boldsymbol{v}_{N,r}\boldsymbol{h}_r^\top, \qquad \eta_{j,t} \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{92}$$

The scale setting below converts $\boldsymbol{\eta}$ into the measurement-noise term $\boldsymbol{\eta}_\rho^{(\mathcal{T})}$ in equation (77). Background and measurement noise are mutually independent and independent of the controlled source activity. Their random realizations are independent between classes and between training and test sets.

Scale setting. The signal-to-noise ratio is controlled task-signal power divided by the total power of the remaining random components. It uses late-template projection power in 1A and sensor-averaged bandpower in 1B. Measurement-noise power is first set to 0.10 times the residual-background power after lead-field propagation, followed by scaling to the selected signal-to-noise ratio. In 1A, the residual background and measurement noise are scaled together. In 1B, the narrowband carrier scale is fixed, and only the other two terms are scaled so that all three attain the specified background power. For each evaluation configuration, we estimate the scales from 10000 unlabeled samples excluded from classifier fitting. These scales are shared across the two candidates, training and test sets, and all conditions.

Sampling and pairing. For each instance, configuration, and condition, generate $\mathcal{D}_0^{\text{tr}}, \mathcal{D}_{\zeta Z}^{\text{tr}}, \mathcal{D}_0^{\text{te}}, \mathcal{D}_{\zeta Z}^{\text{te}}$, with equal class counts in every set. Training and test base draws are independent. Pairing operates at three levels: the same base draw is used in both candidates to form trial-wise model pairs; base draws are reused across conditions; and each model also uses the sign-reversal pairs described below. The first two reduce sampling variation in between-model and between-condition comparisons, respectively.

Within-model sign-reversal pairing negates the random components not assigned to the task. It reverses residual neural background and measurement noise in 1A, and also the narrowband carrier in 1B. In conditions where $\boldsymbol{Z}$ does not change with the pairing sign, this cancels incidental finite-sample cross-moments between background and $\boldsymbol{Z}$. Thus 1000 training base trials per class yield 2000 observations per class, and 2500 test base trials per class yield 5000 observations per class. Base trials and paired observations are distinct counting units. Statistical aggregation resamples source configurations, rather than treating paired observations as independent resampling units (B.3.5).

B.3.3 Target-Trained Classifiers and Bidirectional Exchange Loss

We fit scikit-learn shrinkage-LDA (Pedregosa et al., 2011) separately on $\mathcal{D}_0^{\text{tr}}$ and $\mathcal{D}_{\zeta Z}^{\text{tr}}$ to obtain f_0 and $f_{\zeta Z}$. Instance 1A uses Φ_A, and 1B uses Φ_B; below, f includes both fixed feature extraction and the classification rule. These classifiers compare target rules between the two models; they are not source-trained UDA algorithms.

For the target test set of model W, let $n_{W,y}$ be the number of class-y samples, and define

$$\text{BAcc}_W(f) = \frac{1}{2}\sum_{y=0}^{1}\frac{1}{n_{W,y}}\sum_{i:Y_i=y}\mathbf{1}\{f(\boldsymbol{X}_{W,i}^{(\mathcal{T})}) = y\}, \qquad R_W^{\text{bal}}(f) = 1 - \text{BAcc}_W(f). \tag{93}$$

The bidirectional exchange loss is

$$\alpha_{\text{struct}} = \min\left\{R_{W_0}^{\text{bal}}(f_{\zeta Z}) - R_{W_0}^{\text{bal}}(f_0),\ R_{W_{\zeta Z}}^{\text{bal}}(f_0) - R_{W_{\zeta Z}}^{\text{bal}}(f_{\zeta Z})\right\}. \tag{94}$$

The condition $\alpha_{\text{struct}} > 0$ means that each model's own trained classifier outperforms the exchanged classifier. This measures the interchangeability of the fitted pair, not the Bayes conflict defined over all measurable classifiers in the main text.

Table 3: Four evaluation cells for classifier exchange. Each row compares the classifier trained under the other model with the classifier trained under that row's model. Positive bidirectional exchange loss requires an error increase in both directions.

Target test set	Using f_0	Using $f_{\zeta Z}$
$\mathcal{D}_0^{\mathrm{te}}$	$R_{W_0}^{\mathrm{bal}}(f_0)$	$R_{W_0}^{\mathrm{bal}}(f_{\zeta Z})$
$\mathcal{D}_{\zeta Z}^{\mathrm{te}}$	$R_{W_{\zeta Z}}^{\mathrm{bal}}(f_0)$	$R_{W_{\zeta Z}}^{\mathrm{bal}}(f_{\zeta Z})$

B.3.4 Experimental Conditions and Their Roles

The primary condition is $(\zeta, \rho) = (1, 0.75)$, where we abbreviate $W_{\zeta Z} = W_Z$ and $f_{\zeta Z} = f_Z$. Other comparisons examine where the effect occurs in parameter space and which construction factors affect it. Each instance and configuration has 52 conditions; Table 4 also identifies their result locations.

Table 4: Conditions, questions, and result locations. $\zeta = 0$ and $\rho = 0$ are included in the 25 grid conditions and are not counted again.

Condition group	Count	What changes or stays fixed	Question and results
Full parameter grid	25	$\zeta, \rho \in \{0, 0.25, 0.50, 0.75, 1\}$	Is the primary condition isolated? B.4.1–B.4.2, Figure 5
Rotation-allocation controls	3	Transform only $\boldsymbol{J}_1$, only $\boldsymbol{J}_2$, or use $T_1 = T_2$	Are relative transformation and symmetric angle allocation needed? B.4.3, Table 7
Feature-matching controls	4	Preserve full target EEG differences while matching classifier features	Are full-EEG differences alone sufficient? B.4.4
Label-decoupled matched controls	20	Choose one of two generated reallocation states per base trial, with balanced assignments shared by both classes; preserve the remaining base draws	How does the primary condition compare with matched controls? B.4.4, Table 8
Total	52		

Rotation allocation. The full grid divides the relative angle symmetrically as in equation (80). Under the primary condition, $\boldsymbol{J}_1, \boldsymbol{J}_2$ use $\boldsymbol{R}(+50.625°)$ and $\boldsymbol{R}(-50.625°)$, respectively. The controls transforming only $\boldsymbol{J}_1$ and only $\boldsymbol{J}_2$ use $(\boldsymbol{R}(+101.25°), \boldsymbol{I})$ and $(\boldsymbol{I}, \boldsymbol{R}(-101.25°))$, respectively; the $T_1 = T_2$ control applies $\boldsymbol{R}(+101.25°)$ to both. The first two retain different component-specific target transformations but change angle allocation. The last retains a common cross-domain change while eliminating the component-operator difference.

Feature matching. Each instance has four conditions retaining nonzero full target observation differences. In 1A, adding deterministic spatiotemporal activity with a temporal template orthogonal to the fixed late template $\boldsymbol{h}_D$ makes Φ_A identical for corresponding trials. In 1B, class-specific distribution-preserving orthogonal transformations of the isotropic narrowband carrier are implemented by a corresponding reallocated component. They make the distribution of Φ_B identical within each class, without requiring trial-wise feature equality. These controls match the classifier features rather than the full EEG.

Matched controls. Each configuration uses 20 balanced state assignments. For each base-trial index, one of the two already generated reallocation states is selected and used in both classes. Class labels, task components, and base random draws retain their original values. This removes the association between reallocation state and label without changing the labels. The controls compare the primary condition with label-decoupled results; they do not replace the primary criterion.

B.3.5 Configuration-Level Statistical Inference

For instance $E \in \{A, B\}$ and configuration $b = 1, \ldots, 20$, denote the primary endpoint by $\alpha_{E,b}$ and aggregate it as $\bar{\alpha}_E = 20^{-1} \sum_b \alpha_{E,b}$. We resample configurations with replacement 10000 times and recompute the mean for each resample, following the percentile bootstrap method (Efron, 1981). The

5th percentile of these bootstrap means gives the one-sided 95% lower confidence bound. Instances 1A and 1B share resampling indices to preserve configuration pairing but have separate endpoints and intervals.

Experiment 1 meets the primary criterion only if both instances' lower confidence bounds exceed zero. The two signal instances are not pooled into one effect size. All empirical quantiles use linear interpolation: for n sorted values, the position at level p is $h = 1 + (n - 1)p$.

Matched controls are first summarized within configurations. We take either the maximum or the empirical 95th percentile of the 20 control endpoints, then subtract this reference from the primary endpoint in the same configuration. The same configuration bootstrap estimates the mean difference and its one-sided lower confidence bound. Control comparisons and the primary endpoint thus use the same aggregation unit but answer different questions.

B.4 Results and Mechanism Controls

The primary condition and full parameter grid establish the observed exchange losses. Transformation, feature, and label controls then identify the factors on which they depend.

B.4.1 Bidirectional Exchange Loss at the Primary Condition

Under the primary condition $(\zeta, \rho) = (1, 0.75)$, the mean α_{struct} and its one-sided lower confidence bound are positive in both 1A and 1B, meeting the primary criterion (Table 5). All 20 configurations are positive in 1A; 1B has 19 positive and 1 negative value. The instances are calibrated separately and test rule exchange within each instance, rather than the intrinsic sensitivity of the two types of EEG structure.

Table 5: Results across configurations at the primary condition $(\zeta, \rho) = (1, 0.75)$. Means and one-sided 95% lower confidence bounds use 20 configurations. The instances have separate calibrations and different features, so the values are not used to compare effect sizes between instances.

Instance	Mean α_{struct}	One-sided 95% lower bound	Positive/negative configurations
1A: ERP means	0.0826	0.0708	20 / 0
1B: Spatial bandpower	0.1523	0.1097	19 / 1

Table 6 separates the two exchange directions. In both instances, each model's own classifier has higher mean balanced accuracy. The primary result is therefore not a one-way difference caused by one classifier being weaker under both models.

Table 6: Classifier–target-model cross-evaluation at the primary condition. The first two numeric columns give mean balanced accuracy across configurations; the last gives the mean error increase from replacing the model's own classifier with the other model's classifier. The reported α_{struct} is averaged after taking the smaller increase within each configuration, not the minimum of the two mean increases in this table.

Instance and target model	Using f_0	Using f_Z	Mean exchange error increase
1A: W_0	0.8079	0.7092	0.0987
1A: W_Z	0.6471	0.7442	0.0971
1B: W_0	0.7252	0.4224	0.3028
1B: W_Z	0.4424	0.5947	0.1523

B.4.2 Dependence on Reallocation and Operator Differences

To assess whether the primary condition is isolated, Figure 5 jointly varies ζ and ρ on the same 20 configurations, comparing it with the other combinations in the full grid.

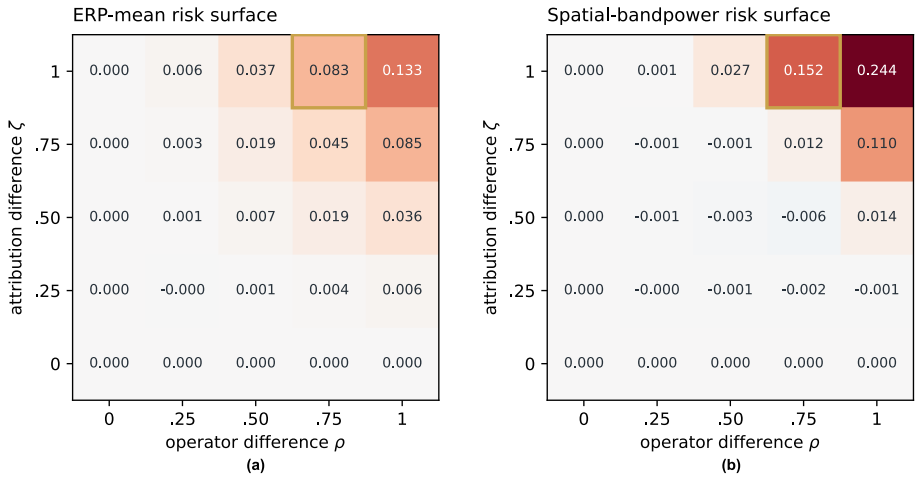

Figure 5: Joint parameter surfaces for reallocation and component-operator differences. Left: 1A; right: 1B. The horizontal axis is ρ, the vertical axis is ζ, and each cell gives mean α_{struct} across 20 configurations. Both panels share a color scale; the gold outline marks the primary condition. The grid describes parameter dependence, while the primary criterion applies only to the primary condition.

The primary condition lies among neighboring positive values and is not an isolated positive result. The grid is not cell-wise monotone: 1A has one slightly negative cell, -0.00043 at $(0.25, 0.25)$; 1B has 9 negative cells, with a minimum of -0.00636 at $(0.50, 0.75)$. These near-zero values are retained without truncation. Both boundaries, $\zeta = 0$ and $\rho = 0$, return exactly to zero in every configuration.

Thus the grid supports a non-isolated positive result at the primary condition, but not strict monotonic growth of exchange loss in both parameters throughout the grid.

B.4.3 Removing Reallocation or Operator Differences and Changing Rotation Allocation

When $\zeta = 0$, $\rho = 0$, or $T_1 = T_2$, the target observation difference in equation (77) vanishes. Both instances yield exactly zero exchange loss under these conditions (Table 7), consistent with the construction's cancellation identity.

The other two controls keep the two component-specific operators unequal while changing how the relative rotation angle is allocated (B.3.4). Mean exchange loss remains positive when only $\boldsymbol{J}_1$ or only $\boldsymbol{J}_2$ is transformed. The first group therefore checks cancellation of observation differences, whereas the second shows that positive results are not limited to symmetric angle allocation.

Table 7: Mean α_{struct} under transformation controls. Every condition uses the same 20 source configurations. The first three rows test cancellation; the last two change rotation allocation.

Condition	1A: ERP means	1B: Spatial bandpower
$\zeta = 0$	0	0
$\rho = 0$	0	0
$T_1 = T_2$	0	0
Transform only $\boldsymbol{J}_1$	0.0824	0.1921
Transform only $\boldsymbol{J}_2$	0.0825	0.1649

B.4.4 Feature Matching and Controls on the Reallocation-State–Label Association

The controls above address attribution and transformation factors. We next distinguish changes in the full EEG from changes in the inputs actually used by the classifier.

Feature matching. The feature-matched controls retain differences in the full target EEG while matching the inputs actually used by the classifier. In 1A, corresponding trials have exactly the same Φ_A, and all four conditions yield zero exchange loss. In 1B, the class-conditional distributions of Φ_B coincide, and the configuration-averaged exchange loss ranges across the four conditions from -0.00154 to -0.00120, with all four means close to zero. Thus, a change in the full observation alone does not explain the exchange loss in the main conditions; what matters is whether the change affects the classifier's features and their class-conditional distributions.

Decoupling from labels. The label-decoupled matched controls use the balanced state assignments described in Section B.3.4 to remove the correspondence between reattribution states and labels, while keeping the classification labels, task components, base samples, and sample sizes unchanged. Table 8 compares the primary endpoint with the 95th percentile or maximum of the reference set for each configuration. In both instances, the mean differences are positive and their lower confidence bounds exceed zero; 1B still has 4 configurations with negative differences. Thus, the primary endpoint exceeds these matched-control references on average across configurations, but the difference is not positive in every configuration.

Table 8: Differences between the main conditions and label-decoupled references. The reference value is subtracted from the primary endpoint within each configuration before aggregation across configurations. The reported interval endpoint is a one-sided 95% lower confidence bound. These control comparisons do not replace the primary criterion in Table 5.

Instance	Reference	Mean difference	Lower confidence bound	Configurations with positive / negative differences
1A	Linearly interpolated 95th percentile	0.0604	0.0473	20 / 0
1A	Maximum	0.0599	0.0468	20 / 0
1B	Linearly interpolated 95th percentile	0.1515	0.1063	16 / 4
1B	Maximum	0.1505	0.1067	16 / 4

By matching classifier features or decoupling reallocation states from labels, these two controls narrow the interpretation from differences in target EEG to class-related differences accessible to the classifier.

Scope of interpretation. The conclusions concern the constructed models, fixed features, and fitted target reference classifiers. α_{struct} measures the interchangeability of these two rules, rather than directly estimating pairwise Bayes conflict. Whether latent attribution in measured EEG takes the same form is outside the scope of this construction. The paired sampling and reference classifiers also differ from the independent sampling and spectral learner in Theorem 2 of the main text, so these results do not experimentally validate that finite-sample upper bound. Instance 1B is outside its ERP mean subclass (Appendices A.6.1–A.6.2).

Within these limits, Experiment 1 provides two statistical instances under a common EEG lead field: reattribution that leaves source observations unchanged trial by trial can produce bidirectional exchange loss between reference classifiers through component-specific target transformations. The full grid, transformation-assignment controls, feature-matched controls, and matched controls further delimit the parameter range and interpretation of this conclusion.

C Experiment 2: Learner-Side Attribution under Real EEG Transfer

This appendix provides the experimental details for Experiment 2 in Section 6.2. Section C.1 defines the data, learners, component-level probe, and statistical analysis underlying the main result. Section C.2 expands the main result, Section C.3 examines two alternative explanations through supplementary analyses, and Section C.4 collects additional experimental details.

C.1 Experimental Setup

Table 9: Dataset structure and DG partitions. Source trial counts are totals across source subjects, whereas target trial counts are per subject. Each trial-count pair lists the positive class followed by the negative class.

Setting	Lee	OpenBMI
Recording structure	30 blocks per subject; 10 positive and 30 negative trials per block	33 offline and 36 online groups per subject; 10 positive and 50 negative trials per group
Source / target subjects	10 / 20	18 / 36
Source data	All 30 blocks	Offline recording
Source split	80% fit and 20% validation, stratified by class within subject	26 fit groups and 7 validation groups
Source fit trials, positive / negative	2,400 / 7,200	4,680 / 23,400
Source validation trials, positive / negative	600 / 1,800	1,260 / 6,300
Target data	All 30 blocks	Offline and online recordings
Target trials per subject, positive / negative	300 / 900	690 / 3,450

C.1.1 Data and Splits

To examine learner-side attribution under real inter-subject transfer, we use subject-wise domain generalization on two independently collected P300 datasets. Treating each subject as a domain allows learners trained on a common source population to be compared on previously unseen individuals. We use the television task from the ERP appliance-control dataset of Lee et al. (2024), abbreviated Lee, and the Session 1 ERP recordings from OpenBMI (Lee et al., 2019). Both datasets use a target-detection task in which target stimuli elicit the P300 response and form the positive class, while nontarget stimuli form the negative class. We index the positive and negative classes by $q = +$ and $q = -$, respectively. The Lee recordings used here consist of 30 online testing blocks per subject. OpenBMI provides separate offline and online recordings for each subject, with each recording organized into stimulus groups of 60 presentations.

Within each dataset, subjects are divided into disjoint source and target sets. To separate learner construction from source-side comparison, the source recordings are further divided into fit and validation sets. Lee trials are split within each source subject using a fixed class-stratified 80%/20% partition. For OpenBMI, only the offline recording is used on the source side, and its stimulus groups are kept intact when 26 groups are assigned to source fit and 7 to source validation. Source fit provides the data used to train the EEGNets; its positive trials are also used to fit the ERP component decomposition described in Section C.1.3. Source validation is held out from training and provides an independent source sample for comparing the trained learners. Because the DG learners are constructed without target data, evaluation uses all selected recordings for each target subject. This includes all 30 Lee blocks and both the offline and online OpenBMI recordings. The two OpenBMI recordings are pooled at the trial level before the target quantities are computed, and the positive and negative classes are evaluated separately.

To retain the post-stimulus ERP activity in a common temporal representation, both datasets follow the same preprocessing procedure. Signals are band-pass filtered from 1 to 10 Hz using a zero-phase

fourth-order Butterworth filter, re-referenced to the common average, and resampled to 128 Hz. Epochs cover $[0, 0.8)$ s after stimulus onset. To reduce trial-specific differences in channel scale, each channel is centered and divided by its temporal standard deviation within each trial. The resulting input dimensions and the denominator floor used in trial standardization are reported in Table 20.

C.1.2 EEGNet Learners

Comparing learner-side attribution requires variation across learners without changing the underlying source learning problem. For each dataset, we therefore train 100 EEGNets (Lawhern et al., 2018) on the same source fit data using the same training procedure, varying only the random seed. We retain the final checkpoint from each run, producing all $\binom{100}{2} = 4{,}950$ unordered learner pairs. The resulting learner library is fixed before target evaluation. Target data play no role in training the networks or selecting checkpoints. During evaluation, model parameters and BatchNorm running statistics remain fixed, and dropout is disabled. Training and evaluation settings and the EEGNet architecture are reported in Tables 18 and 19.

C.1.3 ERP Component Decomposition

Comparing component dependence across learners requires a common set of ERP coordinates. We obtain these coordinates with xDAWN, a standard spatial filtering method for ERP analysis (Rivet et al., 2009), using the implementation in MNE-Python (Gramfort et al., 2013). For each dataset, xDAWN is fitted after learner training using the positive trials in the source fit set. The resulting decomposition is then held fixed across all learners, target subjects, and class-specific analyses.

For $n = 1, \dots, N_+$, let $\boldsymbol{X}_n \in \mathbb{R}^{C\times T}$ denote a positive source fit epoch, where N_+ is the number of such epochs and C and T are the numbers of channels and time samples. Let $\boldsymbol{x}_{n,\tau} \in \mathbb{R}^C$ be the observation at time index $\tau \in \{1, \dots, T\}$ in epoch n. Their event-locked average is $\overline{\boldsymbol{X}} = N_+^{-1} \sum_{n=1}^{N_+} \boldsymbol{X}_n$, with $\overline{\boldsymbol{x}}_\tau$ denoting its observation at time index τ. The calculation uses two channel-covariance matrices. With the input signals treated as zero-centered, the covariance across all positive source fit observations is $\boldsymbol{S}_X = (N_+T)^{-1} \sum_{n=1}^{N_+} \sum_{\tau=1}^{T} \boldsymbol{x}_{n,\tau}\boldsymbol{x}_{n,\tau}^\top$, while the covariance of the event-locked response is $\boldsymbol{S}_{\mathrm{ERP}} = T^{-1} \sum_{\tau=1}^{T} \overline{\boldsymbol{x}}_\tau \overline{\boldsymbol{x}}_\tau^\top$. Both are regularized by the fixed-shrinkage operator $\mathcal{R}_\lambda(\boldsymbol{S}) = (1-\lambda)\boldsymbol{S} + \lambda\, \mathrm{tr}(\boldsymbol{S})\boldsymbol{I}_C/C$, where λ is the fixed shrinkage coefficient and $\boldsymbol{I}_C$ is the $C \times C$ identity matrix. This gives $\widehat{\boldsymbol{\Sigma}}_X = \mathcal{R}_\lambda(\boldsymbol{S}_X)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{ERP}} = \mathcal{R}_\lambda(\boldsymbol{S}_{\mathrm{ERP}})$. Let $\boldsymbol{I}_4$ denote the 4×4 identity matrix. Using these covariance estimates, xDAWN selects four directions that retain strong event-locked activity relative to the total variation in the positive trials.

$$\underset{\widetilde{\boldsymbol{\Omega}}\in\mathbb{R}^{4\times C}}{\text{maximize}} \quad \mathrm{tr}\left(\widetilde{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Sigma}}_{\mathrm{ERP}}\widetilde{\boldsymbol{\Omega}}^\top\right) \qquad \text{subject to} \qquad \widetilde{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Sigma}}_X\widetilde{\boldsymbol{\Omega}}^\top = \boldsymbol{I}_4. \tag{95}$$

The four retained filter directions are normalized and stacked as the rows of $\boldsymbol{\Omega}$. For any trial $\boldsymbol{X}$, the rows of $\boldsymbol{\Omega}\boldsymbol{X}$ are its component time courses. The xDAWN and Haufe numerical settings are reported in Table 20.

The spatial filters define the component time courses, whereas the subsequent intervention requires the contribution of each component in the sensor-time space used by EEGNet. For this purpose, we concatenate the positive source fit observations and center each channel to form $\boldsymbol{X}_{\mathrm{cat}} \in \mathbb{R}^{C\times N_+T}$, with component scores $\boldsymbol{Z}_{\mathrm{cat}} = \boldsymbol{\Omega}\boldsymbol{X}_{\mathrm{cat}}$. The corresponding Haufe forward patterns are defined through the reconstruction problem

$$\boldsymbol{P} = \underset{\widetilde{\boldsymbol{P}}\in\mathbb{R}^{C\times 4}}{\arg\min} \left\| \boldsymbol{X}_{\mathrm{cat}} - \widetilde{\boldsymbol{P}}\boldsymbol{Z}_{\mathrm{cat}} \right\|_F^2. \tag{96}$$

For an arbitrary trial $\boldsymbol{X}$, the k-th component time course and its reconstructed sensor-time contribution are

$$\boldsymbol{z}_k(\boldsymbol{X}) = [\boldsymbol{\Omega}]_{k,:}\boldsymbol{X}, \qquad \boldsymbol{X}_k = [\boldsymbol{P}]_{:,k}\boldsymbol{z}_k(\boldsymbol{X}).$$

Defining the residual after the four reconstructed components gives the exact additive representation

$$\boldsymbol{R} = \boldsymbol{X} - \sum_{k=1}^{4} \boldsymbol{X}_k, \qquad \boldsymbol{X} = \sum_{k=1}^{4} \boldsymbol{X}_k + \boldsymbol{R}. \tag{97}$$

Each $\boldsymbol{X}_k$ has the same sensor-time dimensions as the original trial and can therefore be removed directly from the EEGNet input. Together, the four reconstructed contributions provide the common representation used to compare component-deletion responses across learners in Section C.1.4.

C.1.4 Component Deletion Responses

The shared decomposition allows every learner to be examined through the same component-wise intervention. Let i and j index learners, and let $k \in \{1, \dots, 4\}$ index the retained components. We remove the reconstructed contribution $\boldsymbol{X}_k$ while retaining the other three contributions and the residual:

$$\boldsymbol{X}^{(-k)} = \boldsymbol{X} - \boldsymbol{X}_k.$$

Let $\ell_i(\boldsymbol{X})$ denote the positive-class logit minus the negative-class logit of learner i. For a trial $\boldsymbol{X}$ in class q, the response of learner i to deleting component k is

$$c^q_{i,k}(\boldsymbol{X}) = \ell_i(\boldsymbol{X}) - \ell_i\left(\boldsymbol{X}^{(-k)}\right). \tag{98}$$

The deleted input is evaluated directly by the frozen network without additional normalization. The sign of $c^q_{i,k}(\boldsymbol{X})$ records whether the component increases or decreases the logit margin in the context of the full trial.

The trial-level responses must be aggregated over source validation before learners can be compared. To represent each source subject equally, we first average within subjects and then across subjects. Let $\mathcal{V}^q_a$ denote the source-validation trials of class q from subject a, let $n_{\mathcal{S}}$ be the number of source subjects, and let ψ be any scalar function of a trial. We define

$$\mathbb{E}_{\mathcal{S},q}[\psi(\boldsymbol{X})] = \frac{1}{n_{\mathcal{S}}} \sum_{a=1}^{n_{\mathcal{S}}} \frac{1}{|\mathcal{V}^q_a|} \sum_{\boldsymbol{X} \in \mathcal{V}^q_a} \psi(\boldsymbol{X}). \tag{99}$$

Using this source expectation, we construct the attribution difference in two steps:

$$\begin{aligned} s^q_i &= \left[\frac{1}{4}\mathbb{E}_{\mathcal{S},q} \sum_{k=1}^{4} \left(c^q_{i,k}(\boldsymbol{X})\right)^2\right]^{1/2}, \\ A^{\mathcal{S},q}_{ij} &= \frac{1}{2}\mathbb{E}_{\mathcal{S},q} \sum_{k=1}^{4} \left| \frac{c^q_{i,k}(\boldsymbol{X})}{s^q_i} - \frac{c^q_{j,k}(\boldsymbol{X})}{s^q_j} \right|. \end{aligned} \tag{100}$$

First, s^q_i summarizes the RMS magnitude of learner i's four deletion responses in class q. The same value normalizes all four components, making the comparison invariant to a common positive rescaling of the learner's responses while retaining their relative organization. Second, $A^{\mathcal{S},q}_{ij}$ measures the average distance between the normalized deletion responses of learners i and j. The absolute difference is formed for each trial before averaging, so differences that vary across source trials are retained. The factor $1/2$ fixes the scale of the four-component distance. This quantity is denoted by A in the main text.

C.1.5 Predictive Divergence and Statistical Analysis

Section 3 motivates the empirical question of whether differences in how learners organize predictive activity across components become visible after transfer. We therefore examine whether learner pairs with larger source attribution differences A also show larger target predictive divergences M. The source divergence D describes how differently the same pair predicts on source-validation trials and accounts for predictive differences already present at source. Both D and M are measured using Jensen–Shannon (JS) divergence (Lin, 1991). For two binary probability vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, define

$$\mathrm{JS}(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2}\,\mathrm{KL}(\boldsymbol{u} \| \boldsymbol{m}) + \frac{1}{2}\,\mathrm{KL}(\boldsymbol{v} \| \boldsymbol{m}), \qquad \boldsymbol{m} = \frac{\boldsymbol{u} + \boldsymbol{v}}{2}, \tag{101}$$

using natural logarithms. For target subject t, let $\mathcal{T}^q_t$ contain the evaluation trials of class q. With $p_i(\boldsymbol{X})$ denoting the positive-class probability and $\boldsymbol{\pi}_i(\boldsymbol{X}) = (1 - p_i(\boldsymbol{X}), p_i(\boldsymbol{X}))$, the source and

target measures are

$$D_{ij}^{\mathcal{S},q} = \mathbb{E}_{\mathcal{S},q}\,\mathrm{JS}(\boldsymbol{\pi}_i(\boldsymbol{X}), \boldsymbol{\pi}_j(\boldsymbol{X}))\,,$$
$$M_{ij,t}^{\mathcal{T},q} = \frac{1}{|\mathcal{T}_t^q|} \sum_{\boldsymbol{X}\in\mathcal{T}_t^q} \mathrm{JS}(\boldsymbol{\pi}_i(\boldsymbol{X}), \boldsymbol{\pi}_j(\boldsymbol{X}))\,, \tag{102}$$

Both measures use the full EEG input and compare the complete binary probability vectors on the same trials. Class q specifies the trials over which each divergence is averaged. These quantities correspond to D and M in Section 6.2. The source measures A and D use the same validation trials and the subject weights defined in Eq. (99).

Two DG learners that differ more in their source-validation predictions may also differ more in their source component responses and target predictions. We therefore test whether the association between A and M remains after adjusting for their same-class source predictive divergence D. Because the question concerns whether larger values of A accompany larger values of M, the analysis is based on their ranks. For a fixed target subject and class, we rank A, M, and D across the 4,950 learner pairs. Writing rk for average ranks, we fit regressions with an intercept and compute

$$\mathrm{rk}(A) = \alpha_A + \beta_A\,\mathrm{rk}(D) + e_A,$$
$$\mathrm{rk}(M) = \alpha_M + \beta_M\,\mathrm{rk}(D) + e_M, \tag{103}$$
$$r_{t,q}^{\mathrm{adj}} = \mathrm{Cor}(e_A, e_M).$$

Here, e_A and e_M are the residuals from the two rank regressions. Their Pearson correlation is the partial Spearman correlation used in the main text. It measures the remaining rank association between A and M after adjustment for source divergence in the same class. Each target subject supplies a separate empirical evaluation distribution, so we compute the association separately for every target subject and then give subjects equal weight in the reported mean. Let $n_{\mathcal{T}}$ be the number of target subjects.

$$\bar{r}_q = \frac{1}{n_{\mathcal{T}}} \sum_{t=1}^{n_{\mathcal{T}}} r_{t,q}^{\mathrm{adj}}, \tag{104}$$

For comparison, the unadjusted Spearman correlations between A and M are averaged in the same way.

Because each learner appears in 99 pairs, uncertainty estimation must account for the dependence shared by those pairs. We use a model jackknife over learners (Efron & Stein, 1981), deleting one learner and all 99 pairs involving it. On the remaining 4,851 pairs, we recompute the ranks, regressions, and target averages, obtaining $\bar{r}_q^{(-b)}$ when learner b is removed. The standard error and pointwise 95% confidence interval are

$$\bar{r}_q^{(-\cdot)} = \frac{1}{100} \sum_{b=1}^{100} \bar{r}_q^{(-b)},$$
$$\widehat{\mathrm{SE}}_q = \left[\frac{99}{100} \sum_{b=1}^{100} \left(\bar{r}_q^{(-b)} - \bar{r}_q^{(-\cdot)} \right)^2 \right]^{1/2}, \tag{105}$$
$$\mathrm{CI}_{95\%} = \bar{r}_q \;\pm\; t_{99,0.975} \widehat{\mathrm{SE}}_q,$$

where $t_{99,0.975}$ is the 97.5th percentile of the Student t distribution with 99 degrees of freedom. Treating random seeds as independent realizations of the fixed training procedure, these intervals describe variation among learners conditional on the observed data, subjects, and component probes. Ranges across target subjects describe heterogeneity within the evaluated sample; inference to newly sampled subjects would require a separate estimate of subject-sampling uncertainty.

C.2 Details of the Main Result

The main text reports the adjusted association and its consistent direction across the evaluated target subjects. Tables 10 and 11 place this finding in context by showing how source adjustment changes the association and how its strength varies across subjects, respectively. Adjustment for same-class source predictive divergence reduced the unadjusted mean in every dataset–class combination, indicating

that part of the raw relation covaried with predictive differences already present at source. Even after this reduction, all four adjusted means and their pointwise model-jackknife intervals remained above zero. The target-wise estimates varied in magnitude but retained the same direction for every evaluated subject. Together, these additional results support a consistent relation across both datasets and classes after accounting for measured source predictive divergence.

Table 10: Mean associations between source attribution differences and target predictive divergence. Unadjusted Spearman and adjusted partial Spearman correlations are averaged equally across target subjects. Standard errors and pointwise 95% confidence intervals for the adjusted means are obtained via a model jackknife.

Dataset	Class	Unadjusted mean	Adjusted mean	Jackknife SE	95% interval
Lee	Positive	0.391	0.256	0.030	[0.197, 0.315]
Lee	Negative	0.539	0.248	0.030	[0.189, 0.308]
OpenBMI	Positive	0.329	0.210	0.023	[0.165, 0.256]
OpenBMI	Negative	0.543	0.184	0.027	[0.130, 0.238]

Table 11: Variation across target subjects in the adjusted association between source attribution differences and target predictive divergence. The median and range summarize adjusted subject-wise partial Spearman correlations in the evaluated sample. The final column reports the number of subjects with a positive adjusted correlation out of the total evaluated.

Dataset	Class	Median	Range	Positive / total subjects
Lee	Positive	0.268	[0.140, 0.332]	20 / 20
Lee	Negative	0.256	[0.083, 0.400]	20 / 20
OpenBMI	Positive	0.213	[0.083, 0.310]	36 / 36
OpenBMI	Negative	0.172	[0.031, 0.405]	36 / 36

The main analysis leaves two possible explanations unresolved. The remaining relation could be concentrated among learner pairs that are still relatively dissimilar at source. In addition, A may reflect general learner dissimilarity, leaving its component specificity uncertain. Section C.3 examines the first possibility within source-similarity strata and the second through deletion of individual reconstructed components from target EEG.

C.3 Examining Alternative Explanations

C.3.1 Restricting the Comparison by Source Similarity

Setup. The main analysis adjusts for source predictive divergence across its full observed range, so the remaining association could still be concentrated among learner pairs with relatively large values of $D^{\mathcal{S},q}$. To determine whether the association is also present among learners with more similar source predictions, we construct five strata separately for each dataset and class. For a fixed class q, the learner pairs are ordered by $D^{\mathcal{S},q}$ and divided into $\mathcal{Q}_1, \ldots, \mathcal{Q}_5$. $\mathcal{Q}_1$ contains the 20% with the closest source predictions, and $\mathcal{Q}_5$ contains the 20% with the largest source predictive divergence. For each $g \in \{1, \ldots, 5\}$, the within-stratum association is

$$r_{t,q,g} = \text{partial}\,\text{Spearman}_{(i,j)\in\mathcal{Q}_g}\left(A_{ij}^{\mathcal{S},q}, M_{ij,t}^{\mathcal{T},q} \mid D_{ij}^{\mathcal{S},q}\right).$$

The restriction $(i,j) \in \mathcal{Q}_g$ means that the ranks and both regressions in Eq. (103) are recomputed using only the learner pairs in the current stratum. Stratification narrows the observed range of $D^{\mathcal{S},q}$, and the within-stratum adjustment accounts for the variation that remains in that range. We compute the association separately for each target subject and aggregate the resulting estimates as in Eq. (104). During the model jackknife, the five strata are formed again after each learner is deleted.

Results. Within $\mathcal{Q}_1$, all four mean associations were positive, with pointwise model-jackknife intervals above zero (Figure 6 and Table 12). Estimates in $\mathcal{Q}_2$–$\mathcal{Q}_5$ were also positive, providing descriptive context across the full range of source predictive divergence. The $\mathcal{Q}_1$ result strengthens

the main analysis by showing that its positive association is retained among learner pairs with the closest measured source predictions after adjustment for the remaining source divergence within that stratum.

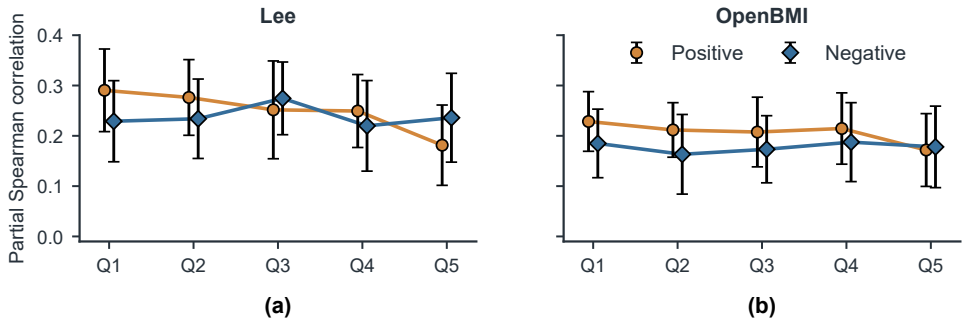

Figure 6: Association profiles across source-divergence strata. Points show mean target-wise partial Spearman correlations, and error bars show pointwise 95% model-jackknife intervals. The strata are constructed separately for each class and order learner pairs from lower source predictive divergence in $\mathcal{Q}_1$ to higher divergence in $\mathcal{Q}_5$.

Table 12: Estimates across source-divergence strata. Entries give mean target-wise partial Spearman correlations with pointwise 95% model-jackknife intervals. The strata are constructed separately for each class.

Dataset	Class	$\mathcal{Q}_1$	$\mathcal{Q}_2$	$\mathcal{Q}_3$	$\mathcal{Q}_4$	$\mathcal{Q}_5$
Lee	Positive	0.290 [0.208, 0.373]	0.276 [0.201, 0.351]	0.252 [0.154, 0.349]	0.249 [0.177, 0.322]	0.181 [0.102, 0.261]
Lee	Negative	0.229 [0.148, 0.310]	0.234 [0.155, 0.313]	0.274 [0.202, 0.347]	0.220 [0.130, 0.310]	0.236 [0.148, 0.324]
OpenBMI	Positive	0.229 [0.169, 0.288]	0.212 [0.157, 0.266]	0.208 [0.138, 0.277]	0.215 [0.144, 0.285]	0.172 [0.099, 0.244]
OpenBMI	Negative	0.185 [0.117, 0.253]	0.163 [0.084, 0.242]	0.173 [0.107, 0.240]	0.187 [0.109, 0.266]	0.178 [0.097, 0.259]

C.3.2 Testing the Component Specificity of Learner-Side Attribution

Setup. Section C.3.1 shows that the association between A and M remains among learner pairs with similar source predictions. However, because A aggregates deletion-response differences across components, it may still act as a broad measure of learner dissimilarity. A component-specific interpretation requires the source attribution difference for component k to be more closely related to target divergence after deleting the matching component than after deleting another component.

For a source component k, we isolate its contribution to the aggregate attribution difference,

$$A_{ij,k}^{\mathcal{S},q} = \frac{1}{2}\mathbb{E}_{\mathcal{S},q}\left|\frac{c_{i,k}^{q}(\boldsymbol{X})}{s_i^q} - \frac{c_{j,k}^{q}(\boldsymbol{X})}{s_j^q}\right|, \qquad A_{ij}^{\mathcal{S},q} = \sum_{k=1}^{4} A_{ij,k}^{\mathcal{S},q}. \tag{106}$$

This quantity measures how differently learners i and j respond to deletion of source component k, using the normalization of the main attribution measure. When the learner pair and class are clear, we write this component term as A_k. We then relate it to the target divergence remaining after component ℓ is removed. Using the shared reconstruction defined in Section C.1.3, we compute

$$M_{ij,t}^{(-\ell),q} = \frac{1}{|\mathcal{T}_t^q|}\sum_{\boldsymbol{X}\in\mathcal{T}_t^q} \mathrm{JS}(\boldsymbol{\pi}_i(\boldsymbol{X}-\boldsymbol{X}_\ell), \boldsymbol{\pi}_j(\boldsymbol{X}-\boldsymbol{X}_\ell))\,. \tag{107}$$

If the main association contains component-specific structure, the relation between $A_{ij,k}^{\mathcal{S},q}$ and the remaining target divergence should depend on which component is removed. Across learner pairs,

$A^{\mathcal{S},q}_{ij,k}$ should therefore have a stronger negative rank association with $M^{(-k),q}_{ij,t}$ than with $M^{(-\ell),q}_{ij,t}$ for $\ell \neq k$. Two DG learners with larger source prediction differences may also differ more in their source component responses and target predictions. In addition, pairs with greater full-input target divergence may remain more divergent after any component is removed. We therefore adjust for both the source predictive divergence $D^{\mathcal{S},q}_{ij}$ and the full-input target divergence $M^{\mathrm{full},q}_{ij,t} = M^{\mathcal{T},q}_{ij,t}$:

$$S_{t,q,k\ell} = -\operatorname{partial\,Spearman}\left(A^{\mathcal{S},q}_{ij,k}, M^{(-\ell),q}_{ij,t} \mid M^{\mathrm{full},q}_{ij,t}, D^{\mathcal{S},q}_{ij}\right). \tag{108}$$

The leading negative sign makes the expected negative association positive. Thus, a positive $S_{t,q,k\ell}$ means that learner pairs with larger source attribution differences for component k tend to retain less target divergence after component ℓ is deleted, conditional on the two controls. The diagonal entries $S_{t,q,kk}$ relate each source component to deletion of the same target component, whereas the off-diagonal entries relate it to deletion of another component. We summarize their difference by

$$\Gamma_{t,q} = \frac{1}{4}\sum_{k=1}^{4} S_{t,q,kk} - \frac{1}{12}\sum_{k\neq\ell} S_{t,q,k\ell}. \tag{109}$$

Positive $\Gamma_{t,q}$ indicates stronger average correspondence between matching component coordinates. Target subjects receive equal weight in the class-level summary,

$$\bar{\Gamma}_q = \frac{1}{n_{\mathcal{T}}}\sum_{t=1}^{n_{\mathcal{T}}} \Gamma_{t,q}.$$

The model jackknife repeats the rank regressions, the component correspondence scores, and their aggregation after each learner is deleted.

Component mapping check. The primary contrast treats A_k and deletion of $\boldsymbol{X}_k$ as a match because both refer to coordinate k in the shared xDAWN decomposition. To determine whether the diagonal advantage follows this predefined correspondence, we compare it with every other one-to-one assignment of the four source attribution terms to the four target deletions. Let σ denote one of the $4! = 24$ permutations of the deletion indices. For target subject t, the contrast under mapping σ is

$$\Gamma^{(\sigma)}_{t,q} = \frac{1}{4}\sum_{k=1}^{4} S_{t,q,k,\sigma(k)} - \frac{1}{12}\sum_{k=1}^{4}\sum_{\ell\neq\sigma(k)} S_{t,q,k\ell}, \qquad \bar{\Gamma}^{(\sigma)}_{q} = \frac{1}{n_{\mathcal{T}}}\sum_{t=1}^{n_{\mathcal{T}}} \Gamma^{(\sigma)}_{t,q}.$$

The same-index mapping is $\sigma_{\mathrm{id}}(k) = k$. We rank $\bar{\Gamma}^{(\sigma_{\mathrm{id}})}_{q}$ among the 24 target-averaged contrasts and, for each target subject, rank $\Gamma^{(\sigma_{\mathrm{id}})}_{t,q}$ among the corresponding 24 target-wise contrasts. Table 15 reports the target-averaged rank and the number of target subjects for which the same-index mapping ranks first. A rank closer to 1 indicates that the observed correspondence follows the shared component coordinates more closely than the alternative assignments.

Deletion effect check. The primary analysis uses $M^{(-\ell),q}_{ij,t}$, which measures how much predictive divergence remains after component ℓ is deleted. This remaining divergence combines the learner pair's full-input disagreement with the change produced by deletion. We therefore examine the change itself to determine whether source attribution differences are related to the target prediction effect of removing the corresponding component. For each learner pair, we define

$$\Delta M^{q}_{ij,t,\ell} = M^{\mathrm{full},q}_{ij,t} - M^{(-\ell),q}_{ij,t}.$$

Positive $\Delta M^{q}_{ij,t,\ell}$ means that deleting $\boldsymbol{X}_\ell$ reduces the predictive divergence between learners i and j. Because the full-input divergence provides the baseline within this difference, the descriptive partial correlation adjusts for source predictive divergence $D^{\mathcal{S},q}_{ij}$:

$$S^{\Delta}_{t,q,k\ell} = \operatorname{partial\,Spearman}\left(A^{\mathcal{S},q}_{ij,k}, \Delta M^{q}_{ij,t,\ell} \mid D^{\mathcal{S},q}_{ij}\right).$$

Replacing $S_{t,q,k\ell}$ with $S^{\Delta}_{t,q,k\ell}$ in Eq. (109) gives $\Gamma^{\Delta}_{t,q}$, and equal averaging across target subjects gives

$$\bar{\Gamma}^{\Delta}_{q} = \frac{1}{n_{\mathcal{T}}}\sum_{t=1}^{n_{\mathcal{T}}} \Gamma^{\Delta}_{t,q}.$$

Agreement between $\bar{\Gamma}_q^{\Delta}$ and the primary $\bar{\Gamma}_q$ would extend the component correspondence from the divergence remaining after deletion to the prediction change produced by deletion itself. The coordinate-isolation check is reported in Table 21.

Results. Component correspondence was clear in both Lee classes, weaker in the OpenBMI positive class, and not stable in the OpenBMI negative class (Figure 7 and Tables 13 and 14). In Lee, all target-averaged diagonal scores were positive, the model-jackknife intervals for $\bar{\Gamma}_q$ were above zero, and $\Gamma_{t,q}$ was positive for every target subject. Learner pairs with larger source attribution differences for component k therefore tended to retain less target disagreement after the matching component $\boldsymbol{X}_k$ was removed, and this relation was stronger than the corresponding mismatched relations. OpenBMI positive retained a positive average contrast with an interval above zero, but one target-averaged diagonal score was negative and the target-wise contrasts were more heterogeneous. The correspondence therefore varied across components and target subjects. For OpenBMI negative, the interval included zero and the correspondence matrix showed no stable diagonal concentration, providing no stable component-specific evidence in this condition.

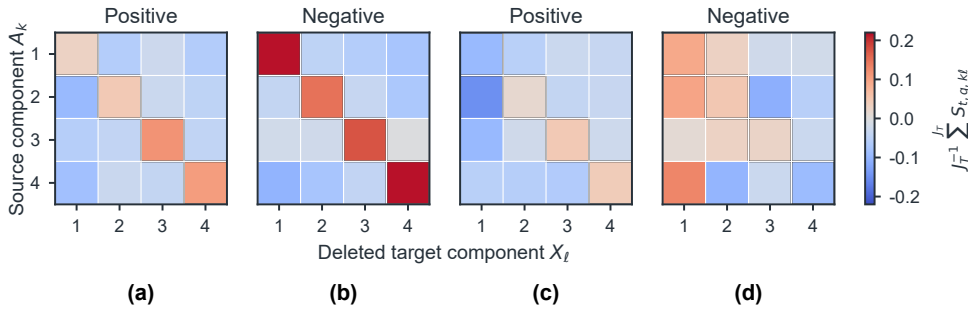

Figure 7: Component correspondence under target deletion. Panels (a)–(d) show Lee positive, Lee negative, OpenBMI positive, and OpenBMI negative, respectively. Rows identify the source attribution term for each component, and columns identify reconstructed target components removed before predictive divergence is measured. Each cell shows the component correspondence score averaged equally across target subjects. Diagonal cells match the same shared component coordinate. All panels use the same color scale.

Table 13: Target-averaged component specificity under target deletion. The diagonal advantage $\bar{\Gamma}_q$ is the equally weighted average of the subject-wise advantages $\Gamma_{t,q}$. Positive values indicate stronger correspondence, on average, for matching than for nonmatching source–target component pairs. Standard errors and pointwise 95% confidence intervals are obtained via a model jackknife.

Dataset	Class	$\bar{\Gamma}_q$	Jackknife SE	95% interval
Lee	Positive	0.1246	0.0116	[0.1015, 0.1476]
Lee	Negative	0.2371	0.0193	[0.1988, 0.2754]
OpenBMI	Positive	0.0564	0.0118	[0.0329, 0.0798]
OpenBMI	Negative	0.0244	0.0238	[−0.0227, 0.0716]

Component mapping check. Descriptively, the same-index mapping produced the strongest target-averaged contrast in both Lee classes and OpenBMI positive (Table 15). It also ranked first for almost every Lee target subject. For OpenBMI positive, the same-index mapping ranked first after target averaging but did so for fewer than half of the individual target subjects. Alternative mappings produced larger target-averaged contrasts in OpenBMI negative. This descriptive pattern indicates that the Lee diagonal advantage follows the shared component coordinates and provides weaker, more heterogeneous support for the same interpretation in OpenBMI positive.

Deletion effect check. The descriptive contrasts based on ΔM were close to the corresponding primary estimates (Table 16). They retained the clear correspondence in Lee, the weaker correspondence in OpenBMI positive, and the absence of stable correspondence in OpenBMI negative. The same

Table 14: Variation across target subjects in component specificity under target deletion. The median and range summarize the subject-wise diagonal advantage $\Gamma_{t,q}$ among the evaluated target subjects. Positive values indicate stronger correspondence for matching than for nonmatching source–target component pairs. The final column reports the number of subjects with $\Gamma_{t,q} > 0$ out of all evaluated target subjects.

Dataset	Class	Median	Range	$\Gamma_{t,q} > 0$
Lee	Positive	0.1213	[0.0249, 0.1745]	20 / 20
Lee	Negative	0.2516	[0.1338, 0.2970]	20 / 20
OpenBMI	Positive	0.0544	[−0.0468, 0.1322]	34 / 36
OpenBMI	Negative	0.0265	[−0.0455, 0.1103]	26 / 36

evidence hierarchy is therefore visible when the target effect is measured as the prediction change produced by deletion.

Table 15: Component mapping check for the predefined same-index assignment. The mapping pairs source attribution component k with deletion of target component k. For each of the $4! = 24$ one-to-one mappings, the diagonal advantage is averaged across target subjects; the same-index mapping is then ranked among these averages (rank 1 is largest). The final column reports the number of target subjects for whom this mapping ranks first among the 24 subject-wise contrasts, out of the total evaluated.

Dataset	Class	Rank of same-index mapping	Targets with same-index mapping ranked first
Lee	Positive	1 / 24	19 / 20
Lee	Negative	1 / 24	20 / 20
OpenBMI	Positive	1 / 24	16 / 36
OpenBMI	Negative	7 / 24	3 / 36

Table 16: Deletion-effect check of component correspondence. The target effect is expressed as the signed change in predictive divergence, $\Delta M = M^{\text{full}} - M^{(-\ell)}$. The table reports the target-averaged diagonal advantage $\bar{\Gamma}^{\Delta}_{q}$, the range of subject-wise advantages, and the number of subjects with $\Gamma^{\Delta}_{t,q} > 0$ out of the total evaluated. Positive diagonal advantages indicate that the partial Spearman association between source attribution differences and ΔM is larger for matching than for nonmatching source–target component pairs.

Dataset	Class	$\bar{\Gamma}^{\Delta}_{q}$	Target-wise range	$\Gamma^{\Delta}_{t,q} > 0$
Lee	Positive	0.1214	[0.0227, 0.1631]	20 / 20
Lee	Negative	0.2305	[0.1275, 0.2826]	20 / 20
OpenBMI	Positive	0.0572	[−0.0370, 0.1228]	34 / 36
OpenBMI	Negative	0.0248	[−0.0434, 0.1068]	26 / 36

Sections C.3.1 and C.3.2 strengthen the main result in two complementary ways. The association remains positive among learner pairs with similar source predictions. Both Lee classes show clear correspondence with the shared ERP component coordinates, while OpenBMI positive provides weaker and more heterogeneous support for the same pattern. These component-level results connect the main association to how learners organize predictive activity across the shared ERP components, adding evidence beyond differences in their overall source predictions.

C.4 Additional Experimental Details

Tables 17–21 collect the learner performance, training procedure, network architecture, component settings, and numerical checks referenced in the preceding sections.

Table 17: Learner performance. Values are the median and range of balanced accuracy across the 100 learners. Source performance is computed from the pooled held-out source-validation trials. Target performance is computed separately for each target subject and then averaged equally across subjects within each learner. It is reported descriptively and was not used for training or checkpoint selection.

Dataset	Source validation BAcc (%)	Full target BAcc (%)
Lee	71.13 [69.94, 72.44]	72.78 [72.13, 73.28]
OpenBMI	87.62 [86.70, 88.34]	83.63 [83.12, 84.08]

Table 18: Training and evaluation settings. All 100 learners within each dataset use these settings and differ only in their random seed.

Setting	Value
Objective	Cross-entropy
Minibatch	32 trials, with 16 per class sampled with replacement
Updates per epoch	300 for Lee; 878 for OpenBMI
Optimizer	AdamW with initial learning rate 10^{-3}
Schedule and duration	Cosine annealing over 200 epochs
Regularization	Zero weight decay; gradient-norm clipping at 4
Initialization	Xavier-uniform convolutional and linear weights; zero classifier bias; BatchNorm scale 1 and offset 0
Checkpoint	Final epoch
Evaluation state	Model parameters and BatchNorm running statistics fixed; dropout disabled

Table 19: EEGNet architecture. The same architecture is used for both datasets. Here, n_{ch} is 31 for Lee and 62 for OpenBMI, and output dimensions exclude the batch dimension.

Block	Operations	Output dimensions
Input	Standardized EEG epoch	$1 \times n_{\mathrm{ch}} \times 102$
Temporal filtering	Temporal convolution with 8 filters, kernel 1×32, padding 16, and no bias; BatchNorm	$8 \times n_{\mathrm{ch}} \times 103$
Spatial filtering	Depthwise convolution with kernel $n_{\mathrm{ch}} \times 1$, 8 groups, depth multiplier 2, maximum weight norm 1, and no bias; BatchNorm; ELU; average pooling 1×4; dropout 0.25	$16 \times 1 \times 25$
Separable temporal filtering	Depthwise convolution with kernel 1×16, padding 8, 16 groups, and no bias; 1×1 pointwise convolution with 16 outputs and no bias; BatchNorm; ELU; average pooling 1×8; dropout 0.25	$16 \times 1 \times 3$
Classifier	Flattening followed by a linear map from 48 features to two logits	2

All BatchNorm layers use affine parameters, momentum 0.01, and $\epsilon = 10^{-3}$. The even-length temporal kernels use symmetric padding, which adds one temporal position before the corresponding pooling operation and accounts for the intermediate lengths of 103 and 26.

Table 20: Preprocessing and component settings.

Stage	Setting	Value
Preprocessed input	Epoch dimensions, channels $\times$ time samples	31×102 for Lee; 62×102 for OpenBMI
Preprocessed input	Denominator floor for within-trial channel standardization	10^{-6}
xDAWN	Implementation	`mne.preprocessing.Xdawn` in MNE-Python
xDAWN	Fit sample	Positive source-fit trials; one decomposition per dataset, shared across learners, classes, and target subjects
xDAWN	Retained components	4
xDAWN	Covariance regularization	Fixed shrinkage $\lambda = 0.001$ applied to both covariance estimates
xDAWN	Overlap correction	Disabled
Haufe mapping	Channel covariance used to construct the forward patterns	Sample covariance of centered positive source-fit observations pooled across trials and time, with divisor $N_+T - 1$
Haufe mapping	Pseudoinverse relative cutoff	10^{-10}
Attribution normalization	Validity criterion for s_i^q	Finite and greater than 10^{-10}

Table 21: Reproducibility checks. Checkpoint replay covers all 100 learners in each dataset. Errors for $A = \sum_k A_k$ are maxima over learner pairs and classes. Errors for the full-input M comparison are maxima over learner pairs, target subjects, and classes. The regression count includes one design for each target subject and class.

Check	Lee	OpenBMI
Checkpoints replayed without model-state changes	100 / 100	100 / 100
Maximum absolute error in $A = \sum_k A_k$	3.89×10^{-15}	6.66×10^{-15}
Maximum absolute difference in full-input M between the main analysis and Section C.3.2	1.25×10^{-15}	1.87×10^{-15}
Full-rank two-control designs	40 / 40	72 / 72
Rank of the retained component covariance	4	4
Condition number of the retained component covariance	2.03	3.84
Coordinate-isolation error $\|\boldsymbol{\Omega P} - \boldsymbol{I}_4\|_F$	1.04×10^{-15}	1.69×10^{-15}

The recomputed A and full-input M values agreed with the main analysis to numerical precision. Because $\boldsymbol{\Omega P}$ equals $\boldsymbol{I}_4$ to numerical precision, subtracting $\boldsymbol{X}_k$ removes coordinate k while leaving the other retained coordinates unchanged.

D Relation to Other Identifiability Problems in Domain Adaptation

Prior identifiability work most closely related to ours asks which cross-domain map or latent structure must be recovered for target prediction to become identifiable. Gulrajani & Hashimoto (2022) locate the unresolved object in the cross-domain correspondence. Multiple admissible maps can match the observed domain marginals while inducing different target labelings, and they derive conditions under which admissible linear maps are uniquely determined. Kong et al. (2022) shift the unresolved object to the latent generative structure. By partitioning the latent space into invariant and changing components and exploiting variation across domains, they establish component-wise identifiability of the changing components and block-wise identifiability of the invariant subspace, and use these results to identify the target joint distribution under their generative assumptions. Ng et al. (2025) expose a finer-grained structural ambiguity. They show that identifying a predictive latent subspace corresponding to the label's Markov blanket can remain too coarse. They further separate the label's parents, children, and spouses and, together with additional assumptions on cross-domain changes, establish identifiability of the target predictive distribution.

Viewed together, these works address the identifiability of transfer-relevant structure. Their route to target prediction proceeds by resolving uncertainty about a cross-domain map, latent partition, or structural role. In contrast, our analysis starts from a different premise. The component-specific cross-domain transformation laws and their correspondence to component types are fixed and exactly known, and we ask whether this transformation information is sufficient to determine the target decision rule. In general, it is not. The same Oracle information can remain compatible with different allocations of source-predictive variation across the fixed component types and hence with different target Bayes rules. If these restrictions were relaxed, ambiguity in the cross-domain map, latent partition, or structural role could coexist with attribution ambiguity; the Oracle setting fixes them precisely to isolate the residual question. A further boundary arises from the forward observation model. The exact source-equivalent reallocation underlying our witness is enabled by the additive, many-to-one aggregation of latent component activity, whereas the latent-variable frameworks of Kong et al. (2022) and Ng et al. (2025) assume invertible latent-to-observation mappings. Attribution ambiguity therefore concerns a residual decision-relevant uncertainty under known cross-domain transformation structure, rather than the recovery of an unidentified transfer structure. This separation is reflected in both directions of our theory. Our negative results show that known transformation structure need not determine the target decision rule, while our positive results in the restricted ERP subclass show that the target decision can be recovered without full latent-attribution identification.