

Optimal Inference with Black-box Predictions

Lucas Kania^{†,*} Abhinav Chakraborty^{◇,*}
Edward Kennedy[†] Larry Wasserman^{†,‡} Sivaraman Balakrishnan^{†,‡}

[†]Department of Statistics and Data Science, Carnegie Mellon University

[◇]Department of Statistics, Columbia University

[‡]Machine Learning Department, Carnegie Mellon University

lucaskania@cmu.edu, ac4662@columbia.edu, {edward,larry,siva}@stat.cmu.edu

August 3, 2026

Abstract

Powerful black-box predictive models have motivated many proposals for combining observed data with predictions to perform valid statistical inference. Despite this progress, the field lacks a unifying principle that explains how hypothesis tests should integrate data and predictions in a way that is both valid and efficient. In this work, we address this gap in the high-dimensional Gaussian sequence model. We characterize the information-theoretic limits of inference with black-box predictions when their accuracies are known and, for orthogonal predictions, when they are unknown. Building on these characterizations, we develop practical hypothesis tests that adapt to the unknown accuracies of the predictions while benefiting from strong alignment among them.

Table of contents

1	Introduction	2
2	Adaptation in the minimax framework	13
3	Oracle inference with multiple predictions	14
4	The cost of adapting to orthogonal predictions	18
5	The cost of adapting to arbitrary predictions	21
6	Simulations	24
7	Discussion	26
	Supplementary Material	28
	References	50

*These authors contributed equally.

1 Introduction

Modern scientific research often centers on evaluating whether available data supports or contradicts a given hypothesis. Statistical hypothesis testing provides a structured framework for addressing this question. At a high level, practitioners specify a reference distribution and apply a test to determine whether the data deviates from this baseline. Each test, either implicitly or explicitly, targets a specific kind of deviation and is designed to detect whether the data aligns more closely with an alternative distribution.

Early developments in parametric and semi-parametric statistics emphasized the design of tests with strong detection power against a small set of well-specified alternative distributions [Bickel et al., 2006, Janssen, 2000]. However, even minor deviations from the assumed alternative distributions can substantially reduce power [Breiman, 2001]. To overcome this limitation, researchers developed nonparametric tests that guarantee detection power across broad classes of alternatives. Yet, spreading power across infinitely many possibilities often weakens performance.

In practice, effective testing strategies should strike a balance between these extremes: concentrating power on alternatives that the data suggests are plausible, while spreading power broadly when no strong alternative is apparent. One promising direction involves assumption-lean methods that leverage black-box predictions to concentrate power on data-driven alternatives adaptively.

Notable examples include prediction-powered inference [Angelopoulos et al., 2023, Zrnic and Candès, 2024], dimension-agnostic inference for M-estimators [Kim and Ramdas, 2024, Takatsu and Kuchibhotla, 2025], and universal inference [Wasserman et al., 2020]. These methods construct tests that integrate observed data with black-box approximations of the data distribution. Many rely on likelihood-ratio tests, either directly or via classifier-based approximations, as in likelihood-free inference [Dalmaso et al., 2023, Gerber et al., 2023]. However, there is no consensus on how to best fuse black-box predictions with empirical data.

In this work, we study how tests can optimally integrate black-box predictions of the data distribution with observed data in the Gaussian sequence model. Most importantly, we focus on test procedures that adapt to the unknown quality of the predictions. These adaptive tests automatically adjust their power allocation: they concentrate power on a narrow range of alternative distributions when the predictions are informative, and distribute it more broadly when the predictions are unreliable.

Outline. Section 1.1 summarizes our main results. Section 2 develops the minimax framework that we use to study adaptation to predictions with unknown quality. Section 3 characterizes the optimal performance of an oracle that knows the accuracy of each prediction. Section 4 then derives an optimal test that adapts to the unknown accuracy of orthogonal predictions. Section 5 extends this approach to arbitrary predictions. Finally, Section 6 illustrates our results through simulations, and Section 7 places our results in the broader context of the field and identifies directions for future research.

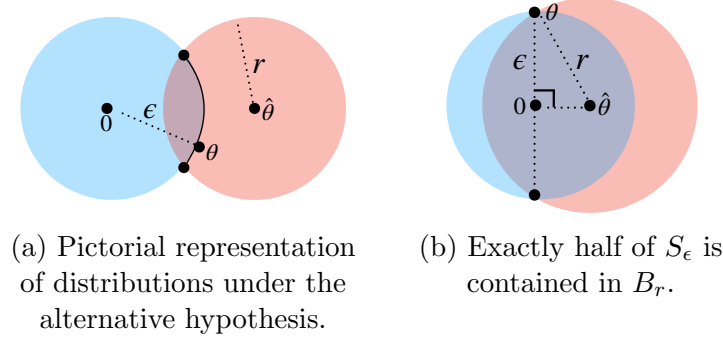


Figure 1: Illustrative scenarios under the alternative hypothesis.

Notation. We write $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \leq C \cdot b_n$ for all sufficiently large n . Analogously, $a_n \asymp b_n$ denotes that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We use $[k] = \{1, \dots, k\}$ to denote the set of the first k natural numbers. Furthermore, we use the notation $I(\text{condition})$ to denote the indicator function that equals 1 if the condition is true and 0 otherwise. Whenever clear from context, we omit stating the random variable in expectations and variances: $E_P[T] = E_{X \sim P}[T(X)]$ and $V_P[T] = V_{X \sim P}[T(X)]$. Furthermore, let $B_2^d(a, r)$ denote the d -dimensional Euclidean ball of radius r centered at $a \in \mathbb{R}^d$: $B_2^d(a, r) = \{\theta \in \mathbb{R}^d : \|\theta - a\|_2 \leq r\}$, and $S_2^{d-1}(a, r)$ denote the surface of the sphere of radius r centered at $a \in \mathbb{R}^d$: $S_2^{d-1}(a, r) = \{\theta \in \mathbb{R}^d : \|\theta - a\|_2 = r\}$. Throughout the paper, we use the inequality sign " $\geq$ " in different contexts. Namely, for vectors $a, b \in \mathbb{R}^d$, we state that $a \geq b$ if the inequality is satisfied entry-wise, while for functions f, g whose domains are the reals, we state that $f \geq g$ if $f(r) \geq g(r)$ for all $r \in \mathbb{R}$. Finally, we let $\text{CS}(a, b)$ denote the cosine similarity between vectors $a, b \in \mathbb{R}^d$, that is, $\text{CS}(a, b) = \langle a, b \rangle / (\|a\|_2 \cdot \|b\|_2)$ if $a \neq 0$ and $b \neq 0$, where $\langle a, b \rangle = a^T b$ is the inner product between the vectors. If either of the vectors is zero, we let $\text{CS}(a, b) = 0$.

1.1 Overview of main results

We study how to optimally aggregate data with auxiliary black-box predictions in the Gaussian sequence model. This model provides a canonical setting for minimax hypothesis testing and a useful benchmark for more general nonparametric regression problems [Ingster and Suslina, 2003, Johnstone, 2019].

In this setting, one is given access to n observations from a standard d -dimensional Gaussian distribution $X_1, \dots, X_n \sim \mathcal{N}(\theta, I_d)$ with some unknown mean $\theta \in \mathbb{R}^d$, and m predictions of the mean $\hat{\theta}_1, \dots, \hat{\theta}_m$ whose accuracy is unknown. The goal is to detect whether the mean of the data is zero ($\theta = 0$), or far from it. In this work,

- We characterize the optimal aggregation strategy of an oracle test that has access to the unknown prediction accuracies (Theorems 1 and 2).
- We develop tests that adapt to unknown accuracy of the predictions and charac-

terize optimal aggregation strategies when predictions are orthogonal (Theorems 3 and 4).

- We show that orthogonal predictions constitute a worst case from the perspective of the geometry of the predictions, and develop adaptive tests that aggregate data and predictions while benefiting from high cosine similarity among the predictions (Theorem 5). This case is practically relevant because predictions often come from the same black-box predictive model and are not orthogonal.

Finally, we remark that our results show that although functional estimation requires access to high-accuracy predictions, testing can use lower-accuracy predictions insofar as they have high cosine similarity with the mean of the underlying data distribution.

Henceforth, we let X denote the average of the n observations: $X \sim \mathcal{N}(\theta, I_d/n)$. Let us first consider the case where we do not have access to any black-box prediction. That is, one must decide if $\theta = 0$ or if it is far away from the origin as measured by the ℓ_2 norm:

$$H_0 : \theta = 0 \text{ versus } H_1 : \|\theta\|_2 \geq \epsilon. \quad (1)$$

It has been known since the work of Ingster [1993] that the hypotheses can be distinguished with nontrivial power if and only if $n \gtrsim \sqrt{d} \cdot \epsilon^{-2}$, which we call the prediction-free rate. The optimal test is based on the chi-squared test statistic:

$$\text{reject } H_0 \text{ if } \|X\|_2^2 \text{ is large,} \quad (2)$$

which intuitively indicates that the strong dependence on the dimension of the data arises due to the optimal test checking for deviations from the null hypothesis in all directions.

In this work, we show that arbitrary predictions can substantially improve the prediction-free rate without requiring strong assumptions about their quality. Informative predictions help the test focus on a smaller set of directions where the null distribution may fail, which increases power when the predictions are reliable. Poor predictions, however, can point the test in unhelpful directions and reduce its effectiveness. We address this tension with tests that adapt to the unknown accuracy of the predictions, which we call *adaptive tests*. As a result, the tests improve over the prediction-free rate when predictions are accurate, while still recovering that rate when predictions are uninformative.

Throughout this work, we consider black-box predictions as fixed non-random objects, which is commonplace in the model selection literature [Juditsky and Nemirovski, 2000, Nemirovski, 2000, Tsybakov, 2003]. Alternatively, one can adopt a conditional point of view [Balakrishnan et al., 2023], in which predictions are obtained from an independent sample of the data. Regardless of the chosen point of view, the focus of this work is on instance-wise optimality, that is, the optimal use of given fixed predictions with an unknown fixed accuracy. That is, we do not exploit any statistical guarantees of the process that generates the predictions, since such guarantees are rarely available in practice.

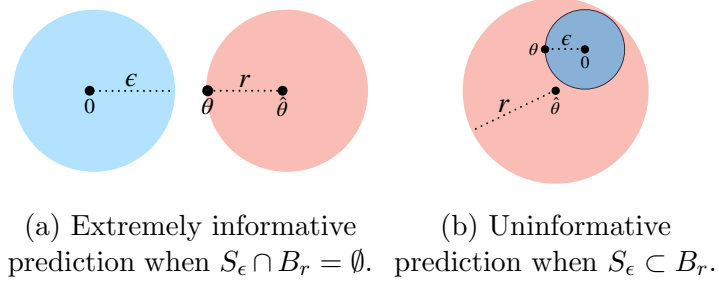


Figure 2: Illustrative scenarios under the alternative hypothesis.

To understand how a prediction can improve inference, let us first consider an oracle setting where we have access to a single black-box prediction $\hat{\theta} \in \mathbb{R}^d$, and we know its accuracy: $\|\theta - \hat{\theta}\|_2 \leq r$. We can incorporate the additional information in the hypotheses (1) as follows:

$$H_0 : \theta = 0, \text{ versus } H_1 : \|\theta\|_2 \geq \epsilon \text{ and } \|\theta - \hat{\theta}\|_2 \leq r. \quad (3)$$

Note that the accuracy of the prediction is not used in the definition of the null hypothesis, which implies that a test must be valid regardless of the quality of a black-box prediction. This setting parallels the attractive feature of methods such as universal or prediction-powered inference [Wasserman et al., 2020, Angelopoulos et al., 2023], whose validity does not depend on the quality of the predictions. Nevertheless, note that the prediction plays a key role in localizing the alternative hypothesis, which implies that the power of any test that distinguishes the hypotheses (3) must be affected by the accuracy of the prediction.

To intuitively understand how a prediction might change the hardness of the testing problem, let us consider an adversary that attempts to confuse a tester by choosing data distributions under the alternative hypothesis that are as close as possible to the null distribution. Going back to the alternative hypothesis in (1), without a prediction, an adversary may choose any point on the surface of the sphere $S_2^{d-1}(0, \epsilon)$, denoted by S_ϵ henceforth. A test must therefore search for deviations from the null hypothesis in every direction on this sphere.

A prediction restricts this geometry. Under the alternative hypothesis of (3), the mean of the data distribution must belong to the ball $B_2^d(\hat{\theta}, r)$, denoted by B_r henceforth. Thus, in the worst case, the adversary can only choose points in $S_\epsilon \cap B_r$, a spherical cap of S_ϵ , as shown in Figure 1a. As this cap shrinks, testing becomes easier because fewer distributions under the alternative hypothesis are closer to the null distribution. Thus, a tester must look into fewer directions to detect deviations from the null distribution.

Let us consider two extreme cases to further clarify the role of the prediction. If $S_\epsilon \subseteq B_r$, see Figure 2b, the prediction does not restrict the adversary. That is, hypotheses (3) reduce to (1), and we expect to attain the prediction-free rate. At the other extreme, if S_ϵ and B_r are disjoint, see Figure 2a, there is a unique point under the alternative hypothesis that is closest to the null distribution. Thus, one can attain parametric

rates by performing a likelihood ratio test between the null distribution and the unique closest distribution under the alternative [Le Cam, 1986]:

$$\theta_*(r) = \arg \inf_{\theta \in B_r} \|\theta\|_2^2. \quad (4)$$

In general, the hardness of distinguishing the hypotheses (3) must interpolate between these extreme cases. An oracle test allocates power according to the informativeness of the prediction. When the prediction is accurate, the test concentrates all the detection power along the predicted direction. Conversely, when the prediction is uninformative, the test spreads power across all directions. Intuitively, the key quantity governing this interpolation should be related to the amount of information that the prediction has about the data distribution. Henceforth, let $\hat{U} = \hat{\theta}/\|\hat{\theta}\|_2$ be the direction of the prediction, and let $\Pi_{\hat{U}}$ denote the projection onto the span of the prediction. We will let $\gamma_*(\epsilon, r)$ denote the length of the smallest projection onto the span of the prediction under the alternative hypothesis

$$\gamma_*^2(\epsilon, r) = \inf_{\theta \in S_\epsilon \cap B_r} \|\Pi_{\hat{U}} \theta\|_2^2. \quad (5)$$

The quantity is fundamentally linked to the amount of information about the data distribution that is contained in the projection onto the prediction's span. Henceforth, let $\text{CS}(\epsilon, r)$ denote the worst cosine similarity between the mean of the data θ and its projection onto the span of the prediction $\Pi_{\hat{U}} \theta$ under the alternative hypothesis:

$$\text{CS}(\epsilon, r) = \inf_{\theta \in S_\epsilon \cap B_r} \text{CS}(\Pi_{\hat{U}} \theta, \theta) = \frac{\gamma_*(\epsilon, r)}{\epsilon} \quad \text{if } S_\epsilon \cap B_r \neq \emptyset, \quad (6)$$

where we define the cosine similarity between two vectors a and b as $\text{CS}(a, b) = \langle a, b \rangle / (\|a\|_2 \cdot \|b\|_2)$. Thus, as $\text{CS}(\epsilon, r)$ increases, the projection onto the prediction's span becomes more informative of the mean of the data distribution. The following theorem summarizes the oracle performance, which, by virtue of knowing the accuracy of the prediction, can decide optimally when to utilize it.

Theorem 1. *Hypotheses (3) can be tested with nontrivial power if and only if*

$$n \gtrsim \begin{cases} \sqrt{d} \cdot \epsilon^{-2} & \text{if } \text{CS}(\epsilon, r) \leq c \cdot d^{-1/4} \text{ and } S_\epsilon \cap B_r \neq \emptyset. \\ \gamma_*^{-2}(\epsilon, r) & \text{if } \text{CS}(\epsilon, r) \in [c \cdot d^{-1/4}, 1] \text{ and } S_\epsilon \cap B_r \neq \emptyset. \\ \|\theta_*(r)\|_2^{-2} & \text{if } S_\epsilon \cap B_r = \emptyset. \end{cases}$$

where c is a positive constant. Furthermore, the rates depend continuously on the intersection $S_\epsilon \cap B_r$, and are ordered from worst to best, from top to bottom.

In broad strokes, Theorem 1 states that if the worst cosine similarity between the mean of the data distribution and its projection onto the prediction is low, then the prediction is not informative enough, and we obtain the prediction-free rate. Geometrically, the first case can be foreseen by noting that when $\text{CS}(\epsilon, r) \lesssim d^{-1/4}$, an adversary

is allowed to choose a data distribution under the alternative hypothesis whose mean is orthogonal to the prediction, see Figure 1b, making the prediction uninformative.

However, if $\text{CS}(\epsilon, r)$ is large enough, in the second case, the oracle can project the data onto the prediction $\Pi_{\hat{U}}X$ and make a decision based on this one-dimensional quantity, avoiding any dependence on the ambient dimension of the data. Finally, in the third case, when a prediction is extremely informative, S_ϵ and B_r are disjoint, and the testing problem reduces to the binary test between $H_0 : \theta = 0$ and $H_1 : \theta = \theta_*$, as defined in (4), which yields dimension-independent parametric rates. Hereafter, we will omit the case where S_ϵ and B_r do not intersect from the introduction, since it is related to the well-studied binary hypothesis testing problem and is relevant to practitioners only when predictions are extremely accurate.

Practitioners may have access to multiple predictions, each of which may contain useful information about the data distribution. To model this setting, we extend the hypotheses in (3) to multiple predictions. Let $r \in \mathbb{R}_+^m$ denote the vector whose entries quantify the accuracy of the m black-box predictions. We then define the hypotheses as follows:

$$H_0 : \theta = 0, \text{ versus } H_1 : \|\theta\|_2 \geq \epsilon(r) \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i \text{ for } i \in [m]. \quad (7)$$

Note that we have written explicitly ϵ as a function of the accuracy vector $r \in \mathbb{R}_+^m$, as one can intuit that the more accurate the predictions are, the smaller $\epsilon(r)$ should be. We call the function $r \mapsto \epsilon(r)$ a separation curve, as it quantifies the separation between the null and alternative hypotheses.

Henceforth, we let B_r denote the intersection of all prediction neighborhoods, $B_r = \cap_{i=1}^m B(\hat{\theta}_i, r_i)$, and let $\Pi_{\hat{U}}\theta$ in (5) and (6) denote the projection onto the span of the predictions, where $\hat{U}$ is a matrix in $\mathbb{R}^{d \times m}$, whose i -th column is given by the i -th predicted direction $\hat{U}_i = \hat{\theta}_i / \|\hat{\theta}_i\|_2$. Since the oracle knows the accuracy of all predictions, it can aggregate them optimally and perform a one-dimensional test by combining their projections:

$$\text{reject } H_0 \text{ if } \langle \Pi_{\hat{U}}X, W \rangle \text{ is large,} \quad (8)$$

where W is a vector of weights that depends on the vector of accuracies r and the separation curve. Thus, when predictions are informative, the performance of the oracle test is independent of the dimension of the data.

Theorem 2 (Optimal performance of oracle with access to arbitrary predictions). *If $\text{rank } \hat{U} \leq d/2$, the hypotheses (7) can be tested with nontrivial power if and only if*

$$\epsilon(r) \gtrsim \inf \left\{ \epsilon \geq 0 : n \gtrsim \frac{1}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \text{ and } S_\epsilon \cap B_r \neq \emptyset \right\}.$$

We remark that, albeit written succinctly, the rates in Theorem 2 match those in Theorem 1 when only one prediction is available. Furthermore, the interpretation remains the same as in the single-prediction setting. Under the alternative hypothesis, when the worst cosine similarity between the mean of the data distribution and its

projection onto the span of the predictions (6) is of order $o(d^{-1/4})$, then predictions do not offer any benefit, and the oracle test disregards them. We note that Theorems 1 and 2 follow from Lemmas 1 and 2 in Section 3, where we further discuss them.

Predictions of unknown quality. In practice, we do not know the accuracy of the predictions. Therefore, the main goal is to construct tests that do not require that information to operate optimally. We formalize this requirement through the adaptivity framework used in minimax theory [Ingster, 2000, Giné and Nickl, 2016], where one aims to distinguish the hypotheses

$$H_0 : \theta = 0 \text{ versus } H_1 : \exists r \in \mathbb{R}_+^m \text{ s.t. } \|\theta\|_2 \geq \epsilon(r) \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i \text{ for } i \in [m]. \quad (9)$$

Unlike the oracle setting (7), the alternative hypothesis in (9) does not reveal the accuracy of the predictions. A test for (9) must therefore adapt to this unknown accuracy, so we call it an *adaptive* test. The key question is the *cost of adaptation*: how many additional observations does the test require because it does not know the true prediction accuracy?

Orthogonal predictions. To illustrate this phenomenon, let us first consider the case of orthogonal predictions when $m \leq d$ to draw a comparison with Theorem 2. The oracle avoids any dimension dependence due to optimally aggregating predictions into a one-dimensional weighted projection (8) using the accuracies of the predictions. An adaptive test lacks this information, so it cannot perform the aggregation. However, since the predictions are orthogonal, it is reasonable for the adaptive test to monitor for deviations from the null distribution in all predicted directions:

$$\text{reject } H_0 \text{ if } \|\Pi_{\hat{U}} X\|_2^2 \text{ is large, where } \|\Pi_{\hat{U}} X\|_2^2 = \sum_{i=1}^m \|\Pi_{\hat{u}_i} X\|_2^2. \quad (10)$$

Note that all projections are equally weighted; thus, the test rejects whenever the data deviates from the null distribution on any of the predicted directions. Given that we hedge our bets in all predicted directions equally, our performance is expected to depend on the number of predictions.

Theorem 3 (Performance of adaptive test with access to orthogonal predictions). *When predictions are orthogonal, the hypotheses (9) can be tested with nontrivial power if*

$$\epsilon(r) \gtrsim \inf \left\{ \epsilon \geq 0 : n \gtrsim \frac{\sqrt{m}}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \text{ and } S_\epsilon \cap B_r \neq \emptyset \right\}.$$

Theorem 3 shows that the adaptive test requires more observations than the oracle test in Theorem 2. The ratio between these sample sizes scales as $\sqrt{m}$, which is called the *adaptivity gap* in the literature. This gap reflects the cost of not knowing in advance which prediction is useful. When only a single prediction is available, the adaptive and oracle tests have the same performance (rate-wise) because they make the same decision: whether to use the prediction. As m grows, however, the adaptive test must account for more candidate predictions. If these predictions are orthogonal, each new prediction

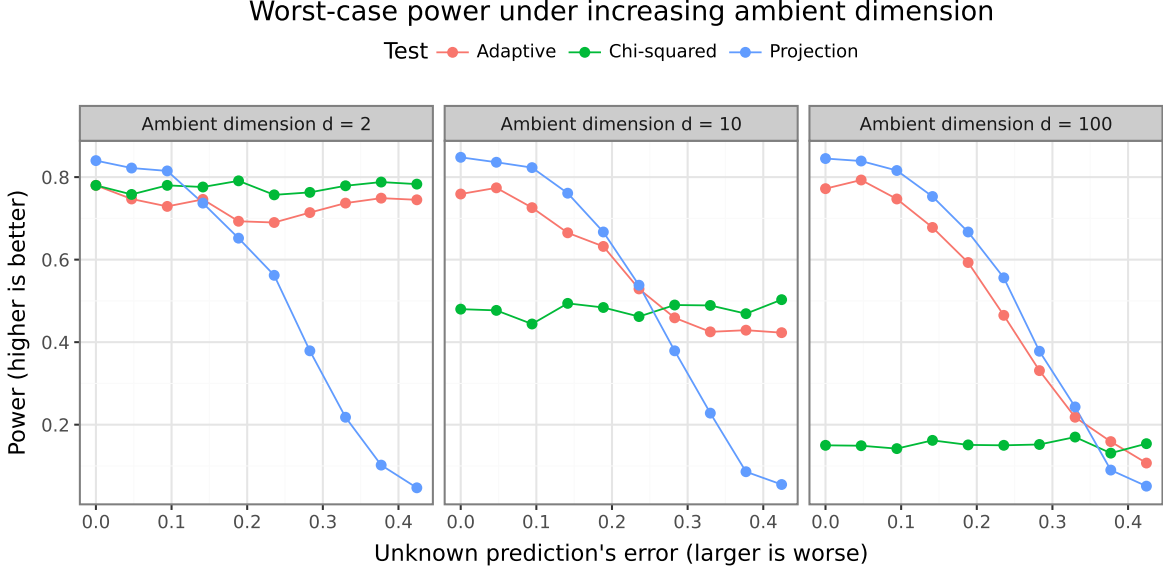


Figure 3: Comparison between the chi-squared test (16), which ignores the prediction, the projection test (24), which always uses the prediction, and the adaptive test (26) that adapts to the unknown prediction’s error.

increases the search space for deviations from the null hypothesis. Thus, having many orthogonal predictions can deteriorate the performance of the adaptive test rather than improve it.

Is there a free lunch? Nevertheless, comparing the performance of the adaptive test in Theorem 3 to the prediction-free rate achieved by the chi-squared test (2) that ignores predictions might lead one to believe that using predictions is always beneficial. This advantage is true only in a rate-wise sense. That is, if the predictions are high-quality, the adaptive test will be more sample-efficient than the chi-squared test. However, if the predictions are low-quality, the adaptive test’s sample efficiency will be at most a constant factor worse than that of the chi-squared test.

This phenomenon is exemplified by the simulation shown in Figure 3, where a single prediction is made available to each test. Details are deferred to Section 6; however, note that when the prediction is of high quality, the *adaptive test* substantially improves power relative to the chi-squared test that ignores the prediction. However, when the prediction is low-quality, the power of the adaptive test is worse than the chi-squared test’s power by a constant. Furthermore, note that when the prediction is high quality, the performance of the adaptive test is independent of the ambient dimension of the data.

Optimality of the adaptive test. Although one can improve Theorem 3 for any fixed $r \in \mathbb{R}_+^m$, the dependence on the number of predictions is optimal among separation curves satisfying a natural invariance property. Namely, looking at the separation curves for both oracle (Theorem 2) and adaptive (Theorem 3) tests, it is clear that they assign the same separation to accuracy vectors that induce projections of the same length

under the alternative hypothesis. The following property quantifies this idea.

Property (Property 1 in Section 4. Same separation for equivalent accuracy vectors). *Let $r \rightarrow \epsilon_*(r)$ be a separation curve. If $\epsilon_*(r) = \epsilon$ for $r \in \mathbb{R}_+^m$, it holds that $\epsilon_*(r') = \epsilon$ for all $r' \in \mathbb{R}_+^m$ such that $\gamma_*(\epsilon, r) = \gamma_*(\epsilon, r')$.*

In other words, the property states that accuracy vectors that lead to equally hard worst-case distributions under the alternative hypotheses get assigned the same separation. The next theorem shows that, under Property 1, the adaptivity gap must grow with the ambient dimension of the prediction's span for orthogonal predictions.

Theorem 4. *If $m \leq d/2$ and the predictions are orthogonal, for any separation curve that satisfies Property 1, the hypotheses (9) can be tested with nontrivial power only if*

$$\epsilon(r) \gtrsim \inf \left\{ \epsilon \geq 0 : n \gtrsim \frac{\sqrt{m}}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \text{ and } S_\epsilon \cap B_r \neq \emptyset \right\}.$$

The proof is based on a variant of Le Cam's random alternative distributions [Tsybakov, 2009], which must be supported on the appropriate spherical cap under the alternative hypothesis. These constructions differ substantially from those used in structure-agnostic functional estimation [Balakrishnan et al., 2023], and they explain the faster rates achieved by the testing procedures proposed in this work in comparison to the functional estimation rates.

Arbitrary predictions. Theorem 3 is most relevant when predictions come from different data sources, each of which provides independent information about the data distribution. In many applications, however, predictions are generated by the same model via resampling, and are not expected to be orthogonal.

A test such as (10) cannot be optimal when predictions are not orthogonal, because it assigns equal weight to every predicted direction. We should rather emphasize detecting deviations in directions most aligned with the majority of predicted directions, and essentially ignore those directions aligned with only a few predictions.

Henceforth, let $\lambda_1 \geq \dots \geq \lambda_k > 0$ denote the nonzero singular values of $\hat{U}$, and $\tilde{u}_1, \dots, \tilde{u}_k$ denote the corresponding orthonormal left singular vectors of $\hat{U}$. The squared singular values $\lambda_1^2 \geq \dots \geq \lambda_k^2 > 0$ are the nonzero eigenvalues of the predictions' cosine similarity matrix $\Sigma = \hat{U}^T \hat{U}$. Our adaptive test relies on an operation that we call the *intrinsic projection* of the data onto the span of the predictions:

$$\text{reject } H_0 \text{ if } \|\Pi_{\hat{U}}^{\text{Int}} X\|_2^2 \text{ is large, where } \|\Pi_{\hat{U}}^{\text{Int}} X\|_2^2 = \sum_{i=1}^k \frac{\lambda_i^2}{\lambda_1^2} \cdot \|\Pi_{\tilde{u}_i} X\|_2^2. \quad (11)$$

The weighting automatically downweights directions that align poorly with the predictions. As a result, the test focuses on fewer directions in which it must detect deviations from the null hypothesis. As mentioned in Section 1.1, the intrinsic projection test owes its name to its performance not depending on the ambient dimension of the span of the predictions, but on a notion of its intrinsic dimension.

We define the intrinsic dimension of Σ^2 as $\text{Int } \Sigma^2 = \text{Tr } \Sigma^2 / \|\Sigma^2\|_{\text{op}}$ where the numerator is the trace of Σ^2 and the denominator is its operator norm [Ipsen and Saibaba, 2024]. This quantity measures the effective degree of orthogonality among the predictions. For instance, the intrinsic dimension of Σ^2 equals 1 when all predictions are parallel, and equals m when they are orthogonal. In this work, we show that the performance of an adaptive test depends on the intrinsic dimension of Σ^2 rather than the ambient dimension of the span of the predictions.

Theorem 5 (Performance of adaptive test with access to arbitrary predictions.). *The hypotheses (9) can be tested with nontrivial power if*

$$\epsilon(r) \gtrsim \inf \left\{ \epsilon \geq 0 : n \gtrsim \frac{\sqrt{\text{Int } \Sigma^2}}{\tilde{\gamma}_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \text{ and } S_\epsilon \cap B_r \neq \emptyset \right\}$$

where $\tilde{\gamma}_*^2(\epsilon, r) = \inf_{\theta \in S_\epsilon \cap B_r} \|\Pi_{\hat{U}}^{\text{Int}} \theta\|_2^2$.

Note that Theorem 3 depends on the ambient dimension of the predictions' span ($\text{rank } \hat{U}$), whereas Theorem 5 depends on the intrinsic dimension of their squared cosine similarity matrix ($\text{Int } \Sigma^2$). It always holds that $\text{Int } \Sigma^2 \leq \text{rank } \hat{U}$, with equality being achieved when all predictions are orthogonal. Thus, orthogonal predictions can be considered a worst-case scenario for Theorem 5.

Under orthogonality, Theorem 5 recovers Theorem 3. Its main purpose, however, is to show that non-orthogonal predictions can improve adaptivity. When the predictions share substantial information, the adaptive test depends on the lower intrinsic dimension of the squared cosine similarity matrix rather than on the full ambient dimension of the predictions' span. This distinction matters in practice because predictions produced from the same predictive model often are not orthogonal, which can substantially reduce the intrinsic dimension. As an illustration, Figure 4 displays a simulation designed to be adversarial to the tests using intrinsic projection, where, nevertheless, if there is high cosine similarity among the predictions, adaptive methods based on intrinsic projection can outperform all competing methods. We defer the details to Section 6.

1.2 Related work

Our work is related to the literature on property testing and goodness-of-fit testing that leverage black-box predictions [Janková et al., 2020, Lundborg et al., 2024, Aliakbarpour et al., 2024]. Although some of them have also obtained minimax limits of inference for their respective settings, our work takes a step further and characterizes the cost of adapting to the unknown accuracy of the predictions.

Our results on adaptation draw on the framework of minimax adaptive hypothesis testing [Spokoiny, 1996, Ingster, 2000], which addresses the issue of testing procedures depending on parameters under the alternative hypothesis, such as function smoothness, that practitioners are unlikely to know in advance. Adaptive tests do not require

Worst-case power as the cosine similarity between predictions increases ($d=10, m=5$)

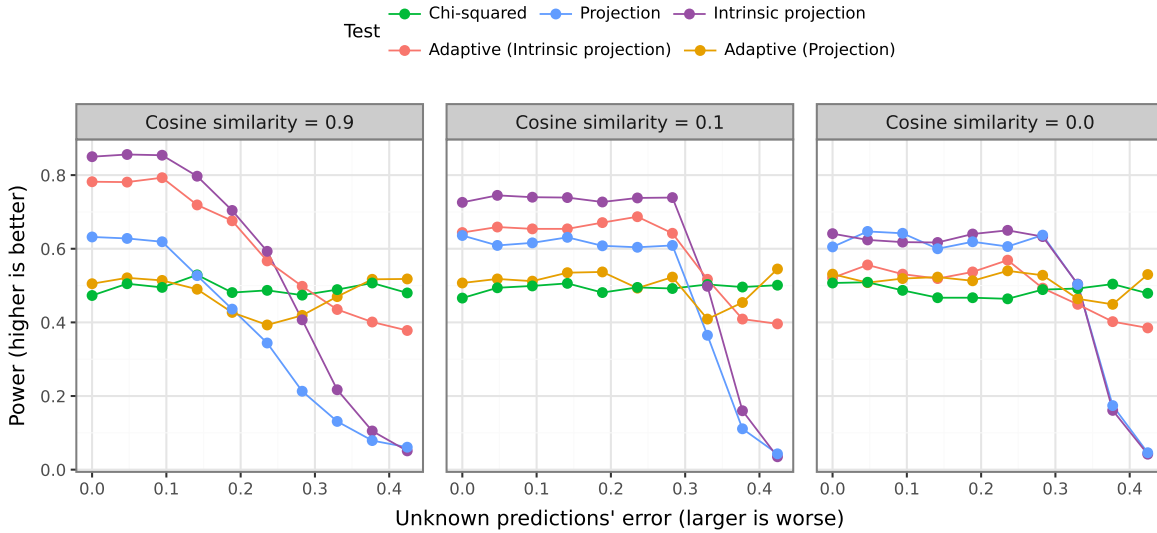


Figure 4: Comparison between the chi-squared test (16), which ignores predictions, tests that always use predictions either via standard (24) or intrinsic (31) projection, and adaptive tests that use standard (26) or intrinsic projection (32).

specifying said parameters, while often retaining nearly the same power as procedures that know them [Ingster and Suslina, 2003, Giné and Nickl, 2016]. In our setting, we show that orthogonal predictions break this usual guarantee. That is, the cost of adaptation is not negligible, and it can increase with the number of available predictions.

More broadly, this paper contributes to the literature on structure-agnostic inference. Prior work has developed this perspective for functional estimation [Balakrishnan et al., 2023, Gu, 2025, Jin and Syrgkanis, 2026], and extended the framework to causal inference [Jin et al., 2025, Jin and Syrgkanis, 2024, Bonvini et al., 2024]. We bring this viewpoint to hypothesis testing with black-box predictions and quantify the cost of adapting to unknown prediction quality.

We are also closely connected to prediction-powered inference and related semisupervised methods [Chakraborty and Cai, 2018, Angelopoulos et al., 2023, Chakraborty et al., 2024, Chakraborty and Dai, 2024, Zou et al., 2026, Xu and Sun, 2026]. In particular, our results support asymptotic arguments of previous works [Zhang et al., 2019, Angelopoulos et al., 2024] that indicated that prediction-based estimators must be aggregated with estimators that do not depend on predictions to control worst-case error while preserving efficiency.

There is also a related strand of work on model aggregation, initiated by Tsybakov [2003], which studies how to combine multiple candidate predictors or estimators to achieve near-oracle performance. This perspective was further developed through mirror-averaging and exponential-weighting methods [Juditsky et al., 2008, Dalalyan and Tsybakov, 2007] and later connected to sparse aggregation [Rigollet and Tsybakov, 2012]. Although this literature primarily studies prediction risk rather than hypothesis

testing, it is conceptually relevant because our procedures also aggregate information across multiple black-box predictions.

Finally, our work connects to a broader literature on statistical procedures that adapt to the different notions of intrinsic dimension or effective rank rather than the ambient dimension of the data. Existing results include asymptotic tests whose performance scales with the intrinsic dimension [Kim and Ramdas, 2024], as well as concentration bounds for self-adjoint operators and random matrices whose rates depend on the intrinsic dimension [Koltchinskii and Lounici, 2015, Wei and Minsker, 2017, Zhivotovskiy, 2024, Martinez-Taboada and Ramdas, 2026].

2 Adaptation in the minimax framework

Given data $X \sim P_\theta$ drawn from the distribution $P_\theta = \mathcal{N}(\theta, I_d/n)$ with unknown mean θ , we wish to test whether $\theta = 0$. The predictions may encode useful structural information about θ , and an effective procedure should exploit this information to increase power. At the same time, the validity of the test must not depend on the accuracy of the prediction. In particular, the testing procedure should remain well defined even when the prediction accuracy is unknown. In this section, we formalize these desiderata within a minimax framework.

Ultimately, we develop methods that optimally adapt to the unknown accuracy of the black-box prediction. However, to establish a baseline performance, let us first consider the case where the accuracy, denoted by r , is known. This setting is usually referred to as an *oracle* setting in the literature, since algorithms may use accuracy to make decisions. That is, we consider the hypotheses (7):

$$H_0 : P_\theta = P_0 \text{ versus } H_1 : P_\theta \in \mathcal{P}(\epsilon, r) \quad (12)$$

where $\mathcal{P}(\epsilon, r) = \left\{ P_\theta : \|\theta\|_2 \geq \epsilon(r) \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i \text{ for } i \in [m] \right\}$ is the localized set of alternative distributions with respect to accuracy vector $r \in \mathbb{R}_+^m$ and the predictions $\hat{\theta}_1, \dots, \hat{\theta}_m$. However, we omit the dependence on the predictions to simplify the notation throughout the paper. This localization may improve power by reducing the size of the alternative set, while the test's validity remains unaffected since the null hypothesis does not depend on the accuracy of the prediction.

We are interested in characterizing the hardness of distinguishing the hypotheses (12) with respect to the quality of the predictions. Let $\Psi_\alpha = \{\psi : P_0(\psi(X) = 1) \leq \alpha\}$ denote the set of all valid tests with type-I error at most α . For any valid test ψ , we define its risk under the alternative hypothesis as the worst-case type-II error:

$$R(\mathcal{P}(\epsilon, r), \psi) = \sup_{P \in \mathcal{P}(\epsilon, r)} P(\psi(X) = 0).$$

In particular, if a valid test ψ controls the worst-case type-II error $R(\mathcal{P}(\epsilon, r), \psi) < \beta$, for $\beta \in (0, 1 - \alpha)$, we say that ψ is powerful. The minimax risk is then the best performance

that we can achieve across all valid tests

$$R_*(\mathcal{P}(\epsilon, r)) = \inf_{\psi \in \Psi_\alpha} R(\mathcal{P}(\epsilon, r), \psi). \quad (13)$$

We can now define the *optimal oracle separation*, also known as critical oracle separation or minimax oracle separation, as the smallest separation between the hypotheses (12) for which there exists a valid and powerful test.

Definition 1 (Optimal oracle separation). *Let us fix $r \in \mathbb{R}_+^m$. We say that $\epsilon_*(r)$ is an oracle separation for hypotheses (12), if there exists a valid test that is powerful for distinguishing the null and alternative hypotheses that are $\epsilon_*(r)$ close. That is, $R_*(\mathcal{P}(\epsilon_*, r)) \leq \beta$.*

Furthermore, we say that ϵ_ is an optimal oracle separation, if no valid test can be powerful for hypotheses (12) if $\epsilon(r) < \epsilon_*(r)$. That is, $R_*(\mathcal{P}(\epsilon, r)) > \beta$ if $\epsilon(r) < \epsilon_*(r)$.*

We now turn to the realistic setting in which the testing algorithm does not know the prediction accuracy. Thus, the goal is to distinguish the hypotheses (9):

$$H_0 : P_\theta = P_0 \text{ versus } H_1 : \exists r \in \mathbb{R}_+^m \text{ s.t. } P_\theta \in \mathcal{P}(\epsilon, r). \quad (14)$$

Note that under the alternative hypothesis, there is no fixed accuracy vector, which accounts for the fact that in practice we do not know the accuracy of the predictions. Analogously to Definition 1, we define an *optimal separation curve* as the smallest separation between hypotheses (14) for which there exists a valid and powerful test.

Definition 2 (Optimal separation curve). *We say that ϵ_* is a separation curve for hypotheses (14), if there exists a valid test that distinguishes alternative hypotheses that are ϵ_* away. That is, $R_*(\mathcal{P}(\epsilon_*)) \leq \beta$ where $\mathcal{P}(\epsilon_*) = \bigcup_{r \in \mathbb{R}_+^m} \mathcal{P}(\epsilon_*, r)$.*

Furthermore, we say that ϵ_ is an optimal separation curve among a set of functions $\mathcal{E}$, if for any $\epsilon \in \mathcal{E}$ that significantly improves over ϵ_* , no valid test can be powerful for hypotheses (14). That is, if there exists a positive constant C that depends only on α and β such that for every $\epsilon \in \mathcal{E}$, if there exists $r \in \mathbb{R}_+^m$ satisfying $\epsilon(r) < C \cdot \epsilon_*(r)$, then $R_*(\mathcal{P}(\epsilon)) > \beta$ where $\mathcal{P}(\epsilon) = \bigcup_{r \in \mathbb{R}_+^m} \mathcal{P}(\epsilon, r)$.*

3 Oracle inference with multiple predictions

With a single prediction, there is no adaptivity gap. That is, a data-driven test can achieve the same rates as an oracle test that knows the predictions' accuracy. The structure of the problem forces a binary decision: either exploit the prediction with a projection test or ignore it with a chi-squared test. However, both tests must be used since the mean of the data distribution might be orthogonal to the prediction. Thus, the only advantage of the oracle test is knowing exactly when to run each test, which yields at most a constant-factor improvement in efficiency over non-oracle tests.

The picture changes when several predictions are available. Each additional prediction introduces a new direction along which deviations from the null hypothesis may occur. As a result, a test must determine not only whether to incorporate predictions, but also which ones to use to avoid diluting its power too much. An oracle, with knowledge of the accuracy of each prediction, can identify to what degree each prediction is informative and aggregate them optimally. A non-oracle test that lacks this information must hedge across all predicted directions and therefore cannot reliably achieve the same performance. As a result, oracle and non-oracle tests generally emphasize different directions in which to detect deviations from the null hypothesis, and this mismatch creates a non-constant performance gap.

Consider first the oracle setting, where under the alternative hypothesis we know the accuracies of the predictions:

$$H_0 : \theta = 0 \text{ versus } H_1 : \|\theta\|_2 \geq \epsilon \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i \text{ for } i \in [m], \quad (15)$$

Henceforth, let $B_r = \bigcap_{i=1}^m B_2^d(\hat{\theta}_i, r_i)$ denote the intersection of all the neighborhoods around the predictions. As commented in Section 1.1, an oracle can detect when the predictions are uninformative, which geometrically corresponds to $S_\epsilon \subset B_r$, in which case hypotheses (15) reduce to the well-studied hypotheses (1). In this case, the chi-squared test (2), which discards all predictions, is optimal. That is, we reject whenever the ℓ_2 norm of the data is larger than the expected deviations under the null hypothesis:

$$\psi_{\chi^2}^\alpha(X) = I\left(n \cdot \|X\|_2^2 \geq \chi_{d,1-\alpha}^2\right), \quad (16)$$

where $\chi_{d,1-\alpha}^2$ is the $(1 - \alpha)$ -quantile of the chi-squared distribution with d degrees of freedom.

Conversely, consider the case in which the predictions are highly informative; geometrically, this corresponds to the alternative hypothesis containing a single point closest to the null hypothesis. This happens whenever $S_\epsilon \cap B_r$ is empty or a singleton. In this case, using a likelihood-ratio test between $H_0 : \theta = 0$ and $H_1 : \theta = \theta_*(r)$, see (4), is optimal. The likelihood ratio test is equivalent to rejecting the null hypothesis whenever the projection of the data onto the direction of $\theta_*(r)$ is large enough:

$$\psi_{\text{LR}}^\alpha(X) = I\left(\sqrt{n} \cdot \left\langle X, \frac{\theta_*(r)}{\|\theta_*(r)\|_2} \right\rangle \geq z_{1-\alpha}\right). \quad (17)$$

In general, using projections onto the predicted directions is a sensible choice, since under the alternative hypothesis, each predicted direction can serve as a means of detecting deviations from the null hypothesis.

Remark 1. Let $h_i = \|\hat{\theta}_i\|_2$ be the length of the i -th prediction, and $\hat{u}_i = \hat{\theta}_i/h_i$ its direction. Define $A_i(\epsilon, r) = (\epsilon^2 + h_i^2 - r_i^2)/(2h_i)$. Then, under the alternative hypothesis (15), the length of the projection onto the i -th prediction is lower bounded: $\|\Pi_{\hat{u}_i}\theta\|_2^2 \geq A_{i,+}^2(\epsilon, r)$ where $A_{i,+}(\epsilon, r) = \max(0, A_i(\epsilon, r))$ for $i \in [m]$.

Given the geometric restrictions imposed by each prediction, it is not clear if the lower bounds in Remark 1 can be achieved. Henceforth, let γ_* be the length of the smallest projection onto the span of the predictions under the alternative hypothesis

$$\gamma_*^2(\epsilon, r) = \inf_{\theta \in B_r \cap S_\epsilon} \|\Pi_{\hat{U}} \theta\|_2^2 \quad \text{where } \hat{U} = [\hat{u}_1, \dots, \hat{u}_m], \quad (18)$$

and let $\text{rank } \hat{U} = k$. If $B_r \cap S_\epsilon \neq \emptyset$ and $k < d$, it holds that there exists a unique vector $A_*(\epsilon, r) \in \mathbb{R}^k$ that characterizes the projection of θ onto the span of the predictions, i.e., $\gamma_*^2(\epsilon, r) = \|A_*(\epsilon, r)\|_2^2$. In particular, if all predictions are orthogonal, it holds that $A_*(\epsilon, r) = A_+(\epsilon, r)$, meaning that the lower bounds in Remark 1 are achieved.

Furthermore, if $B_r \cap S_\epsilon \neq \emptyset$, hypotheses (15) can be reduced to testing the hypotheses:

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \in \arg \inf_{\theta \in S_\epsilon \cap B_r} \text{CS}(\Pi_{\hat{U}} \theta, \theta). \quad (19)$$

Therefore, to detect the alternative hypothesis, the oracle test only needs to measure how aligned $\Pi_{\hat{U}} X$ is with any element on the alternative hypothesis (19). This choice is immaterial because their projection onto the span of the predictions is fixed and given by $A_*(\epsilon, r)$. Thus, the natural test statistic for the oracle test is proportional to the plug-in estimator of the worst-case cosine similarity:

$$\psi_{\text{CS}}^\alpha(X) = I \left(\sqrt{n} \cdot \left\langle \Pi_{\hat{U}} X, \frac{\tilde{\theta}}{\|\tilde{\theta}\|_2} \right\rangle \geq z_{1-\alpha} \right) \quad \text{where } \tilde{\theta} = \tilde{U}_k A_*(\epsilon, r). \quad (20)$$

where $\tilde{U}_k = [\tilde{u}_1, \dots, \tilde{u}_k]$ are the k left singular vectors of $\hat{U}$ associated with the non-zero singular values.

In summary, if S_ϵ and B_r intersect, the oracle test runs the chi-squared test (16) whenever the minimum projection onto the span of the predictions is too small, and the worst-case cosine similarity test (20) otherwise. Alternatively, if S_ϵ and B_r do not intersect, a likelihood ratio test (17) suffices since there is a unique point under the alternative that is closest to the null hypothesis. We define our oracle test as follows:

$$\psi_{\text{Oracle}}(X) = \begin{cases} \psi_{\chi^2}^{\alpha/2}(X) \vee \psi_{\text{CS}}^{\alpha/2}(X) & \text{if } S_\epsilon \cap B_r \neq \emptyset \\ \psi_{\text{LR}}^\alpha(X) & \text{if } S_\epsilon \cap B_r = \emptyset \end{cases} \quad (21)$$

Henceforth, let B_ϵ° be the interior of the ball B_ϵ , i.e., $B_\epsilon^\circ = B_\epsilon \setminus S_\epsilon$. The following lemma summarizes the performance achieved by the oracle test. The proof is deferred to Appendix C.1 of the supplementary material.

Lemma 1 (Performance of oracle test with multiple predictions). *For hypotheses (15), the alternative hypothesis is non-vacuous if $B_r \not\subset B_\epsilon^\circ$. Furthermore, if $\text{rank } \hat{U} < d$, there exists a positive constant C that depends only on α and β such that the test (21) is powerful whenever*

$$n \geq C \cdot \begin{cases} \frac{1}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} & \text{if } S_\epsilon \cap B_r \neq \emptyset \\ \frac{1}{\|\theta_*(r)\|_2^2} & \text{if } S_\epsilon \cap B_r = \emptyset \end{cases}$$

We note that the rates are continuous since $\|\theta_*(r)\|_2^2 = \inf_{\theta \in B_r} \|\theta\|_2^2 = \inf_{\theta \in B_r} \|\Pi_{\hat{U}} \theta\|_2^2$ if S_ϵ and B_r do not intersect.

The first case shows that the oracle test achieves the prediction-free rate when predictions are uninformative, which has full dependence on the ambient dimension of the data. Conversely, when predictions are informative, the oracle test avoids any dependence on the ambient dimension of the data (d) or the dimension of the span of the predictions ($\text{rank } \hat{U}$). This improvement should come as no surprise, given that the oracle test relies on a one-dimensional projection that aggregates all predictions (20). The second case shows that when there is a single closest distribution under the alternative hypothesis, the oracle test achieves parametric rates.

3.1 Optimality of the oracle test

Let us first recall what is known regarding distinguishing the hypotheses (1), for which the chi-squared test (16) is optimal [Ingster, 1993]. An optimality proof follows from proving that no test can distinguish the null distribution P_0 from those distributions in the alternative hypotheses that are closest. Given that any point on the S_ϵ sphere is equally challenging to distinguish, we put a uniform measure on the sphere.

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta \sim \text{Uniform}[S_\epsilon]. \quad (22)$$

We can guarantee that the above construction is valid under hypotheses (15) whenever $S_\epsilon \subseteq B_r$, see Figure 2b. Thus, for big enough B_r , the lower bound from Ingster [1993] must hold.

Remark 2. *For hypotheses (15), there exists a positive constant C that depends only on α and β , such that no valid test can be powerful whenever*

$$n \leq C \cdot \frac{\sqrt{d}}{\epsilon^2} \text{ and } S_\epsilon \subseteq B_r.$$

Conversely, let us consider the case where the predictions are highly informative. Geometrically, this corresponds to the intersection of S_ϵ and B_r containing at most one point, implying that, under the alternative hypothesis, there is a unique point closest to the null hypothesis. Consequently, distinguishing the hypotheses (15) is equivalent to distinguishing the simple null hypothesis $H_0 : \theta = 0$ from the simple alternative hypothesis $H_1 : \theta = \theta_*(r)$, see (4).

Remark 3. *For hypotheses (15), there exists a positive constant C that depends only on α and β , such that no valid test can be powerful whenever*

$$n \leq C \cdot \frac{1}{\|\theta_*(r)\|_2^2} \text{ and } S_\epsilon \cap B_r \text{ is a singleton or the empty set.}$$

In summary, whenever the predictions are uninformative, we can use the unrestricted construction (22) since $S_\epsilon \subseteq B_r$, and consequently, we obtain the prediction-free rate by constructing a uniform measure on S_ϵ under the alternative hypothesis. Conversely, when the prediction is highly informative, we can use a simple binary hypothesis-testing construction, akin to assigning a *uniform measure* supported on a single point under the alternative.

For the general case, see Figure 1a, based on (22), a natural idea is to put a uniform measure on the spherical cap $S_\epsilon \cap B_r$. That is, using the alternative hypothesis $H_1 : \theta \sim \text{Uniform}[S_\epsilon \cap B_r]$, which interpolates the previously mentioned cases. Although that construction is valid, a sufficient approach is to focus only on those points of the spherical cap that minimize the length of the projection onto the span of the predictions (18). Namely, we look to distinguish

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta \sim \text{Uniform} \left[\{ \theta \in S_\epsilon \cap B_r : \|\Pi_{\hat{U}} \theta\|_2^2 = \gamma_*^2(\epsilon, r) \} \right].$$

The distributions under the alternative hypothesis have both the property of being as close as possible to the null hypothesis, and minimizing the amount of information that predictions provide on them. Based on this construction, we prove the following limit on the performance of any test that attempts to distinguish hypotheses (15). The proof of the following lemma is deferred to Appendix C.2 of the supplementary material.

Lemma 2. *For $\text{rank } \hat{U} \leq d/2$ and hypotheses (15), there exists a positive constant C that depends only on α and β , such that no valid test can be powerful whenever*

$$n \leq C \cdot \left[\frac{1}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \right] \text{ and } S_\epsilon \cap B_r \neq \emptyset$$

Together, Remarks 2 and 3, and Lemma 2 show that the performance of the oracle test in Lemma 1 is optimal up to constants. Furthermore, the remarks and lemmas 1 and 2 prove Theorems 1 and 2 in Section 1.1.

4 The cost of adapting to orthogonal predictions

When predictions are orthogonal, each prediction contains some new information about the underlying data distribution. From the perspective of the adaptivity gap, this is a worst-case scenario. While the oracle test can optimally aggregate the predictions and avoid any dependence on the dimension of the span, non-oracle procedures might have to consider all possible predictions since each one of them potentially contains information about the data distribution not available in the others.

As discussed in Section 2, we model the need to adapt to the unknown accuracies by requiring a test to distinguish between the hypotheses (15) for any accuracy vector. Formally, we seek a test that distinguishes

$$H_0 : \theta = 0 \text{ versus } H_1 : \exists r \in \mathbb{R}_+^m \text{ s.t. } \|\theta\|_2 \geq \epsilon(r) \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i \text{ for } i \in [m]. \quad (23)$$

where the predictions are assumed to be orthogonal. Note that the separation curve $r \mapsto \epsilon(r)$ also depends on the sample size n . However, since n remains fixed throughout this work, we omit this dependence from the notation.

We first approximate the projection statistic that drives the oracle test in (20). Under the alternative, by virtue of knowing the vector of accuracies, the oracle test can compute the worst cosine similarity (20). Without knowledge of the accuracy vector r , this information is not available. Nevertheless, we can still approximate the cosine similarity between $\Pi_{\hat{U}}\theta$ and θ by measuring the length of the projection of the data onto the span of the predictions:

$$\psi_{\text{Proj}}^\alpha(X) = I\left(n \cdot \|\Pi_{\hat{U}}X\|_2^2 \geq \chi_{\text{rank } \hat{U}, 1-\alpha}^2\right). \quad (24)$$

Analogously to the oracle test, we must also protect ourselves against the case where the predictions are uninformative. However, rather than running the chi-squared test (16), since $\|X\|_2^2 = \|\Pi_{\hat{U}}X\|_2^2 + \|\Pi_{\hat{U}^\perp}X\|_2^2$, we can avoid reusing the data by simply rejecting the null hypothesis whenever the length of the projection onto the orthogonal complement of the predictions' span is large:

$$\psi_{\text{Proj}^\perp}^\alpha(X) = I\left(n \cdot \|\Pi_{\hat{U}^\perp}X\|_2^2 \geq \chi_{d-\text{rank } \hat{U}, 1-\alpha}^2\right). \quad (25)$$

Thus, we might replace the oracle test (21) with the following Bonferroni aggregation of (24) and (25):

$$\psi^\alpha(X) = \max\left(\psi_{\text{Proj}}^{\alpha/2}(X), \psi_{\text{Proj}^\perp}^{\alpha/2}(X)\right). \quad (26)$$

The next lemma provides a sharp characterization of the performance of this test. Crucially, the required sample size depends on the dimension m of the span of the predictions. Compared with the oracle bound in Lemma 1, the required sample size increases by a factor of $\sqrt{m}$. This is expected since (26) relies on a chi-squared test on the span of the predictions, which looks for deviations from the null hypothesis in all directions of the predictions' span. The proof is deferred to Appendix D.1 of the supplementary material.

Lemma 3 (The cost of adaptation to orthogonal predictions). *For hypotheses (23), if the predictions are orthogonal, there exists a positive constant C that depends only on α and β such that the test (26) is valid and powerful for the separation curve*

$$\epsilon_*(r) = \inf\{\epsilon \geq 0 : n \geq C \cdot g(\epsilon, r)\} \quad (27)$$

where for $m \leq d/2$,

$$g(\epsilon, r) = \begin{cases} \frac{\sqrt{m}}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} & \text{if } S_\epsilon \cap B_r \neq \emptyset \\ \frac{\sqrt{m}}{\|\theta_*(r)\|_2^2} & \text{if } S_\epsilon \cap B_r = \emptyset \end{cases}$$

and for $m > d/2$, $g(\epsilon, r) = \sqrt{d} \cdot \epsilon^{-2}$.

4.1 Characterization of optimal adaptation

In this section, we show that Lemma 3 is optimal in the relevant adaptive sense. The separation curve (27), achieved by the adaptive test (26), is pointwise improvable. This is typical in adaptive testing problems [Giné and Nickl, 2016]. For example, fix an accuracy vector $r_* \in \mathbb{R}_+^m$. A Bonferroni correction allows us to combine the adaptive test (26) with the oracle test (21) designed for the alternative $H_1 : \|\theta\|_2 \geq \epsilon_*(r_*)$ and $\theta \in B_{r_*}$. This combined test has greater power at r_* . Such a gain, however, can occur only at a fixed number of accuracy vectors, and therefore does not contradict adaptive optimality.

We therefore prove optimality within a class of separation curves. This class contains curves that assign the same separation to accuracy vectors that yield the same minimal projection onto the span of the predictions. In other words, if two accuracy vectors lead to worst-case projections of equal length onto the predictions' span, then their induced alternative hypotheses are considered equally hard to distinguish from the null hypothesis.

Property 1 (Same separation for equivalent accuracy vectors). *Let $\tilde{\epsilon}$ be a separation curve. If $\tilde{\epsilon}(r) = \epsilon$ for $r \in \mathbb{R}_+^m$, then $\tilde{\epsilon}(r') = \epsilon$ for all $r' \in \mathbb{R}_+^m$ such that $\gamma_*(\epsilon, r) = \gamma_*(\epsilon, r')$.*

Note that the property is natural, given that both the separation curves associated with the oracle test in Lemma 1 and the adaptive test in Lemma 3 satisfy this property. The next lemma shows that, among all separation curves satisfying Property 1, the separation curve in Lemma 3 cannot be meaningfully improved.

Lemma 4 (Optimality of the adaptive test (26)). *For $m \leq d/2$ and orthogonal predictions, consider hypotheses (23) with a separation curve $\tilde{\epsilon}$ that has Property 1. If there exists $r_* \in \mathbb{R}_+^m$ such that $\tilde{\epsilon}(r_*) < C \cdot \epsilon_*(r_*)$ where*

$$\epsilon_*(r) = \inf \left\{ \epsilon \geq 0 : n \geq C \cdot \left[\frac{\sqrt{m}}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \right] \text{ and } S_\epsilon \cap B_r \neq \emptyset \right\}, \quad (28)$$

then no valid test can be powerful.

The proof is deferred to Appendix D.2 of the supplementary material. Furthermore, note that Lemma 4 proves that Lemma 3 is sharp up to constants for the $S_\epsilon \cap B_r \neq \emptyset$ case. Together Lemmas 4 and 3 prove Theorems 3 and 4 in Section 1.1.

Lemma 4 strengthens the usual adaptivity guarantee, which only establishes optimality of the separation curve ϵ_* against any *fully dominated* separation curve [Giné and Nickl, 2016]. Such guarantees rule out any valid test with nontrivial power against a candidate curve $\tilde{\epsilon}$ satisfying $\tilde{\epsilon}(r) < C \cdot \epsilon_*(r)$ for every $r \in \mathbb{R}_+^m$. Our lower bound reaches the same conclusion under a weaker requirement: ϵ_* only needs to dominate $\tilde{\epsilon}$ on a particular subset of accuracy vectors.

To see why the dependence on the dimension of the prediction span is sharp, consider a weak oracle that, under the alternative hypothesis, knows the norm of the projection

onto the prediction span but not its direction. This oracle must distinguish

$$H_0 : \hat{U}^T \theta = 0 \text{ versus } H_1 : \exists y \in S_2^{m-1} \text{ s.t. } \|\hat{U}^T \theta\|_2 = \gamma \text{ and } \hat{U}^T \theta = \gamma \cdot y \quad (29)$$

where γ is known. An adversary can make this problem difficult by randomizing the direction of $\hat{U}^T \theta$. In particular, under the alternative, the adversary may choose $H_1 : \hat{U}^T \theta / \gamma \sim \text{Uniform}(S_2^{m-1})$. This distribution spreads the projected signal uniformly across all directions in the prediction span. A useful test must therefore search over the entire unit sphere in $\mathbb{R}^m$ rather than inspect a known direction. This mirrors the lower bound in (22), where randomizing the alternative forces the test to look for deviations in all directions. Consequently, the oracle can distinguish the hypotheses only when $n \geq \sqrt{m} \cdot \gamma^{-2}$. Thus, no test for (29) can remove the $\sqrt{m}$ dependence.

A related construction yields a lower bound for the original problem. Under the hypotheses (23), an adversary can draw random points under the alternative hypothesis satisfying

$$H_1 : \|\theta\|_2 = \epsilon \text{ and } \frac{\hat{U}^T \theta}{\|\hat{U}^T \theta\|_2} \sim \text{Uniform}(S_2^{m-1}). \quad (30)$$

This distribution captures the correct $\sqrt{m}$ dependence but does not correctly encode the role of $\gamma_*(\epsilon, r) = \|A_+(\epsilon, r)\|_2$. The issue arises because $\hat{U}^T \theta$ may contain negative coordinates, which increase $\|\hat{U}^T \theta\|_2$ relative to the optimal norm $\|(\hat{U}^T \theta)_+\|_2$. To obtain a tight lower bound, we must restrict the distribution to be supported on the positive orthant. However, replacing S_2^{m-1} by $S_2^{m-1} \cap \{\theta \in \mathbb{R}^m : \theta \geq 0\}$ in (30) does not work because the uniform measure on the positive orthant remains too diffuse. That is, the concentration of one-dimensional projections of said random vectors depends on the dimension m . Instead, we must support the distribution on a smaller structured subset of the positive orthant. Let $\mathcal{S}$ denote the collection of all subsets of size $\sqrt{m}$ from $[m]$. Under the alternative hypothesis, the adversary chooses

$$H_1 : \|\theta\|_2 = \epsilon \text{ and } \frac{\hat{U}^T \theta}{\|\hat{U}^T \theta\|_2} = \mathbb{1}_S \cdot m^{-1/4} \text{ where } S \sim \text{Uniform}(\mathcal{S})$$

and $[\mathbb{1}_S]_i = I(i \in S)$. Lemma 4 utilizes this random alternative to prove that the rates attained by test (26) are optimal.

5 The cost of adapting to arbitrary predictions

In the previous section, we considered only orthogonal predictions. In that setting, the test's performance depends on the ambient dimension of the span of predictions, because each prediction provides information about the alternative that the others cannot represent. This becomes overly pessimistic when predictions are not orthogonal, since many predictions may cluster tightly, especially when they are obtained by retraining the same predictive method on different subsamples of the available data.

Henceforth, we consider the realistic scenario in which the accuracy of the predictions is unknown, and the predictions are not assumed to be orthogonal. That is, we

distinguish hypotheses (23) where the predictions span a k -dimensional subspace. That is, $\text{rank } \hat{U} = k \leq \min(m, d)$.

To understand how non-orthogonality changes the geometry of the problem, consider using the projection test (26) we employed for the case of orthogonal predictions. The following lemma shows that the performance of the test depends on the ambient dimension of the predictions' span, even if all predictions are highly aligned. The proof is deferred to Appendix E.1 of the supplementary material.

Lemma 5 (Performance of the adaptive test that uses standard projection). *For hypotheses (23), there exists a positive constant C that depends only on α and β such that the adaptive test using standard projection (26) is powerful for the separation curve*

$$\epsilon_*(r) = \inf\{\epsilon \geq 0 : n \geq C \cdot g(\epsilon, r)\}$$

where

$$g(\epsilon, r) = \begin{cases} \frac{\sqrt{k}}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} & \text{if } S_\epsilon \cap B_r \neq \emptyset \\ \frac{\sqrt{k}}{\|\theta_*(r)\|_2^2} & \text{if } S_\epsilon \cap B_r = \emptyset \end{cases}$$

We could have foreseen these results by noting that the test discards useful information by considering all the dimensions of the span of the predictions equally important, since $\|\Pi_{\hat{U}} X\|_2^2 = \sum_{i=1}^k \|\Pi_{\tilde{u}_i} X\|_2^2$ where $\tilde{u}_1, \dots, \tilde{u}_k$ are the left singular vectors of $\hat{U}$, and therefore must pay for looking for deviations from the null on each of them.

A better approach to reduce the strong dependence on the dimension of predictions' span is to hedge our bets and rely mostly on those directions in the span of predictions that are aligned with most of the predictions. As mentioned in the introduction, this idea leads to the test statistic that we call *intrinsic projection* of the data onto the span of the predictions (11): $\|\Pi_{\hat{U}}^{\text{Int}} X\|_2^2 = \sum_{i=1}^k (\lambda_i^2 / \lambda_1^2) \cdot \|\Pi_{\tilde{u}_i} X\|_2^2$. When the predictions are orthogonal, intrinsic projection coincides with standard projection $\|\Pi_{\hat{U}}^{\text{Int}} X\|_2^2 = \|\Pi_{\hat{U}} X\|_2^2$ since the cosine similarity matrix becomes the identity matrix. However, when this is not the case, the weighting automatically disregards directions that are not aligned with most of the predictions, thereby avoiding reliance on the span of the predictions' dimension.

Under the null hypothesis, the test statistic follows a mixture of chi-squared distributions for which the appropriate quantile can be computed. Thus, we can use the following test on the intrinsic projection.

$$\psi_{\text{Int}}^\alpha(X) = I(n \cdot \|\Pi_{\hat{U}}^{\text{Int}} X\|_2^2 \geq q_{1-\alpha}) \quad (31)$$

where $q_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of $\sum_{i=1}^k (\lambda_i^2 / \lambda_1^2) \cdot \chi_1^2$. The term intrinsic reflects the fact that the performance of the test depends not on the dimension of the predictions' span itself, but on the intrinsic dimension of the squared cosine similarity matrix:

$$\text{Int } \Sigma^2 = \frac{\text{Tr } \Sigma^2}{\|\Sigma^2\|_{\text{op}}} = \sum_{i=1}^k \frac{\lambda_i^4}{\lambda_1^4}$$

This quantity provides a stable estimate of the dimensionality of the span of the predictions, as it uses the full singular value distribution rather than only counting nonzero directions. If all predictions coincide, their span has both an intrinsic and an ambient dimension of 1. A small perturbation can sharply increase the ambient dimension even when the predictions remain highly aligned. However, the intrinsic dimension of Σ^2 will remain close to 1 since its eigenvalues will decay rapidly.

Analogously to the case of orthogonal predictions, we expect the intrinsic projection to have no power when the alternative lives in the orthogonal complement of the span of the predictions, which can be detected by using the chi-squared test. Thus, we consider the aggregation test

$$\psi^\alpha(X) = \max \left(\psi_{\text{Int}}^{\alpha/2}(X), \psi_{\chi^2}^{\alpha/2}(X) \right). \quad (32)$$

The following lemma summarizes the rates achieved by the test (32). The proof is deferred to Appendix E.1 of the supplementary material. Furthermore, we note that Theorem 5 in Section 1.1 is a consequence of the following lemma.

Lemma 6 (Performance of the adaptive test that uses intrinsic projection). *For hypotheses (23), there exists a positive constant C that depends only on α and β such that the adaptive test using intrinsic projection (32) is powerful for the separation curve*

$$\epsilon_*(r) = \inf \{ \epsilon \geq 0 : n \geq C \cdot g(\epsilon, r) \} \quad (33)$$

where

$$g(\epsilon, r) = \begin{cases} \frac{\sqrt{\text{Int } \Sigma^2}}{\tilde{\gamma}_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} & \text{if } S_\epsilon \cap B_r \neq \emptyset \\ \frac{\sqrt{\text{Int } \Sigma^2}}{\|\tilde{\theta}_*(r)\|_2^2} & \text{if } S_\epsilon \cap B_r = \emptyset \end{cases}$$

and

$$\tilde{\gamma}_*^2(\epsilon, r) = \inf_{\theta \in S_\epsilon \cap B_r} \|\Pi_{\hat{U}}^{\text{Int}} \theta\|_2^2 \quad \text{and} \quad \|\tilde{\theta}_*(r)\|_2^2 = \inf_{\theta \in B_r} \|\Pi_{\hat{U}}^{\text{Int}} \theta\|_2^2.$$

When the predictions are orthogonal, the intrinsic projection reduces to the standard projection. Consequently, Lemma 6 reduces to Lemma 3, which can be observed since when predictions are orthogonal it holds that $\tilde{\gamma}_*^2(\epsilon, r) = \|A_+(\epsilon, r)\|_2^2$ and $\text{Int } \Sigma^2 = \text{rank } \hat{U} = k$.

Comparing the performance of intrinsic projection in Lemma 6 to that of the standard projection in Lemma 5, the dependence on the dimension of the space is improved since it always holds that $\text{Int } \Sigma^2 \leq \text{rank } \hat{U} = k$. However, this comes at the cost of a smaller projection since $\|\Pi_{\hat{U}}^{\text{Int}} \theta\|_2^2 \leq \|\Pi_{\hat{U}} \theta\|_2^2$ for every $\theta \in \mathbb{R}^d$.

5.1 Dependence on the intrinsic dimension

From an information standpoint, the projection test (24) discards valuable information about the relative importance of each orthogonal direction within the predictions' span.

The intrinsic projection test (31) does not discard this information. Intuitively, we expect the test based on intrinsic projection to perform better than the test based on standard projection when the predictions are aligned, since their span will have a low intrinsic dimension.

To intuitively understand why we should expect its performance to depend on the intrinsic dimension of Σ^2 , note that the intrinsic projection test (31) measures the norm of the random vector $(\Lambda_k/\lambda_1) \cdot \tilde{U}_k^T X$ whose mean is $(\Lambda_k/\lambda_1) \cdot \tilde{U}_k^T \theta$ and covariance matrix $(\Lambda_k^2/\lambda_1^2)/n$. Based on the work of [Laurent et al. \[2012\]](#), we expect to be able to detect a deviation on the intrinsic projection of the data ($H_0 : \theta = 0$ versus $H_1 : \|\Pi_{\tilde{U}}^{\text{Int}} \theta\|_2 \geq \gamma$) as soon as $n \geq \sqrt{\text{Tr } \Lambda_k^4/\lambda_1^4}/\gamma^2$, where $\text{Tr } \Lambda_k^4/\lambda_1^4 = \text{Int } \Sigma^2$. The constructions used by [Laurent et al. \[2012\]](#) do not directly apply to hypotheses (23) but can be adapted to our problem. First, consider all the separation curves that satisfy the following property.

Property 2. *Let $\tilde{\epsilon}$ be a separation curve, and $\tilde{\gamma}(\epsilon, r) = \|A(\epsilon, r)\|_2/\|\hat{U}\|_{\text{op}}$. If $\tilde{\epsilon}(r) = \epsilon$ for $r \in \mathbb{R}_+^m$, then $\tilde{\epsilon}(r') = \epsilon$ for all $r' \in \mathbb{R}_+^m$ such that $\tilde{\gamma}(\epsilon, r) = \tilde{\gamma}(\epsilon, r')$.*

The following lemma provides some evidence for the potential dependence on the intrinsic dimension for adaptive tests. Compared to Lemma 6, the lower bound on the separation curve provided by Lemma 7 is loose since it depends on $\tilde{\gamma}$ rather than $\tilde{\gamma}_*$. Nevertheless, it conveys the message that the performance of tests should depend on the intrinsic dimension of Σ^2 . The proof is deferred to Appendix E.2 of the supplementary material.

Lemma 7 (Dependence of any adaptive test on the intrinsic dimension). *For $\text{rank } \hat{U} \leq d/2$, consider hypotheses (23) with a separation curve $\tilde{\epsilon}$ that satisfies Property 2. If there exists $r \in \mathbb{R}_+^m$ such that $\tilde{\epsilon}(r) < C \cdot \tilde{\epsilon}_*(r)$ where*

$$\tilde{\epsilon}_*(r) = \inf \left\{ \epsilon \geq 0 : n \geq C' \cdot \left(\frac{\sqrt{\text{Int } \Sigma^2}}{\tilde{\gamma}^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \right) \text{ and } \epsilon \geq \sqrt{\frac{m}{\text{Int } \Sigma^2}} \cdot \tilde{\gamma}(\epsilon, r) \right\}, \quad (34)$$

then no test can be valid and powerful.

6 Simulations

In this section, we compare tests that differ in how they use the available predictions. The chi-squared test (16) ignores predictions, whereas the projection test (24) and intrinsic projection test (31) always use the predictions regardless of their quality. We also consider two adaptive tests: one based on standard projection (26) and one based on the intrinsic projection (32).

Given a set of m predicted directions $\{\hat{u}_1, \dots, \hat{u}_m\}$, a trial consists of obtaining an observation X from a least favorable distribution under the alternative hypothesis, providing the observation and set of predictions to a test, and recording if the test

rejects the null hypothesis. Assuming that the span of the predictions has dimension m , we sample an adversarially chosen θ based on the construction used in Lemma 2:

$$\theta = \tilde{U}_m A_*(\epsilon, r) + \tilde{U}_m^\perp z \cdot \sqrt{\epsilon^2 - \|A_*(\epsilon, r)\|_2^2} \quad (35)$$

where the columns of $\tilde{U}_m$ are a basis for the span of $\hat{u}_1, \dots, \hat{u}_m$, the columns of $\tilde{U}_m^\perp$ are a basis for its orthogonal complement, $z \sim \text{Uniform}(S_2^{d-m-1})$ and $A_*(\epsilon, r)$ is chosen to be most adversarial for a projection test (24). Namely, $A_*(\epsilon, r)$ is the unique m -dimensional vector such that $\|A_*(\epsilon, r)\|_2^2 = \inf_{\theta \in S_\epsilon \cap B_r} \langle \Pi_{\tilde{U}} \theta, \theta \rangle$ which guarantees any sampled θ from (35) has the worst cosine similarity with its projection onto the span of the predictions. Note that this is most adversarial to tests that rely on standard projection, like (24) and (26). Finally, we sample $X \sim N(\theta, I_d/n)$, and provide the observation X and the set of predictions to the test under evaluation.

Consider the case of a single prediction. We sample $\hat{u}_1$ uniformly from the unit sphere, and run a trial. We perform 1,000 trials per test. Figure 3 showcases the results. Note that the chi-squared test suffers from the curse of dimensionality, while the projection test (24) does not, so long as the prediction is of high quality. However, for a low-quality prediction, the projection test performs worse than the chi-squared test. This is expected, as the projection test alone is not adaptive. A good trade-off between the performance of the projection and chi-squared tests is obtained by the adaptive test (26), which matches the power of the projection test when the prediction is good, but remains a constant away from the power of the chi-squared test when the prediction is uninformative. This simulation illustrates that the adaptive test (26) achieves a rate-wise free lunch, as described in Section 1.1.

For multiple orthogonal predictions, we sample $\{\hat{u}_1, \dots, \hat{u}_m\}$ from the unit sphere such that they are orthogonal and span an m -dimensional subspace. Again, we perform 1,000 trials per test. Figure 5 showcases the results. Note that we progressively increase the number of orthogonal predicted directions until the number of predictions matches half of the ambient dimension of the data, since at that point Lemma 3 indicates that the adaptive test offers no advantage over the chi-squared test. Since predictions are orthogonal, we are in the worst case from a geometric point of view, as this forces both the projection test (24) and the adaptive test (26) to look for deviations from the null hypothesis in all predicted directions. Thus, power overall decreases as the number of predictions increases. Nevertheless, when predictions are high-quality, we can still outperform the chi-squared test that ignores them.

For non-orthogonal predictions, we modify the construction in (35). Namely, we replace A_* to make the distribution of θ most adversarial for the intrinsic projection (31) rather than the standard projection test. That is, A_* is replaced by $\tilde{A}_*(\epsilon, r)$, which is the unique m -dimensional vector that satisfies $\|\tilde{A}_*(\epsilon, r)\|_2^2 / \|\tilde{U}\|_{\text{op}}^2 = \inf_{\theta \in S_\epsilon \cap B_r} \|\Pi_{\tilde{U}}^{\text{Int}} \theta\|_2^2$. For the simulation, we sample $\{\hat{u}_1, \dots, \hat{u}_m\}$ from the unit sphere while making sure that they span an m -dimensional subspace, and all pairwise cosine similarities are equal. Thus, by controlling their cosine similarity, we can control the intrinsic dimension of Σ^2 . For every test, we perform these 1,000 trials.

Figure 4 showcases the results. First, note that when the predictions are orthogo-

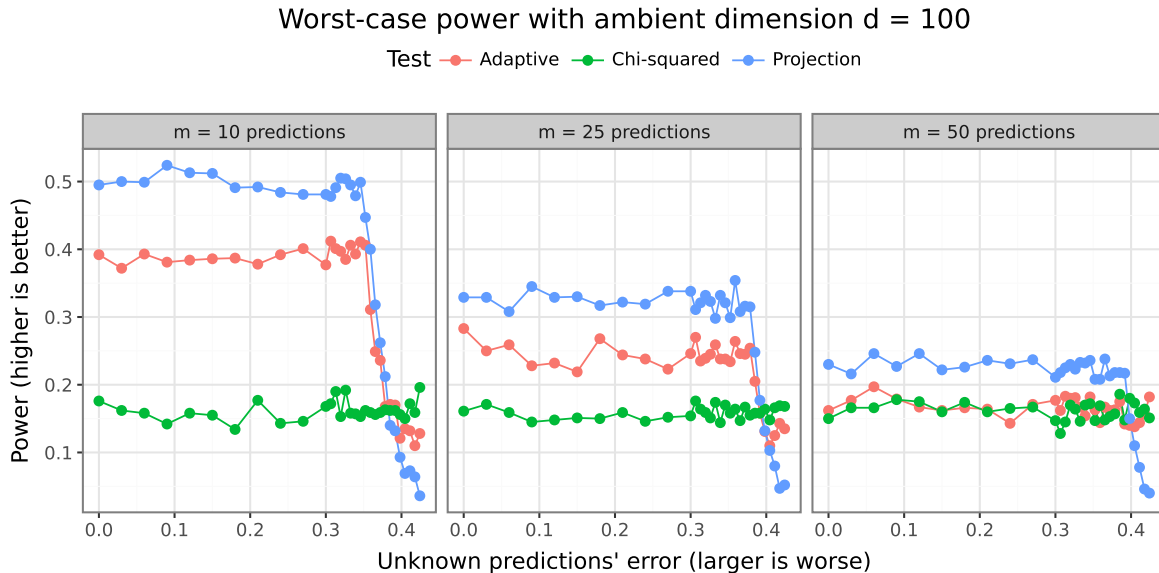


Figure 5: Comparison between the chi-squared test (16), which ignores predictions, the projection test (24), which always uses all predictions, and the adaptive test (26), which optimally adapts to the predictions' error.

nal, methods based on intrinsic or standard projection perform equally. However, even though the simulation has been constructed to be most adversarial to intrinsic projection, when there is high cosine similarity among the predictions, tests based on intrinsic projection, like (31) and (32), take advantage of this and outperform those that use standard projection. Furthermore, the adaptive test that uses intrinsic projection (32) appropriately uses or discards the predictions, adapting to their unknown accuracy, and providing a good trade-off between the intrinsic projection test (31) and the chi-squared test (16).

7 Discussion

In this work, we have sought to provide answers to the question of optimal data aggregation with predictions, such as those produced by machine learning models. Although our results concern the Gaussian sequence model, the fundamental ideas are geometric in nature and therefore provide a way to generalize to other models and geometries. That is, insofar as a space is equipped with an inner product, both standard and intrinsic projections can be realized.

At a high level, it is worth recalling that the problems of estimating and testing with access to black-box predictions are substantially different. As expected, when testing with black-box predictions, we are achieving faster rates than comparable estimation problems with similar guarantees. More importantly, testing requires only information about the *direction* of a prediction. As a result, a prediction may remain useful for

testing even when it lies far from the target, provided that it aligns well with the target of inference. Thus, decision-making can rely on predictions with weaker guarantees than those needed for estimation.

This observation suggests a practical criterion for training predictive models when their outputs feed into downstream decisions. In that setting, high cosine similarity with the target should matter more than distance to the target itself. This shift in evaluation could also yield computational gains, since in many modern machine learning pipelines, the cost of generating a prediction often increases with its quality. Therefore, decision procedures that use weaker yet well-aligned predictions may reduce computational cost without sacrificing operational performance.

Finally, other works have approached the problem of decision-making with black-box predictions by constructing confidence sets and then making decisions based on them. Although equivalent to the testing framework explored in this work, due to the duality between confidence sets and testing procedures, the width of said confidence sets is not fully indicative of the performance of decision-making based on them. This is because their width is fundamentally dominated by the rates of functional estimation, which typically differ from those of testing.

Confidence sets remain highly useful for human-in-the-loop decision-making because they communicate information about the uncertainty involved in the decision-making process. However, as we move into automatic decision-making applications, the testing perspective becomes increasingly relevant, as it is solely concerned with the operational performance of large volumes of decisions without the need to interpret any single decision. Thus, we expect that the testing perspective will be a worthwhile framework for advancing research on large-scale automated decision-making with black-box predictions.

Acknowledgments The authors thank Alexandra Carpentier, Ryan Tibshirani, Arun Kumar Kuchibhotla, Mikael Kuusela, Chad Schafer, Tudor Manole, and Richard Samworth for useful comments and discussions during the development of this work.

Table of contents

A	General upper bound on the performance of tests	29
A.1	Bonferroni correction	30
B	General lower bound on the performance of tests	30
C	Oracle inference	31
C.1	Performance of the oracle test (Lemma 1)	31
C.2	Lower bound on the performance of oracle tests (Lemma 2)	34
D	Adapting to orthogonal predictions	36
D.1	Performance of the adaptive test (Lemma 3)	36
D.2	Lower bound on the performance of adaptive tests (Lemma 4)	38
E	Adapting to arbitrary predictions	44
E.1	Performance of adaptive tests (Lemmas 5 and 6)	44
E.2	Lower bound on the performance of adaptive tests (Lemma 7)	46
	References	50

A General upper bound on the performance of tests

The following proposition provides a means to determine when a test controls both type-I and type-II errors uniformly over the null and alternative hypotheses, respectively. In particular, (A.2) is good enough to capture, up to constants that depend on type-I and type-II errors, the minimum distance between the hypotheses required for a test to distinguish them.

Proposition A.1. *Let $\mathcal{P}_0$ and $\mathcal{P}_1(\epsilon)$ be sets of distributions, and given data $X \sim P$ consider the hypotheses*

$$H_0 : P \in \mathcal{P}_0 \text{ versus } H_1 : P \in \mathcal{P}_1(\epsilon).$$

Let T be a test statistic, and $q_{1-\alpha}(P, T)$ denote the $(1 - \alpha)$ -quantile of the test statistic under the distribution P

$$q_{1-\alpha}(P, T) = \inf \{u : P(T > u) \leq \alpha\}.$$

Define the test

$$\psi(X) = I(T(X) > t) \text{ where } t = \sup_{P \in \mathcal{P}_0} q_{1-\alpha}(P, T),$$

then the test controls the type-I error uniformly over the null hypothesis:

$$\sup_{P \in \mathcal{P}_0} P(\psi(X) = 1) \leq \alpha.$$

Furthermore, if it holds that

$$t \leq \inf_{P \in \mathcal{P}_1} E_P[T(X)] - \sqrt{\frac{V_P[T(X)]}{\beta}}, \quad (\text{A.1})$$

the test controls the type-II error uniformly over the alternative hypothesis

$$\sup_{P \in \mathcal{P}_1} P(\psi(X) = 0) \leq \beta.$$

Consequently, the critical separation ϵ_ is upper bounded by ϵ :*

$$\epsilon_* = \inf \{ \epsilon \geq 0 : \inf_{\psi \in \Psi_\alpha(\mathcal{P}_0)} \sup_{P \in \mathcal{P}_1(\epsilon)} P(\psi(X) = 0) \leq \beta \} \leq \epsilon.$$

Finally, since $q_{1-\alpha}(P, T) \leq E_P[T(X)] + \sqrt{V_P[T(X)]/\alpha}$, it follows that (A.1) is implied by

$$\sup_{P \in \mathcal{P}_0} E_P[T(X)] + \sqrt{\frac{V_P[T(X)]}{\alpha}} \leq \inf_{P \in \mathcal{P}_1} E_P[T(X)] - \sqrt{\frac{V_P[T(X)]}{\beta}}. \quad (\text{A.2})$$

A.1 Bonferroni correction

Proposition A.2. *Let $\mathcal{P}_0$ and $\mathcal{P}_1(\epsilon)$ be sets of distributions, and given data $X \sim P$ consider the hypotheses*

$$H_0 : P \in \mathcal{P}_0 \text{ versus } H_1 : P \in \mathcal{P}_1(\epsilon).$$

Let ψ_i^α for $i \in \{1, 2\}$ be tests such that they control the type-I error uniformly by α . Additionally, they control the type-II error uniformly by β whenever $\epsilon \geq \epsilon_i$ respectively. Then, the test

$$\psi^\alpha(X) = \psi_1^{\alpha \cdot C}(X) \vee \psi_2^{\alpha \cdot (1-C)}(X) \quad \text{for } C \in (0, 1)$$

controls the type-I error uniformly by α . Furthermore, it controls the type-II error uniformly by β whenever $\epsilon \geq \epsilon_1 \wedge \epsilon_2$.

Proof. Under the null hypothesis, we have that

$$\begin{aligned} \sup_{P \in \mathcal{P}_0} E_P [\psi^\alpha] &= \sup_{P \in \mathcal{P}_0} E_P [\psi_1^{\alpha \cdot C} \vee \psi_2^{\alpha \cdot (1-C)}] \\ &= \sup_{P \in \mathcal{P}_0} E_P [\psi_1^{\alpha \cdot C}] + E_P [\psi_2^{\alpha \cdot (1-C)}] - E_P [\psi_1^{\alpha \cdot C} \wedge \psi_2^{\alpha \cdot (1-C)}] \\ &\leq \alpha \end{aligned}$$

Under the alternative hypothesis, it follows that

$$\inf_{P \in \mathcal{P}_1} E_P [\psi^\alpha] \geq \inf_{P \in \mathcal{P}_1(\epsilon)} \left(E_P [\psi_1^{\alpha \cdot C}] \vee E_P [\psi_2^{\alpha \cdot (1-C)}] \right) \geq 1 - \beta$$

whenever $\epsilon \geq \epsilon_1 \wedge \epsilon_2$. □

B General lower bound on the performance of tests

Proposition B.1 (Ingster and Suslina [2003], Tsybakov [2009]). *Let P_0 be a distribution and $\mathcal{P}_1(\epsilon)$ be a set of distributions, and given data $X \sim P$ consider the hypotheses*

$$H_0 : P = P_0 \text{ versus } H_1 : P \in \mathcal{P}_1(\epsilon).$$

Let π be a distribution supported on $\mathcal{P}_1(\epsilon)$. It follows that minimax risk (13) is lower bounded

$$R_*(\mathcal{P}_1(\epsilon)) > \beta \text{ if } \chi^2(E_{P \sim \pi} P, P_0) \leq C_{\alpha, \beta}$$

where $C_{\alpha, \beta}$ is a positive constant that depends only on α and β .

C Oracle inference

C.1 Performance of the oracle test (Lemma 1)

The proof of Lemma 1 depends on the performance of the chi-squared test (Lemma C.1) and the quantification of the worst-case projection onto the span of the predictions under the alternative hypothesis (Lemma C.2). In the following, we state both lemmas, and prove Lemma 1. The proof of Lemma C.2 is at the end of this appendix.

The following lemma quantifies the performance of the chi-squared test.

Lemma C.1 (Ingster [1993]). *For hypotheses (15), there exists a positive constant C that depends only on α and β such that the chi-squared test (16) is powerful whenever*

$$n \geq C \cdot \frac{\sqrt{d}}{\epsilon^2}.$$

The next lemma quantifies the worst-case projection onto the span of the predictions under the alternative hypothesis.

Lemma C.2. *If $k < d$ and $S_\epsilon \cap B_r \neq \emptyset$, it follows that there exists a unique k -dimensional vector $A_*(\epsilon, r)$ such that*

$$A_*(\epsilon, r) = \arg \inf_{a \in \mathcal{A}(\epsilon, r)} \|a\|_2^2 \text{ where } \mathcal{A}(\epsilon, r) = \left\{ \tilde{U}_k^T \theta : \theta \in S_\epsilon \cap B_r \right\} \subseteq \mathbb{R}^k. \quad (\text{C.1})$$

We proceed to prove Lemma 1.

Proof of Lemma 1. In the following, we indicate the conditions under which the chi-squared test (16), the worst-case cosine similarity test (20), and the likelihood-ratio test (17) are valid and powerful. Finally, we combine the conditions to state when the oracle test (21) is valid and powerful.

Throughout the proof, we let $\text{rank } \hat{U} = k$, and let the reduced SVD form of $\hat{U}$ be

$$\hat{U} = \tilde{U}_k \Lambda_k \tilde{V}_k^T,$$

where the columns of $\tilde{U}_k$ span the same k -dimensional subspace spanned by $\hat{U}$. Furthermore, let the columns of $\tilde{U}_k^\perp$ be a basis for its orthogonal complement, a $(d - k)$ -dimensional subspace.

Analysis when S_ϵ and B_r are not disjoint. If $S_\epsilon \cap B_r \neq \emptyset$, the oracle test (21) performs a Bonferroni correction between the chi-squared test (16) and the worst-case cosine similarity test (20):

$$\psi(X) = \psi_{\chi^2}^{\alpha/2}(X) \vee \psi_{\text{CS}}^{\alpha/2}(X) \quad \text{if } S_\epsilon \cap B_r \neq \emptyset. \quad (\text{C.2})$$

Let us first analyze the chi-squared test. By Lemma C.1, it follows that the oracle test is powerful whenever

$$n \geq C_0 \cdot \frac{\sqrt{d}}{\epsilon^2} \text{ and } S_\epsilon \cap B_r \neq \emptyset. \quad (\text{C.3})$$

where C_0 is a constant that depends only on β and α . Turning to the worst cosine similarity test (20), it holds that $\psi_{\text{CS}}^{\alpha/2}(X) = I(T_r(X) \geq z_{1-\alpha/2})$ where

$$T_r(X) = \sqrt{n} \cdot \left\langle \tilde{U}_k^T X, \tilde{u}(\epsilon, r) \right\rangle \text{ and } \tilde{u}(\epsilon, r) = \frac{A_*(\epsilon, r)}{\|A_*(\epsilon, r)\|_2}.$$

By (C.2), Proposition A.1, and the fact that

$$E_{P_\theta}[T_r(X)] = \sqrt{n} \cdot \left\langle \tilde{U}_k^T \theta, \tilde{u}(\epsilon, r) \right\rangle \text{ and } V_{P_\theta}[T_r(X)] = 1 \quad \forall \theta \in \mathbb{R}^d,$$

it holds that the oracle test is powerful whenever

$$\inf_{\theta \in H_1(\epsilon, r)} E_{P_\theta}[T_r(X)] \geq C_1 \cdot 1 \text{ and } S_\epsilon \cap B_r \neq \emptyset \quad (\text{C.4})$$

where $H_1(\epsilon, r) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \geq \epsilon \text{ and } \theta \in B_r\}$ and C_1 is a constant that depends on β and α . Since $S_\epsilon \cap B_r \neq \emptyset$, it holds that

$$\inf_{\theta \in H_1(\epsilon, r)} E_{P_\theta}[T_r(X)] = \inf_{\theta \in S_\epsilon \cap B_r} E_{P_\theta}[T_r(X)].$$

We compute the RHS of (C.4). Since $k < d$ and $S_\epsilon \cap B_r \neq \emptyset$, it follows by Lemma C.2 that there exists a unique k -dimensional vector $A_*(\epsilon, r)$ such that

$$\|A_*(\epsilon, r)\|_2^2 = \inf_{\theta \in B_r \cap S_\epsilon} \|\Pi_{\tilde{U}} \theta\|_2^2 = \inf_{\theta \in B_r \cap S_\epsilon} \|\tilde{U}_k^T \theta\|_2^2 = \gamma_*^2(\epsilon, r).$$

Consequently, it holds that

$$\left\langle \tilde{U}_k^T \theta - A_*(\epsilon, r), A_*(\epsilon, r) \right\rangle \geq 0 \quad \forall \theta \in B_r \cap S_\epsilon,$$

and

$$\begin{aligned} \inf_{\theta \in B_r \cap S_\epsilon} \left\langle \tilde{U}_k^T \theta, A_*(\epsilon, r) \right\rangle &= \inf_{\theta \in B_r \cap S_\epsilon} \|A_*(\epsilon, r)\|_2^2 + \left\langle \tilde{U}_k^T \theta - A_*(\epsilon, r), A_*(\epsilon, r) \right\rangle \\ &= \|A_*(\epsilon, r)\|_2^2. \end{aligned}$$

Therefore, it holds that

$$\inf_{\theta \in S_\epsilon \cap B_r} E_{P_\theta}[T_r(X)] = \sqrt{n} \cdot \|A_*(\epsilon, r)\|_2 = \sqrt{n} \cdot \gamma_*(\epsilon, r). \quad (\text{C.5})$$

Thus, by (C.4) and (C.5), it holds that the oracle test (21) is powerful whenever

$$\sqrt{n} \cdot \gamma_*(\epsilon, r) \geq C_1 \text{ and } S_\epsilon \cap B_r \neq \emptyset, \quad (\text{C.6})$$

where C_1 is a positive constant that depends only on α and β . Finally, by (C.3), (C.6) and Proposition A.2, the oracle test is valid and powerful whenever

$$n \geq C_2 \cdot \left(\frac{1}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \right) \text{ and } S_\epsilon \cap B_r \neq \emptyset, \quad (\text{C.7})$$

where C_2 is a positive constant that depends only on α and β .

Analysis of oracle test when S_ϵ and B_r are disjoint. When $S_\epsilon \cap B_r = \emptyset$, there is a unique point under the alternative hypothesis that is closest to the null hypothesis, and the oracle test performs a likelihood ratio between them:

$$\psi(X) = \psi_{\text{LR}}^\alpha(X) \quad \text{if } S_\epsilon \cap B_r = \emptyset.$$

It follows that the oracle test is powerful whenever

$$n \geq C_3 \cdot \|\theta_*(r)\|_2^{-2} \quad \text{and } S_\epsilon \cap B_r = \emptyset, \quad (\text{C.8})$$

where C_3 is a positive constant that depends only on α and β .

In summary, by (C.7) and (C.8), it holds that the oracle test (21) is valid and powerful whenever

$$n \geq C \cdot \begin{cases} \frac{1}{\gamma_*^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} & \text{if } S_\epsilon \cap B_r \neq \emptyset \\ \frac{1}{\|\theta_*(r)\|_2^2} & \text{if } S_\epsilon \cap B_r = \emptyset \end{cases}$$

where C is a positive constant that depends only on α and β . □

We finish this appendix by proving Lemma C.2.

Proof of Lemma C.2. Since $S_\epsilon \cap B_r \neq \emptyset$, it follows that $\mathcal{A}(\epsilon, r) \neq \emptyset$. We prove that if $k < d$, it holds that

$$\mathcal{A}(\epsilon, r) = \left\{ a \in \mathbb{R}^k : \|a\|_2 \leq \epsilon \text{ and } \hat{U}^T \tilde{U}_k a \geq A(\epsilon, r) \right\}.$$

Thus, $\mathcal{A}(\epsilon, r)$ is the intersection of a convex ball and closed half-spaces, so it is closed and convex. Consequently, the minimizer of (C.1) is unique.

First, we prove that for any $a \in \mathbb{R}^k$ such that $\|a\|_2 \leq \epsilon$ and $\hat{U}^T \tilde{U}_k a \geq A(\epsilon, r)$, there exists a $\theta \in S_\epsilon \cap B_r$ such that $\tilde{U}_k^T \theta = a$. Since $k < d$, there exists unit vector $z \in \text{span } \tilde{U}_k^\perp$. Let

$$\theta = \tilde{U}_k a + z \cdot \sqrt{\epsilon^2 - \|a\|_2^2}.$$

By definition, it follows that $\|\theta\|_2 = \epsilon$ and $\tilde{U}_k^T \theta = a$. Furthermore, for $i \in [m]$, it holds that

$$\begin{aligned} \|\theta - \hat{\theta}_i\|_2^2 &= \|\theta\|_2^2 + h_i^2 - 2h_i \cdot \langle \theta, \hat{u}_i \rangle \\ &= \|\theta\|_2^2 + h_i^2 - 2h_i \cdot \left\langle \tilde{U}_k a, \hat{u}_i \right\rangle \\ &= \|\theta\|_2^2 + h_i^2 - 2h_i \cdot [\hat{U}^T \tilde{U}_k a]_i \end{aligned}$$

Since $\hat{U}^T \tilde{U}_k a \geq A(\epsilon, r)$, it follows that

$$\|\theta - \hat{\theta}_i\|_2^2 \leq \epsilon^2 + h_i^2 - 2h_i \cdot A_i(\epsilon, r) = r_i^2.$$

Thus, $\theta \in B_r$, and consequently, $\tilde{U}_k^T \theta = a$ and $\theta \in B_r \cap S_\epsilon$.

For the reverse inclusion, we prove that for any $\theta \in S_\epsilon \cap B_r$ such that $\tilde{U}_k^T \theta = a$, it holds that $\|a\|_2 \leq \epsilon$ and $\hat{U}^T \tilde{U}_k a \geq A(\epsilon, r)$. Decompose θ as

$$\theta = \tilde{U}_k a + z \text{ where } z \in \text{span } \tilde{U}_k^\perp.$$

Since $\theta \in S_\epsilon$, it holds that $\epsilon^2 = \|\theta\|_2^2 = \|a\|_2^2 + \|z\|_2^2$. Thus, $\|a\|_2 \leq \epsilon$. Furthermore, since $\theta \in B_r$, for $i \in [m]$, it holds that

$$r_i^2 \geq \|\theta - \hat{\theta}_i\|_2^2 = \epsilon^2 + h_i^2 - 2h_i \cdot [\hat{U}^T \tilde{U}_k a]_i,$$

which is equivalent to

$$[\hat{U}^T \tilde{U}_k a]_i \geq A_i(\epsilon, r) \text{ for } i \in [m].$$

That is, $\hat{U}^T \tilde{U}_k a \geq A(\epsilon, r)$. In summary, $\|a\|_2 \leq \epsilon$ and $\hat{U}^T \tilde{U}_k a \geq A(\epsilon, r)$. □

C.2 Lower bound on the performance of oracle tests (Lemma 2)

We will need the following result before proving Lemma 2.

Proposition C.1 (Ingster and Suslina [2003]). *Let P_θ denote the distribution $\mathcal{N}(\theta, I_d/n)$ for $\theta \in \mathbb{R}^d$. Consider P_0 and $P_\pi = E_{\theta \in \pi}[P_\theta]$. It holds that*

$$1 + \chi^2(P_\pi, P_0) = E_{\theta, \theta' \stackrel{\text{iid}}{\sim} \pi} [h(\theta, \theta')] \text{ where } h(\theta, \theta') = \exp(n \cdot \langle \theta, \theta' \rangle).$$

We proceed to prove Lemma 2.

Proof of Lemma 2. Let $\text{rank } \hat{U} = k$, and let the reduced SVD form of $\hat{U}$ be

$$\hat{U} = \tilde{U}_k \Lambda_k \tilde{V}_k^T \text{ where } \Lambda_k = \text{Diagonal}(\lambda_1, \dots, \lambda_k),$$

$\lambda_1^2 \geq \dots \geq \lambda_k^2 > 0$, $\tilde{U}_k$ spans the same k -dimensional subspace spanned by $\hat{U}$. Furthermore, let $\tilde{U}_k^\perp$ be a basis for its orthogonal complement, a $(d-k)$ -dimensional subspace.

Since $k \leq d/2$ and $S_\epsilon \cap B_r \neq \emptyset$, it follows by Lemma C.2 that there exists a unique k -dimensional vector $A_*(\epsilon, r)$ that satisfies

$$\|A_*(\epsilon, r)\|_2^2 = \min_{\theta \in S_\epsilon \cap B_r} \|\Pi_{\hat{U}} \theta\|_2^2 = \min_{\theta \in S_\epsilon \cap B_r} \|\tilde{U}_k^T \theta\|_2^2. \quad (\text{C.9})$$

Define π to be the distribution of θ where

$$\theta = \tilde{U}_k \cdot A_*(\epsilon, r) + \frac{\rho(\epsilon)}{\sqrt{d-k}} \cdot \tilde{U}_k^\perp \cdot z, \quad (\text{C.10})$$

where $z \sim \text{Uniform}(\sqrt{d-k} \cdot S_2^{d-k-1})$ and $\rho(\epsilon) = \sqrt{\epsilon^2 - \|A_*(\epsilon, r)\|_2^2}$. Note that by (C.9),

$$\|A_*(\epsilon, r)\|_2^2 = \min_{\theta \in S_\epsilon \cap B_r} \|\Pi_{\tilde{U}_k} \theta\|_2^2 \leq \min_{\theta \in S_\epsilon \cap B_r} \|\theta\|_2^2 = \epsilon^2.$$

Thus, $\rho(\epsilon)$ is well defined. Furthermore, by the definition of π , it holds that

$$\|\theta\|_2 = \epsilon \text{ and } \hat{U}^T \theta = \tilde{V}_k \Lambda_k A_*(\epsilon, r) \quad \forall \theta \in \text{supp } \pi.$$

By (C.9), it follows that there exists $\tilde{\theta} \in B_r \cap S_\epsilon$ such that

$$A_*(\epsilon, r) = \tilde{U}_k^T \tilde{\theta}.$$

Consequently, we have that

$$\|\theta\|_2 = \epsilon \text{ and } \hat{U}^T \theta = \tilde{V}_k \Lambda_k \tilde{U}_k^T \tilde{\theta} = \hat{U}^T \tilde{\theta} \geq A(\epsilon, r) \quad \forall \theta \in \text{supp } \pi,$$

since any $\tilde{\theta} \in S_\epsilon \cap B_r$ satisfies the constraint $\hat{U}^T \tilde{\theta} \geq A(\epsilon, r)$. This implies that π is supported on the alternative hypothesis since

$$\|\theta\|_2 = \epsilon \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i \text{ for } i \in [m] \iff \|\theta\|_2 = \epsilon \text{ and } \hat{U}^T \theta \geq A(\epsilon, r).$$

By Proposition C.1 and (C.10), it holds that

$$1 + \chi^2(P_\pi, P_0) = E_{\theta, \theta' \stackrel{\text{iid}}{\sim} \pi} \exp(n \cdot \langle \theta, \theta' \rangle)$$

where

$$\langle \theta, \theta' \rangle = \|A_*(\epsilon, r)\|_2^2 + \frac{\rho^2(\epsilon)}{d-k} \cdot \langle z, z' \rangle \text{ and } z, z' \stackrel{\text{iid}}{\sim} \text{Uniform}(\sqrt{d-k} \cdot S_2^{d-k-1}).$$

Hence, we have that

$$1 + \chi^2(P_\pi, P_0) = \exp(n \cdot \|A_*(\epsilon, r)\|_2^2) \cdot E_{z, z' \stackrel{\text{iid}}{\sim} \text{Uniform}(\sqrt{d-k} \cdot S_2^{d-k-1})} \exp\left(\frac{n \rho^2(\epsilon)}{d-k} \cdot \langle z, z' \rangle\right).$$

Since the distribution of z is C_0 -sub-Gaussian, where C_0 is a positive constant that does not depend on d , it follows that

$$1 + \chi^2(P_\pi, P_0) \leq \exp(n \cdot \|A_*(\epsilon, r)\|_2^2) \cdot E_{z'} \exp\left(\frac{C_0^2}{2} \cdot \frac{n^2}{(d-k)^2} \cdot \rho^4(\epsilon) \cdot \|z'\|_2^2\right)$$

where $\|z'\|_2^2 = d-k$. Thus, for any $\eta > 0$, there exists C_η such that $\chi^2(P_\pi, P_0) \leq \eta$, whenever

$$n \cdot \|A_*(\epsilon, r)\|_2^2 + \frac{C_0^2}{2} \cdot \frac{n^2}{d-k} \cdot \rho^4(\epsilon) \leq C_\eta.$$

Using the assumption that $\text{rank } \hat{U} = k \leq d/2$ and the fact that $\rho^4(\epsilon) \leq \epsilon^4$, the above equation is implied by

$$n \cdot \|A_*(\epsilon, r)\|_2^2 + C_0^2 \cdot \frac{n^2}{d} \cdot \epsilon^4 \leq C_\eta.$$

The statement of the lemma follows by Proposition B.1. $\square$

D Adapting to orthogonal predictions

D.1 Performance of the adaptive test (Lemma 3)

If tighter control on the type-I error is needed, we can alternatively aggregate the test statistics of the test (24) and (25):

$$\psi^\alpha(X) = I \left(n \cdot \frac{\|\Pi_{\hat{U}} X\|_2^2}{\sqrt{\text{rank } \hat{U}}} + n \cdot \frac{\|\Pi_{\hat{U}^\perp} X\|_2^2}{\sqrt{d - \text{rank } \hat{U}}} \geq q_{1-\alpha} \right) \quad (\text{D.1})$$

where $q_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of the mixture

$$\frac{\chi_{\text{rank } \hat{U}}^2}{\sqrt{\text{rank } \hat{U}}} + \frac{\chi_{d - \text{rank } \hat{U}}^2}{\sqrt{d - \text{rank } \hat{U}}}.$$

The test (D.1) achieves the same type-I and type-II error guarantees as the test (26) up to constants. Thus, we prove Lemma 3 using the test (D.1), but note that the same result holds for the test (26).

Proof of Lemma 3. Since the predictions are mutually orthogonal, $\text{rank } \hat{U} = k = m$. Thus, k and m are equal. Throughout the proof, we use k rather than m .

Power analysis for the test (D.1). For the test statistic of the adaptive test (D.1), it holds that

$$n \cdot \frac{\|\Pi_{\hat{U}} X\|_2^2}{\sqrt{k}} + n \cdot \frac{\|\Pi_{\hat{U}^\perp} X\|_2^2}{\sqrt{d - k}} \sim \frac{\chi_k^2 (n \cdot \|\Pi_{\hat{U}} \theta\|_2^2)}{\sqrt{k}} + \frac{\chi_{d-k}^2 (n \cdot \|\Pi_{\hat{U}^\perp} \theta\|_2^2)}{\sqrt{d - k}} \quad \forall \theta \in \mathbb{R}^d.$$

Let H_1 denote the set of distributions under the alternative hypothesis

$$H_1 = \bigcup_{r \in \mathbb{R}_+^k} H_1(\epsilon(r), r)$$

$$\text{where } H_1(\epsilon, r) = \left\{ \theta \in \mathbb{R}^d : \|\theta\|_2 \geq \epsilon \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i \text{ for } i \in [k] \right\}.$$

By Proposition A.1, it follows that the test (D.1) controls the type-II error by β whenever

$$n \cdot \inf_{\theta \in H_1} f(\theta) \geq C \text{ where } f(\theta) = \frac{\|\Pi_{\hat{U}} \theta\|_2^2}{\sqrt{k}} + \frac{\|\Pi_{\hat{U}^\perp} \theta\|_2^2}{\sqrt{d - k}} \quad (\text{D.2})$$

and C is a constant that only depends on α and β .

Minimum projection. We proceed to characterize the minimum projection onto the span of $\hat{U}$. For any $\theta \in H_1$, there exists $r \in \mathbb{R}_+^k$ such that $\theta \in H_1(\epsilon(r), r)$. Furthermore, note that

$$\gamma_*^2(\epsilon(r), r) \leq \epsilon^2(r) \iff S_{\epsilon(r)} \cap B_r \neq \emptyset \text{ where } B_r = \bigcap_{i=1}^k B_2^d(\hat{\theta}_i, r_i).$$

In particular, if $\gamma_*^2(\epsilon(r), r) \leq \epsilon^2(r)$, we have that

$$\inf_{\theta \in H_1(\epsilon(r), r)} f(\theta) = \inf_{\theta \in H_1(\epsilon(r), r)} \frac{\|\Pi_{\hat{U}}\theta\|_2^2}{\sqrt{k}} + \frac{\epsilon^2(r) - \|\Pi_{\hat{U}}\theta\|_2^2}{\sqrt{d-k}} \quad \text{and} \quad \inf_{\theta \in H_1(\epsilon(r), r)} \|\Pi_{\hat{U}}\theta\|_2^2 = \gamma_*^2(\epsilon(r), r).$$

Therefore, the solution to the optimization is

$$\begin{aligned} \inf_{\theta \in H_1(\epsilon(r), r)} f(\theta) &= \inf_{\|\Pi_{\hat{U}}\theta\|_2^2 \in [\gamma_*^2(\epsilon(r), r), \epsilon^2(r)]} \frac{\|\Pi_{\hat{U}}\theta\|_2^2}{\sqrt{k}} + \frac{\epsilon^2(r) - \|\Pi_{\hat{U}}\theta\|_2^2}{\sqrt{d-k}} \\ &= \begin{cases} \frac{\epsilon^2(r)}{\sqrt{k}} & \text{if } k > d/2 \\ \frac{\gamma_*^2(\epsilon(r), r)}{\sqrt{k}} + \frac{\epsilon^2(r) - \gamma_*^2(\epsilon(r), r)}{\sqrt{d-k}} & \text{if } k \leq d/2 \end{cases} \\ &\geq \begin{cases} \frac{\epsilon^2(r)}{\sqrt{d}} & \text{if } k > d/2 \text{ or } (k \leq d/2 \text{ and } \gamma_*^2(\epsilon(r), r) \leq \epsilon^2(r) \cdot c_k) \\ \frac{\gamma_*^2(\epsilon(r), r)}{\sqrt{k}} & \text{if } k \leq d/2 \text{ and } \epsilon^2(r) \cdot c_k \leq \gamma_*^2(\epsilon(r), r) \leq \epsilon^2(r) \end{cases} \end{aligned}$$

where $c_k = \left[\sqrt{\frac{d}{k} - 1} + 1 \right]^{-1}$. Equivalently, it holds that

$$\inf_{\theta \in H_1(\epsilon(r), r)} f(\theta) \geq \begin{cases} \frac{\epsilon^2(r)}{\sqrt{d}} & \text{if } k > d/2 \\ \frac{\gamma_*^2(\epsilon(r), r)}{\sqrt{k}} \wedge \frac{\epsilon^2(r)}{\sqrt{d}} & \text{if } k \leq d/2 \text{ and } S_{\epsilon(r)} \cap B_r \neq \emptyset \end{cases} \quad (\text{D.3})$$

Disjoint neighborhoods. Conversely, it holds that

$$\gamma_*^2(\epsilon(r), r) > \epsilon^2(r) \iff S_{\epsilon(r)} \cap B_r = \emptyset \quad \text{where} \quad B_r = \bigcap_{i=1}^k B_2^d(\hat{\theta}_i, r_i)$$

Thus, for all $\theta \in B_r$ it holds that $\|\theta\|_2 > \epsilon(r)$. Thus, $H_1(\epsilon(r), r) = B_r$, and the weighted minimum projection of θ onto the spans of $\hat{U}$ and $\hat{U}^\perp$ does not depend on $\epsilon(r)$:

$$\inf_{\theta \in H_1(\epsilon(r), r)} f(\theta) = \inf_{\theta \in B_r} f(\theta) = \inf_{\theta \in B_r} \frac{\|\Pi_{\hat{U}}\theta\|_2^2}{\sqrt{k}} + \frac{\|\Pi_{\hat{U}^\perp}\theta\|_2^2}{\sqrt{d-k}}$$

We can rewrite this optimization. Namely, let

$$\theta = \hat{U}x + \hat{U}^\perp z \quad \text{where } x \in \mathbb{R}^k \text{ and } z \in \mathbb{R}^{d-k}.$$

Since $\hat{U}$ has orthonormal columns, it follows that

$$\|\theta\|_2^2 = \|x\|_2^2 + \|z\|_2^2 \quad \text{and} \quad \|\theta - \hat{\theta}_i\|_2^2 = \|x - h_i \cdot e_i\|_2^2 + \|z\|_2^2 \quad \text{for } i \in [k].$$

Thus, we can rewrite the optimization problem as

$$\inf_{x \in \mathbb{R}^k, z \in \mathbb{R}^{d-k}} \frac{\|x\|_2^2}{\sqrt{k}} + \frac{\|z\|_2^2}{\sqrt{d-k}} \text{ s.t. } \|x\|_2^2 \geq \sum_{i=1}^k \left(\frac{\|x\|_2^2 + h_i^2 - r_i^2 + \|z\|_2^2}{2h_i} \right)_+^2.$$

Note that $z \neq 0$ increases $\|z\|_2^2$ and the norm of x due to the constraint. Thus, the optimizer sets $z = 0$. Therefore, we have that

$$\inf_{\theta \in H_1(\epsilon(r), r)} f(\theta) = \inf_{\theta \in B_r} f(\theta) = \inf_{\theta \in B_r} \frac{\|\Pi_{\hat{U}} \theta\|_2^2}{\sqrt{k}} + \frac{\|\Pi_{\hat{U}^\perp} \theta\|_2^2}{\sqrt{d-k}} = \inf_{\theta \in B_r} \frac{\|\Pi_{\hat{U}} \theta\|_2^2}{\sqrt{k}}$$

Recalling that $\inf_{\theta \in B_r} \|\Pi_{\hat{U}} \theta\|_2^2 = \inf_{\theta \in B_r} \|\theta\|_2^2$, it follows that

$$\inf_{\theta \in H_1(\epsilon(r), r)} f(\theta) = \inf_{\theta \in B_r} f(\theta) = \inf_{\theta \in B_r} \frac{\|\theta\|_2^2}{\sqrt{k}} = \frac{\|\theta_*(r)\|_2^2}{\sqrt{k}} \quad (\text{D.4})$$

where $\theta_*(r)$ is the unique closest point to the null hypothesis, defined in (4).

In summary, by (D.3) and (D.4), for any $\theta \in H_1$, there exists $r \in \mathbb{R}_+^k$ such that $\theta \in H_1(\epsilon(r), r)$, and it holds that

$$\inf_{\theta \in H_1(\epsilon(r), r)} f(\theta) \geq \begin{cases} \frac{\epsilon^2(r)}{\sqrt{d}} & \text{if } k > d/2 \\ \frac{\gamma_*^2(\epsilon(r), r)}{\sqrt{k}} \wedge \frac{\epsilon^2(r)}{\sqrt{d}} & \text{if } k \leq d/2 \text{ and } S_{\epsilon(r)} \cap B_r \neq \emptyset \\ \frac{\|\theta_*(r)\|_2^2}{\sqrt{k}} & \text{if } S_{\epsilon(r)} \cap B_r = \emptyset \end{cases} \quad (\text{D.5})$$

Consequently, by (D.5) and the definition of $\epsilon(r)$, see (27), it follows that

$$n \cdot \inf_{\theta \in H_1} f(\theta) = \inf_{r \in \mathbb{R}_+^k} \left[n \cdot \inf_{\theta \in H_1(\epsilon(r), r)} f(\theta) \right] \geq \inf_{r \in \mathbb{R}_+^k} C \geq C.$$

Thus, by (D.2), the type-II error is uniformly controlled under the alternative hypothesis by the projection test (D.1). □

D.2 Lower bound on the performance of adaptive tests (Lemma 4)

Proof of Lemma 4. Given any separation curve $\tilde{\epsilon}$, define

$$\mathcal{P}(\tilde{\epsilon}, r) = \left\{ P_\theta : \|\theta\|_2 \geq \tilde{\epsilon}(r) \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i \text{ for } i \in [m] \right\},$$

and let $\mathcal{P}(\tilde{\epsilon})$ denote the set of distributions under the alternative hypothesis

$$\mathcal{P}(\tilde{\epsilon}) = \bigcup_{r \in \mathbb{R}_+^m} \mathcal{P}(\tilde{\epsilon}, r). \quad (\text{D.6})$$

For any $\tilde{\epsilon}$ separation curve that satisfies Property 1, we need to show that if

$$\exists r \in \mathbb{R}_+^m \text{ s.t. } \tilde{\epsilon}(r) \leq C \cdot \epsilon_*(r) \implies R_*(\mathcal{P}(\tilde{\epsilon})) > \beta.$$

We prove this statement using the contrapositive. For any $\tilde{\epsilon}$ separation curve that satisfies Property 1, we show that if

$$R_*(\mathcal{P}(\tilde{\epsilon})) \leq \beta \implies \forall r \in \mathbb{R}_+^m, \tilde{\epsilon}(r) \geq C \cdot \epsilon_*(r).$$

Consider a separation curve $\tilde{\epsilon}$ that satisfies Property 1 and $R_*(\mathcal{P}(\tilde{\epsilon})) \leq \beta$. Fix $r \in \mathbb{R}_+^m$. Then by Lemma D.2, for $\epsilon = \tilde{\epsilon}(r)$ it holds that

$$n \geq C \cdot \left(\frac{\sqrt{d}}{\epsilon^2} \wedge \frac{\sqrt{m}}{\gamma_*^2(\epsilon, r)} \right) \text{ and } \gamma_*(\epsilon, r) \leq \epsilon \quad (\text{D.7})$$

and C is a positive constant that depends only on α and β . Since $S_\epsilon \cap B_r \neq \emptyset \iff \gamma_*(\epsilon, r) \leq \epsilon$, (D.7) is equivalent to

$$n \geq C \cdot \left(\frac{\sqrt{d}}{\epsilon^2} \wedge \frac{\sqrt{m}}{\gamma_*^2(\epsilon, r)} \right) \text{ and } S_\epsilon \cap B_r \neq \emptyset.$$

By definition of $\epsilon_*(r)$, see (28), it holds that

$$\epsilon_*(r) = \inf \left\{ \epsilon \geq 0 : n \geq C \cdot \left(\frac{\sqrt{d}}{\epsilon^2} \wedge \frac{\sqrt{m}}{\gamma_*^2(\epsilon, r)} \right) \text{ and } S_\epsilon \cap B_r \neq \emptyset \right\}.$$

where we used the fact that $S_\epsilon \cap B_r \neq \emptyset \iff \gamma_*(\epsilon, r) \leq \epsilon$. Thus, it must hold that $\tilde{\epsilon}(r) = \epsilon \geq \epsilon_*(r)$. Since $r \in \mathbb{R}_+^m$ in the previous argument was arbitrary, the argument is valid for all $r \in \mathbb{R}_+^m$, which concludes the proof. $\square$

Before proving Lemma D.2, we prove a looser result to warm up and get a simplified idea of the lower bound construction. First, we introduce a looser version of Property 1:

Property 3 (Same separation for equivalent accuracy vectors). *Let $\gamma(\epsilon, r) = \|A(\epsilon, r)\|_2$. If $\tilde{\epsilon}(r) = \epsilon$ for $r \in \mathbb{R}_+^m$, then $\tilde{\epsilon}(r') = \epsilon$ for all $r' \in \mathbb{R}_+^m$ such that $\gamma(\epsilon, r) = \gamma(\epsilon, r')$.*

Note that the only difference between the sharp result of Lemma D.2 and the looser result of Lemma D.1 is that the former uses $\|A_+(\epsilon, r)\|_2$ in the rates while the latter uses the sub-optimal choice of $\|A(\epsilon, r)\|_2$.

Lemma D.1. For $m \leq d/2$, given separation curve $\tilde{\epsilon}$ with Property 3 and $r \in \mathbb{R}_+^m$, it holds that if

$$n < C \cdot \left(\frac{\sqrt{m}}{\gamma^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \right) \text{ and } \gamma(\epsilon, r) \leq \epsilon$$

where $\epsilon = \tilde{\epsilon}(r)$, $\gamma(\epsilon, r) = \|A(\epsilon, r)\|_2$, and C is a positive constant that depends on α and β , then the minimax risk (13) over $\mathcal{P}(\tilde{\epsilon})$ (D.6) is not controlled: $R_*(\mathcal{P}(\tilde{\epsilon})) > \beta$.

Proof of Lemma D.1. Since the predictions are mutually orthogonal, $\text{rank } \hat{U} = k = m$. Thus, k and m are equal. Throughout the proof, we use k rather than m . Henceforth, fix $r_* \in \mathbb{R}_+^k$ and let

$$\epsilon = \tilde{\epsilon}(r_*) \text{ and } \gamma = \gamma(\epsilon, r_*) = \|A(\epsilon, r_*)\|_2.$$

Constructing accuracy vectors that match the desired norm. Given $y \in S_2^{k-1}(0, 1)$, γ and ϵ , define k -dimensional vector $r(\epsilon, \gamma, y)$ as follows

$$r_i = \sqrt{\epsilon^2 + h_i^2 - 2h_i y_i \gamma} \text{ for } i \in [k].$$

If $0 \leq \gamma \leq \epsilon$, it follows that

$$\epsilon^2 + h_i^2 - 2h_i y_i \gamma \geq (\epsilon - h_i)^2 \geq 0 \text{ for } i \in [k],$$

which guarantees that $r(\epsilon, \gamma, y) \in \mathbb{R}_+^k$. Furthermore, it holds that

$$A(\epsilon, r) = \gamma \cdot y \text{ and } \|A(\epsilon, r)\|_2 = \gamma \text{ for } r = r(\epsilon, \gamma, y).$$

Constructing distribution supported on the alternative hypothesis. Recall that $\hat{U}$ spans a k -dimensional subspace. Henceforth, let $\hat{U}^\perp$ be a basis for its orthogonal complement, a $(d - k)$ -dimensional subspace. Let π be the distribution of

$$\theta = \hat{U}A(\epsilon, r) + \hat{U}^\perp z \cdot \frac{\rho(\gamma)}{\sqrt{d - k}} \quad (\text{D.8})$$

where $r = r(\epsilon, \gamma, \frac{y}{\sqrt{k}})$, $\rho(\gamma) = \sqrt{\epsilon^2 - \gamma^2}$ and

$$y \sim \text{Uniform}(\sqrt{k} \cdot S_2^{k-1}), z \sim \text{Uniform}(\sqrt{d - k} \cdot S_2^{d-k-1}) \text{ and } y \perp z.$$

By definition of π it holds that

$$\forall \theta \in \text{supp } \pi, \exists y \in S_2^{k-1} \text{ s.t. } \hat{U}^T \theta = A(\epsilon, r(\epsilon, \gamma, y)) \text{ and } \|\theta\|_2 = \epsilon,$$

thus

$$\forall \theta \in \text{supp } \pi, \exists y \in S_2^{k-1} \text{ s.t. } \|\theta\|_2 = \epsilon \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i(\epsilon, \gamma, y) \text{ for } i \in [k].$$

Finally, Property 3 implies that π is supported on the alternative hypothesis.

Chi-squared distance between null hypothesis and mixture. By Proposition C.1 and (D.8), it holds that

$$1 + \chi^2(P_\pi, P_0) = E_{\theta, \theta' \stackrel{\text{iid}}{\sim} \pi} \exp(n \cdot \langle \theta, \theta' \rangle)$$

where

$$\langle \theta, \theta' \rangle = \frac{\gamma^2}{k} \langle y, y' \rangle + \frac{\rho^2(\gamma)}{d-k} \cdot \langle z, z' \rangle.$$

Hence, we have that

$$1 + \chi^2(P_\pi, P_0) = E_{y, y'} \exp\left(\frac{n\gamma^2}{k} \cdot \langle y, y' \rangle\right) \cdot E_{z, z'} \exp\left(\frac{n\rho^2(\gamma)}{d-k} \cdot \langle z, z' \rangle\right).$$

Since the distributions of y and z are C_0 -sub-Gaussian, where C_0 is a positive constant that does not depend on d , it follows that

$$1 + \chi^2(P_\pi, P_0) \leq E_{y'} \exp\left(\frac{C_0^2}{2} \cdot \frac{n^2}{k^2} \cdot \gamma^4 \cdot \|y'\|_2^2\right) \cdot E_{z'} \exp\left(\frac{C_0^2}{2} \cdot \frac{n^2}{(d-k)^2} \cdot \rho^4(\gamma) \cdot \|z'\|_2^2\right)$$

where $\|y'\|_2^2 = k$ and $\|z'\|_2^2 = d - k$. Thus, for any $\eta > 0$, there exists C_η such that $\chi^2(P_\pi, P_0) \leq \eta$, whenever

$$n^2 \cdot \frac{\gamma^4}{k} + \frac{n^2}{d-k} \cdot \rho^4(\gamma) \leq C_\eta$$

where C_η is a positive constant that depends only on η . Thus, by Proposition B.1, it follows that there exists a positive constant C that depends only on α and β such that whenever

$$n < C \cdot \min\left(\frac{\sqrt{k}}{\gamma^2}, \frac{\sqrt{d-k}}{\epsilon^2 - \gamma^2}\right) \text{ and } \gamma^2 \leq \epsilon^2,$$

it follows that $R_*(\mathcal{P}(\tilde{\epsilon})) > \beta$. The statement of the lemma follows by noting that for $k \leq d/2$, it holds that

$$\min\left(\frac{\sqrt{k}}{\gamma^2}, \frac{\sqrt{d-k}}{\epsilon^2 - \gamma^2}\right) \geq \tilde{C} \cdot \min\left(\frac{\sqrt{k}}{\gamma^2}, \frac{\sqrt{d}}{\epsilon^2}\right)$$

where $\tilde{C}$ is a positive constant. □

We are ready to improve the lower bound argument by proving Lemma D.2. The construction in Lemma D.1 is loose because for some θ in the support of π , some of the entries of $\hat{U}^T \theta$ are allowed to be negative. Thus, we should restrict the support of π so that $\hat{U}^T \theta \geq 0$ element-wise. A natural idea is to restrict y in (D.8) to belong to the positive orthant, but doing so would destroy the sub-Gaussian property on which the lemma relies. This is because $y \sim \text{Uniform}(\{x \in \mathbb{R}^m \text{ s.t. } \|x\| = \sqrt{m} \text{ and } x \geq 0\})$ is C -sub-Gaussian but C depends on the dimension m .

To correct this issue in the proof of Lemma 4, we will construct the support of π using *sparse* y vectors that belong to the positive orthant. We will use the standard fact about the hypergeometric and binomial distributions.

Proposition D.1 (Convex ordering of hypergeometric and binomial random variables). *For $X \sim \text{Hypergeometric}(N, k, n)$ and $Y \sim \text{Binomial}(n, \frac{k}{N})$, it holds that $E[f(X)] \leq E[f(Y)]$ for every convex function f . Consequently, for every $\lambda \geq 0$ it holds that*

$$E[\exp(\lambda X)] \leq E[\exp(\lambda Y)] = \left(1 - \frac{k}{N} + \frac{k}{N} \exp(\lambda)\right)^n \leq \exp\left(\frac{k \cdot n}{N} (e^\lambda - 1)\right)$$

In the following, we prove Lemma D.2.

Lemma D.2. *For $m \leq d/2$, given separation curve $\tilde{\epsilon}$ with Property 1 and $r \in \mathbb{R}_+^m$, it holds that if*

$$n < C \cdot \left(\frac{\sqrt{d}}{\epsilon^2} \wedge \frac{\sqrt{m}}{\gamma_*^2(\epsilon, r)} \right) \text{ and } \gamma_*(\epsilon, r) \leq \epsilon,$$

where $\epsilon = \tilde{\epsilon}(r)$, $\gamma_*(\epsilon, r) = \|A_+(\epsilon, r)\|_2$, and C is a positive constant that depends on α and β , then the minimax risk (13) over $\mathcal{P}(\tilde{\epsilon})$ (D.6) is not controlled: $R_*(\mathcal{P}(\tilde{\epsilon})) > \beta$.

Proof. Since the predictions are mutually orthogonal, $\text{rank } \hat{U} = k = m$. Thus, k and m are equal. Throughout the proof, we use k rather than m . Henceforth, fix $r_* \in \mathbb{R}_+^k$, and let

$$\epsilon = \tilde{\epsilon}(r_*) \text{ and } \gamma = \gamma_*(\epsilon, r_*) = \|A_+(\epsilon, r_*)\|_2.$$

Constructing accuracy vectors that match the desired norm. Given a subset $S \subseteq \{1, \dots, k\}$ of size $M \in [k]$, γ and ϵ , define the k -dimensional vector $r(\epsilon, \gamma, S)$ as follows

$$r_i = \sqrt{\epsilon^2 + h_i^2 - 2h_i\gamma \cdot \frac{I(i \in S)}{\sqrt{M}}} \text{ for } i \in [k].$$

If $0 \leq \gamma \leq \epsilon$, it follows that

$$\epsilon^2 + h_i^2 - 2h_i\gamma \cdot \frac{I(i \in S)}{\sqrt{M}} \geq (\epsilon - h_i)^2 \geq 0 \text{ for } i \in [k],$$

which guarantees that $r(\epsilon, \gamma, S) \in \mathbb{R}_+^k$. Furthermore, it holds that

$$A_i(\epsilon, r) = \gamma \cdot \frac{I(i \in S)}{\sqrt{M}} \geq 0 \text{ and } \|A_+(\epsilon, r)\|_2 = \gamma \text{ for } r = r(\epsilon, \gamma, S).$$

Constructing a distribution supported on the alternative hypothesis. Recall that $\hat{U}$ spans a k -dimensional subspace. Henceforth, let $\hat{U}^\perp$ be a basis for its orthogonal complement, a $(d - k)$ -dimensional subspace. Furthermore, let $\mathcal{S}$ denote the collection of all $\binom{k}{M}$ subsets of $[k]$ having size M . Let π be the distribution of

$$\theta = \hat{U}A(\epsilon, r) + \hat{U}^\perp z \cdot \frac{\rho(\gamma)}{\sqrt{d - k}} \quad (\text{D.9})$$

where $r = r(\epsilon, \gamma, S)$, $\rho(\gamma) = \sqrt{\epsilon^2 - \gamma^2}$, and

$$S \sim \text{Uniform}(\mathcal{S}), z \sim \text{Uniform}\left(\sqrt{d - k} \cdot S_2^{d-k-1}\right) \text{ and } S \perp\!\!\!\perp z.$$

By definition of π it holds that

$$\forall \theta \in \text{supp } \pi, \exists S \in \mathcal{S} \text{ s.t. } \hat{U}^T \theta = A(\epsilon, r(\epsilon, \gamma, S)) \text{ and } \|\theta\|_2 = \epsilon,$$

thus

$$\forall \theta \in \text{supp } \pi, \exists S \in \mathcal{S} \text{ s.t. } \|\theta\|_2 = \epsilon \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i(\epsilon, \gamma, S) \text{ for } i \in [k].$$

Due to Property 1, it holds that π is supported on the alternative hypothesis.

Chi-squared distance between null hypothesis and mixture. By Proposition C.1 and (D.9), it holds that

$$1 + \chi^2(P_\pi, P_0) = E_{\theta, \theta' \stackrel{\text{iid}}{\sim} \pi} \exp(n \cdot \langle \theta, \theta' \rangle)$$

where

$$\langle \theta, \theta' \rangle = \frac{\gamma^2}{M} \cdot |S \cap S'| + \frac{\rho^2(\gamma)}{d-k} \cdot \langle z, z' \rangle.$$

Hence, we have that

$$1 + \chi^2(P_\pi, P_0) = E_{S, S'} \exp\left(\frac{n\gamma^2}{M} \cdot |S \cap S'|\right) \cdot E_{z, z'} \exp\left(\frac{n\rho^2(\gamma)}{d-k} \cdot \langle z, z' \rangle\right). \quad (\text{D.10})$$

Bound for second term in RHS of (D.10). Since the distribution of z is C_0 -sub-Gaussian, where C_0 is a positive constant that does not depend on d , it follows that

$$\begin{aligned} E_{z, z'} \exp\left(\frac{n\rho^2(\gamma)}{d-k} \cdot \langle z, z' \rangle\right) &\leq E_{z'} \exp\left(\frac{C_0^2}{2} \cdot \frac{n^2}{(d-k)^2} \cdot \rho^4(\gamma) \cdot \|z'\|_2^2\right) \\ &= \exp\left(\frac{C_0^2}{2} \cdot \frac{n^2}{d-k} \cdot \rho^4(\gamma)\right) \end{aligned}$$

where we used the fact that $\|z'\|_2^2 = d - k$ in the last equality.

Bound for first term in RHS of (D.10). Fix S' , and note that $|S \cap S'|$ can be rewritten as follows

$$|S \cap S'| = \sum_{i \in S'} I(i \in S)$$

Fix S' . Consider an urn containing k balls labeled by $[k]$, with the $M = |S'|$ balls indexed by S' colored red and the remaining $k - M$ balls colored blue. Drawing $M = |S|$ balls uniformly without replacement produces the subset S . Therefore, $|S \cap S'|$ is the number of red balls drawn, and hence

$$|S \cap S'| \mid S' \sim \text{Hypergeometric}(k, M, M).$$

Since the RHS does not depend on S' , it follows that

$$|S \cap S'| \sim \text{Hypergeometric}(k, M, M).$$

By Proposition D.1, it holds that

$$E_{S,S'} \exp \left(\frac{n\gamma^2}{M} \cdot |S \cap S'| \right) \leq \exp \left(\frac{M^2}{k} \cdot \left(\exp \left(\frac{n\gamma^2}{M} \right) - 1 \right) \right)$$

Bound for (D.10). Choosing $M = \sqrt{k}$, it follows that

$$1 + \chi^2(P_\pi, P_0) \leq \exp \left(\exp \left(\frac{n\gamma^2}{\sqrt{k}} \right) - 1 + \frac{C_0^2}{2} \cdot \frac{n^2}{d-k} \cdot \rho^4(\gamma) \right)$$

Thus, for any $\eta > 0$, there exists C_η such that $\chi^2(P_\pi, P_0) < \eta$, whenever

$$\exp \left(\frac{n\gamma^2}{\sqrt{k}} \right) - 1 + \frac{C_0^2}{2} \cdot \frac{n^2}{d-k} \cdot \rho^4(\gamma) < C_\eta.$$

where C_η is a positive constant that depends on η . Thus, by Proposition B.1, it follows that there exists a positive constant C that depends on α and β such that whenever

$$n < C \cdot \min \left(\frac{\sqrt{k}}{\gamma^2}, \frac{\sqrt{d-k}}{\epsilon^2 - \gamma^2} \right) \text{ and } \gamma^2 \leq \epsilon^2,$$

it follows that $R_*(\mathcal{P}(\tilde{\epsilon})) > \beta$. The statement of the lemma follows by noting that for $k \leq d/2$, it holds that

$$\min \left(\frac{\sqrt{k}}{\gamma^2}, \frac{\sqrt{d-k}}{\epsilon^2 - \gamma^2} \right) \geq \tilde{C} \cdot \min \left(\frac{\sqrt{k}}{\gamma^2}, \frac{\sqrt{d}}{\epsilon^2} \right)$$

where $\tilde{C}$ is a positive constant. □

E Adapting to arbitrary predictions

E.1 Performance of adaptive tests (Lemmas 5 and 6)

In the following, we provide a proof of Lemma 6. We omit the proof of Lemma 5 since it is analogous.

Proof of Lemma 6. Distribution of test statistic of the intrinsic projection test (31). Recall that $\hat{U}$ is the $d \times m$ matrix where the i -th column is the prediction $\hat{u}_i$. Let $k = \text{rank}(\hat{U}) \leq \min(m, d)$. Then by the singular value decomposition we have that

$$\hat{U} = \tilde{U}_k \Lambda_k \tilde{V}_k^T \text{ where } \tilde{U}_k \in \mathbb{R}^{d \times k}, \Lambda_k \in \mathbb{R}^{k \times k} \text{ and } \tilde{V}_k \in \mathbb{R}^{m \times k} \quad (\text{E.1})$$

where both $\tilde{U}_k$ and $\tilde{V}_k$ have orthonormal columns, and

$$\Lambda_k = \text{Diagonal}(\lambda_1, \dots, \lambda_k) \text{ s.t. } \lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_k^2 > 0.$$

Therefore, it holds that

$$T_{\hat{U}}(X) = \frac{n}{\|\hat{U}\|_{\text{op}}^2} \cdot \|\hat{U}^T X\|_2^2 = \sum_{i=1}^k \frac{\lambda_i^2}{\lambda_1^2} \cdot \|\sqrt{n} \cdot \Pi_{\tilde{u}_i} X\|_2^2.$$

It always holds that

$$T_{\hat{U}}(X) \sim \sum_{i=1}^k \frac{\lambda_i^2}{\lambda_1^2} \cdot \chi_1^2(n \cdot \|\Pi_{\tilde{u}_i} \theta\|_2^2) \quad \forall \theta \in \mathbb{R}^d$$

where each term is independent. Thus, the expectation and variance are

$$\begin{aligned} \lambda_1^2 \cdot E_{P_\theta}[T_{\hat{U}}(X)] &= \|\lambda\|_2^2 + n \cdot \sum_{i=1}^k \lambda_i^2 \cdot \|\Pi_{\tilde{u}_i} \theta\|_2^2 \\ &= \|\lambda\|_2^2 + n \cdot \|\Lambda_k \tilde{U}_k^T \theta\|_2^2 \\ &= \|\lambda\|_2^2 + n \cdot \|\hat{U}^T \theta\|_2^2, \end{aligned} \tag{E.2}$$

$$\begin{aligned} \lambda_1^4 \cdot V_{P_\theta}[T_{\hat{U}}(X)] &= 2 \cdot \|\lambda\|_4^4 + 4n \cdot \sum_{i=1}^k \lambda_i^4 \cdot \|\Pi_{\tilde{u}_i} \theta\|_2^2 \\ &\leq 2 \cdot \|\lambda\|_4^4 + 4 \cdot \lambda_1^2 \cdot n \cdot \|\Lambda_k \tilde{U}_k^T \theta\|_2^2 \\ &= 2 \cdot \|\lambda\|_4^4 + 4 \cdot \lambda_1^2 \cdot n \cdot \|\hat{U}^T \theta\|_2^2. \end{aligned} \tag{E.3}$$

Condition for the intrinsic projection test to be powerful. Henceforth, let $\epsilon(r)$ be an arbitrary separation curve, and let H_1 denote the set of distributions under the alternative hypothesis

$$H_1 = \bigcup_{r \in \mathbb{R}_+^m} H_1(\epsilon(r), r) \quad \text{where } H_1(\epsilon, r) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \geq \epsilon \text{ and } \theta \in B_r\}.$$

By (E.2), (E.3), and Proposition A.1, it follows that the intrinsic projection test (31) is powerful whenever

$$n \cdot \inf_{\theta \in H_1} \|\hat{U}^T \theta\|_2^2 \geq C \cdot (\|\lambda\|_4^2 + \lambda_1^2)$$

where C is a constant that only depends on α and β . Noting that $\|\lambda\|_4^2 \geq \lambda_1^2$, the above is implied by

$$n \cdot \inf_{\theta \in H_1} \|\hat{U}^T \theta\|_2^2 \geq 2C \cdot \|\lambda\|_4^2.$$

Equivalently, the condition can be written as follows

$$n \cdot \inf_{\theta \in H_1} \frac{\|\hat{U}^T \theta\|_2^2}{\lambda_1^2} \geq 2C \cdot \sqrt{\text{Int } \Sigma^2} \quad \text{since } \frac{\|\lambda\|_4^2}{\lambda_1^2} = \sqrt{\frac{\text{Tr } \Sigma^2}{\|\Sigma^2\|_{\text{op}}}} = \sqrt{\text{Int } \Sigma^2}. \tag{E.4}$$

The minimum intrinsic projection. For any $\theta \in H_1$, there exists $r \in \mathbb{R}_+^m$ such that $\theta \in H_1(\epsilon(r), r)$. Since $\theta \in H_1(\epsilon(r), r)$, it holds that $H_1(\epsilon(r), r) \neq \emptyset$ and consequently $B_r \not\subset B_{\epsilon(r)}^\circ$. Therefore, one of two situations must happen: either $B_r \cap S_{\epsilon(r)} \neq \emptyset$, or $B_r \cap S_{\epsilon(r)} = \emptyset$. Consequently,

$$\inf_{\theta \in H_1(\epsilon(r), r)} \frac{\|\hat{U}^T \theta\|_2^2}{\lambda_1^2} = \begin{cases} \tilde{\gamma}_*^2(\epsilon(r), r) & \text{if } B_r \cap S_{\epsilon(r)} \neq \emptyset \\ \|\tilde{\theta}_*(r)\|_2^2 & \text{if } B_r \cap S_{\epsilon(r)} = \emptyset \end{cases} \quad (\text{E.5})$$

Condition for the aggregated test to be powerful. Since

$$\inf_{\theta \in H_1} \|\hat{U}^T \theta\|_2^2 = \inf_{r \in \mathbb{R}_+^m} \inf_{\theta \in H_1(\epsilon(r), r)} \|\hat{U}^T \theta\|_2^2,$$

by (E.4) and (E.5), the test is powerful whenever

$$2C \leq n \cdot \inf_{r \in \mathbb{R}_+^m} \frac{1}{\sqrt{\text{Int } \Sigma^2}} \cdot \begin{cases} \tilde{\gamma}_*^2(\epsilon(r), r) & \text{if } B_r \cap S_{\epsilon(r)} \neq \emptyset \\ \|\tilde{\theta}_*(r)\|_2^2 & \text{if } B_r \cap S_{\epsilon(r)} = \emptyset \end{cases} \quad (\text{E.6})$$

Furthermore, by Lemma C.1, we know that the chi-squared test is powerful whenever

$$n \cdot \inf_{r \in \mathbb{R}_+^m} \frac{\epsilon^2(r)}{\sqrt{d}} \geq C' \quad (\text{E.7})$$

where C' is a positive constant that depends on α and β . Consequently, by (E.6), (E.7), and Proposition A.2 it holds that the aggregated test (32) is powerful whenever

$$2C \vee C' \leq n \cdot \left(\left[\inf_{r \in \mathbb{R}_+^m} \frac{\epsilon^2(r)}{\sqrt{d}} \right] \wedge \inf_{r \in \mathbb{R}_+^m} \frac{1}{\sqrt{\text{Int } \Sigma^2}} \cdot \begin{cases} \tilde{\gamma}_*^2(\epsilon(r), r) & \text{if } B_r \cap S_{\epsilon(r)} \neq \emptyset \\ \|\tilde{\theta}_*(r)\|_2^2 & \text{if } B_r \cap S_{\epsilon(r)} = \emptyset \end{cases} \right)$$

which is implied by

$$2C \vee C' \leq n \cdot \inf_{r \in \mathbb{R}_+^m} \left(\frac{\epsilon^2(r)}{\sqrt{d}} \wedge \frac{1}{\sqrt{\text{Int } \Sigma^2}} \cdot \begin{cases} \tilde{\gamma}_*^2(\epsilon(r), r) & \text{if } B_r \cap S_{\epsilon(r)} \neq \emptyset \\ \|\tilde{\theta}_*(r)\|_2^2 & \text{if } B_r \cap S_{\epsilon(r)} = \emptyset \end{cases} \right)$$

The lemma is proved by noting that $\epsilon(r) = \epsilon_*(r)$, see (33), satisfies the above condition. $\square$

E.2 Lower bound on the performance of adaptive tests (Lemma 7)

The proof of Lemma 7 follows by an argument analogous to that used for Lemma 4.

Proof of Lemma 7. Given any separation curve $\tilde{\epsilon}$, let

$$\mathcal{P}(\tilde{\epsilon}, r) = \left\{ P_\theta : \|\theta\|_2 \geq \tilde{\epsilon}(r) \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i \text{ for } i \in [m] \right\},$$

and

$$\mathcal{P}(\tilde{\epsilon}) = \bigcup_{r \in \mathbb{R}_+^m} \mathcal{P}(\tilde{\epsilon}, r). \quad (\text{E.8})$$

For any $\tilde{\epsilon}$ separation curve that satisfies Property 2, we need to show that if

$$\exists r \in \mathbb{R}_+^m \text{ s.t. } \tilde{\epsilon}(r) \leq C \cdot \tilde{\epsilon}_*(r) \implies R_*(\mathcal{P}(\tilde{\epsilon})) > \beta.$$

We prove this statement using the contrapositive. For any $r \mapsto \tilde{\epsilon}(r)$ separation curve that satisfies Property 2, we show that if

$$R_*(\mathcal{P}(\tilde{\epsilon})) \leq \beta \implies \forall r \in \mathbb{R}_+^m, \tilde{\epsilon}(r) \geq C \cdot \tilde{\epsilon}_*(r).$$

Consider a separation curve $\tilde{\epsilon}$ that satisfies Property 2 and $R_*(\mathcal{P}(\tilde{\epsilon})) \leq \beta$. Fix $r \in \mathbb{R}_+^m$; then, by Lemma E.1, for $\epsilon = \tilde{\epsilon}(r)$, it holds that

$$n \geq C \cdot \left(\frac{\sqrt{\text{Int } \Sigma^2}}{\tilde{\gamma}^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \right) \text{ and } \tilde{\gamma}(\epsilon, r) \cdot \sqrt{\frac{m}{\text{Int } \Sigma^2}} \leq \epsilon,$$

where C is a positive constant that depends only on α and β . Finally, by definition of $\tilde{\epsilon}_*(r)$, see (34), it holds that

$$\tilde{\epsilon}_*(r) = \inf \left\{ \epsilon \geq 0 : n \geq C \cdot \left(\frac{\sqrt{\text{Int } \Sigma^2}}{\tilde{\gamma}^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \right) \text{ and } \tilde{\gamma}(\epsilon, r) \cdot \sqrt{\frac{m}{\text{Int } \Sigma^2}} \leq \epsilon \right\},$$

Thus, it must hold that $\tilde{\epsilon}(r) \geq \tilde{\epsilon}_*(r)$. Since $r \in \mathbb{R}_+^m$ in the previous argument was arbitrary, the argument is valid for all $r \in \mathbb{R}_+^m$, which concludes the proof. $\square$

In the following, we prove the necessary Lemma E.1.

Lemma E.1. *For $k \leq d/2$, given separation curve $\tilde{\epsilon}$ with Property 2 and $r \in \mathbb{R}_+^m$, it holds that if*

$$n < C \cdot \left(\frac{\sqrt{\text{Int } \Sigma^2}}{\tilde{\gamma}^2(\epsilon, r)} \wedge \frac{\sqrt{d}}{\epsilon^2} \right) \text{ and } \tilde{\gamma}(\epsilon, r) \cdot \sqrt{\frac{m}{\text{Int } \Sigma^2}} \leq \epsilon$$

where $\epsilon = \tilde{\epsilon}(r)$, $\tilde{\gamma}^2(\epsilon, r) = \|A(\epsilon, r)\|_2^2 / \|\hat{U}\|_{\text{op}}^2$, and C is a positive constant that depends on α and β , then the minimax risk (13) on $\mathcal{P}(\tilde{\epsilon})$ (E.8) is not controlled: $R_*(\mathcal{P}(\tilde{\epsilon})) > \beta$.

Proof of Lemma E.1. Henceforth, fix $r_* \in \mathbb{R}_+^m$ and let

$$\epsilon = \tilde{\epsilon}(r_*) \text{ and } \gamma = \gamma(\epsilon, r_*) = \|A(\epsilon, r_*)\|_2.$$

Constructing accuracy vectors that match the desired norm. Given $y \in S_2^{m-1}$, γ and ϵ , define the m -dimensional vector $r(\epsilon, \gamma, y)$ as follows

$$r_i = \sqrt{\epsilon^2 + h_i^2 - 2h_i y_i \gamma} \text{ for } i \in [m].$$

Since $\max_{i \in [m]} |y_i| \leq 1$, if $0 \leq \gamma \leq \epsilon$, it follows that

$$\epsilon^2 + h_i^2 - 2h_i y_i \gamma \geq (\epsilon - h_i)^2 \geq 0 \text{ for } i \in [m],$$

which guarantees that $r(\epsilon, \gamma, y) \in \mathbb{R}_+^m$. Furthermore, it holds that

$$A(\epsilon, r) = \gamma \cdot y \text{ and } \|A(\epsilon, r)\|_2 = \gamma \text{ for } r = r(\epsilon, \gamma, y).$$

Constructing distribution supported on the alternative hypothesis. Recall that $\hat{U} = \tilde{U}_k \Lambda_k \tilde{V}_k^T$ is the reduced-form singular value decomposition of $\hat{U}$, see (E.1), where $\tilde{U}_k$ spans a k -dimensional subspace. Henceforth, let $\tilde{U}_k^\perp$ be a basis for its orthogonal complement, a $(d - k)$ -dimensional subspace. Let π be the distribution of

$$\theta = \tilde{U}_k \Lambda_k^{-1} \cdot y_w \cdot \gamma + \tilde{U}_k^\perp z \cdot \frac{\rho(\gamma)}{\sqrt{d - k}} \quad (\text{E.9})$$

where

$$\rho(\gamma) = \sqrt{\epsilon^2 - \|\Lambda_k^{-1} y_w\|_2^2 \cdot \gamma^2} \text{ and } z \sim \text{Uniform}(\sqrt{d - k} \cdot S_2^{d-k-1})$$

and

$$y_w = \frac{\Lambda_k^2 w}{\|\lambda\|_4^2}, \quad w \in \{-1, 1\}^k \text{ and } w_i \stackrel{\text{iid}}{\sim} \text{Uniform}(\{-1, 1\}) \text{ for } i \in [k].$$

We require that $\epsilon^2 \geq \|\Lambda_k^{-1} y_w\|_2^2 \cdot \gamma^2$. Noting that

$$\|\Lambda_k^{-1} y_w\|_2^2 = \frac{\|\lambda\|_2^2}{\|\lambda\|_4^4} = \frac{\text{Tr } \Sigma}{\text{Tr } \Sigma^2} = \frac{m}{\text{Tr } \Sigma^2} = \frac{m}{m + \sum_{i \neq j} \langle \hat{u}_i, \hat{u}_j \rangle^2},$$

said condition is satisfied whenever

$$\epsilon^2 \geq \frac{m}{\text{Int } \Sigma^2} \cdot \frac{\gamma^2}{\|\hat{U}\|_{\text{op}}^2}.$$

By definition of π it holds that, for all $\theta \in \text{supp } \pi$,

$$\exists w \in \{-1, 1\}^k \text{ s.t. } \|\theta\|_2 = \epsilon \text{ and } \hat{U}^T \theta = \gamma \cdot \tilde{V}_k y_w = A(\epsilon, r),$$

where $r = r(\epsilon, \gamma, \tilde{V}_k y_w)$ and $\|\tilde{V}_k y_w\|_2 = 1$ since $\|y_w\|_2^2 = 1$ and $\tilde{V}_k$ has orthonormal columns. Thus

$$\forall \theta \in \text{supp } \pi, \exists w \in \{-1, 1\}^k \text{ s.t. } \|\theta\|_2 = \epsilon \text{ and } \|\theta - \hat{\theta}_i\|_2 \leq r_i(\epsilon, \gamma, \tilde{V}_k y_w) \text{ for } i \in [m]$$

Finally, by Property 2, it holds that π is supported on the alternative hypothesis.

Chi-squared distance between null hypothesis and mixture. By Proposition C.1 and (E.9), it holds that

$$1 + \chi^2(P_\pi, P_0) = E_{\theta, \theta' \stackrel{\text{iid}}{\sim} \pi} \exp(n \cdot \langle \theta, \theta' \rangle)$$

where

$$\langle \theta, \theta' \rangle = \frac{\gamma^2}{\|\lambda\|_4^4} \cdot \sum_{i=1}^k \lambda_i^2 \cdot w_i w'_i + \frac{\rho^2(\gamma)}{d-k} \cdot \langle z, z' \rangle.$$

Hence, we have that

$$1 + \chi^2(P_\pi, P_0) = E_{w, w'} \exp \left(\frac{n\gamma^2}{\|\lambda\|_4^4} \cdot \sum_{i=1}^k \lambda_i^2 \cdot w_i w'_i \right) \cdot E_{z, z'} \exp \left(\frac{n\rho^2(\gamma)}{d-k} \cdot \langle z, z' \rangle \right).$$

Since the distribution of z is C_0 -sub-Gaussian, where C_0 is a positive constant that does not depend on d , it follows that

$$E_{z, z'} \exp \left(\frac{n\rho^2(\gamma)}{d-k} \cdot \langle z, z' \rangle \right) \leq E_{z'} \exp \left(\frac{C_0^2}{2} \cdot \frac{n^2}{(d-k)^2} \cdot \rho^4(\gamma) \cdot \|z'\|_2^2 \right),$$

where $\|z'\|_2^2 = d - k$. Furthermore, it holds that

$$E_{w, w'} \exp \left(\frac{n\gamma^2}{\|\lambda\|_4^4} \cdot \sum_{i=1}^k \lambda_i^2 \cdot w_i w'_i \right) = \prod_{j=1}^k \cosh \left(\frac{n\gamma^2 \lambda_j^2}{\|\lambda\|_4^4} \right) \leq \exp \left(\frac{n^2 \gamma^4}{2\|\lambda\|_4^4} \right).$$

Thus, for any $\eta > 0$, there exists C_η such that $\chi^2(P_\pi, P_0) \leq \eta$, whenever

$$n^2 \cdot \frac{\gamma^4}{\|\lambda\|_4^4} + \frac{n^2}{d-k} \cdot \rho^4(\gamma) \leq C_\eta,$$

where C_η is a positive constant that depends on η . Using the assumption that $k \leq d/2$, by Proposition B.1, it follows that there exists a positive constant C that depends on α and β such that whenever

$$n < C \cdot \min \left(\frac{\|\lambda\|_4^2}{\gamma^2}, \frac{\sqrt{d}}{\epsilon^2} \right), \quad \epsilon^2 \geq \frac{m}{\text{Int } \Sigma^2} \cdot \frac{\gamma^2}{\|\hat{U}\|_{\text{op}}^2}, \quad \text{and } k \leq d/2,$$

it follows that $R_*(\mathcal{P}(\tilde{\epsilon})) > \beta$. The statement of the lemma follows by noting that

$$\frac{\|\lambda\|_4^2}{\gamma^2} = \frac{\sqrt{\text{Int } \Sigma^2}}{\gamma^2 / \|\Sigma\|_{\text{op}}} = \frac{\sqrt{\text{Int } \Sigma^2}}{\gamma^2 / \|\hat{U}\|_{\text{op}}^2} \text{ since } \text{Int } \Sigma^2 = \frac{\text{Tr } \Sigma^2}{\|\Sigma^2\|_{\text{op}}} = \frac{\|\lambda\|_4^4}{\|\Sigma\|_{\text{op}}^2}.$$

□

References

- Maryam Aliakbarpour, Piotr Indyk, Ronitt Rubinfeld, and Sandeep Silwal. Optimal Algorithms for Augmented Testing of Discrete Distributions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-Powered Inference, 2023.
- Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient Prediction-Powered Inference, 2024.
- Sivaraman Balakrishnan, Edward H. Kennedy, and Larry Wasserman. The Fundamental Limits of Structure-Agnostic Functional Estimation, 2023.
- Peter J. Bickel, Ya’acov Ritov, and Thomas M. Stoker. Tailor-made tests for goodness of fit to semiparametric hypotheses. *The Annals of Statistics*, 34(2), 2006.
- Matteo Bonvini, Edward H. Kennedy, Oliver Dukes, and Sivaraman Balakrishnan. Doubly-robust inference and optimality in structure-agnostic models with smoothness, 2024.
- Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.
- Abhishek Chakraborty and Tianxi Cai. Efficient and Adaptive Linear Regression in Semi-Supervised Settings. *The Annals of Statistics*, 46, 2018.
- Abhishek Chakraborty and Guorong Dai. A General Framework for Treatment Effect Estimation in Semi-Supervised and High Dimensional Settings, 2024.
- Abhishek Chakraborty, Guorong Dai, and Raymond J. Carroll. Semi-Supervised Quantile Estimation: Robust and Efficient Inference in High Dimensional Settings, 2024.
- Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *International Conference on Computational Learning Theory*, pages 97–111. Springer, 2007.
- Niccolò Dalmaso, Luca Masserano, David Zhao, Rafael Izbicki, and Ann B. Lee. Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning for Reliable Simulator-Based Inference, 2023.
- Patrik Róbert Gerber, Tianze Jiang, Yury Polyanskiy, and Rui Sun. Kernel-Based Tests for Likelihood-Free Hypothesis Testing, 2023.
- Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016.

- Yihong Gu. Open Problem: Structure-Agnostic Minimax Risk for Partial Linear Model. In *Proceedings of Thirty Eighth Conference on Learning Theory*. PMLR, 2025.
- Yuri I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II and III. *Math. Methods Statist*, 2(2):85–114, 1993.
- Yuri I. Ingster. Adaptive chi-square tests. *Journal of Mathematical Sciences*, 99(2): 1110–1119, 2000.
- Yuri I. Ingster and Irina A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169. Springer, 2003.
- Ilse C. F. Ipsen and Arvind K. Saibaba. Stable Rank and Intrinsic Dimension of Real and Complex Matrices, 2024.
- Jana Janková, Rajen D. Shah, Peter Bühlmann, and Richard J. Samworth. Goodness-of-fit Testing in High Dimensional Generalized Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82, 2020.
- Arnold Janssen. Global Power Functions of Goodness of Fit Tests. *The Annals of Statistics*, 28(1):239–253, 2000.
- Jikai Jin and Vasilis Syrgkanis. Structure-agnostic Optimality of Doubly Robust Learning for Treatment Effect Estimation, 2024.
- Jikai Jin and Vasilis Syrgkanis. Sharp Structure-Agnostic Lower Bounds for General Linear Functional Estimation, 2026.
- Jikai Jin, Lester Mackey, and Vasilis Syrgkanis. It’s Hard to Be Normal: The Impact of Noise on Structure-agnostic Estimation, 2025.
- Iain M. Johnstone. Gaussian estimation: Sequence and wavelet models. Unpublished manuscript, 2019. URL https://imjohnstone.su.domains/GE_09_16_19.pdf.
- Anatoli Juditsky and Arkadii Nemirovski. Functional aggregation for nonparametric regression. *The Annals of Statistics*, 28(3):681–712, 2000.
- Anatoli Juditsky, Philippe Rigollet, and Alexandre B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36, 2008.
- Ilmun Kim and Aaditya Ramdas. Dimension-agnostic inference using cross U-statistics, 2024.
- Vladimir Koltchinskii and Karim Lounici. Asymptotics and Concentration Bounds for Bilinear Forms of Spectral Projectors of Sample Covariance, 2015.
- Béatrice Laurent, Jean-Michel Loubes, and Clément Marteau. Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electronic Journal of Statistics*, 6, 2012.

- Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer New York, 1986.
- Anton Rask Lundborg, Ilmun Kim, Rajen D. Shah, and Richard J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52, 2024.
- Diego Martinez-Taboada and Aaditya Ramdas. Intrinsic-dimension empirical Bernstein inequalities for bounded self-adjoint operators, 2026.
- Arkadi Nemirovski. Topics in non-parametric. *Ecole d’Eté de Probabilités de Saint-Flour*, 28:85, 2000.
- Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, pages 558–575, 2012.
- Vladimir G. Spokoiny. Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24(6):2477–2498, 1996.
- Kenta Takatsu and Arun Kumar Kuchibhotla. Bridging Root- n and Non-standard Asymptotics: Dimension-agnostic Adaptive Inference in M-Estimation, 2025.
- Alexandre B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 303–313. Springer, 2003.
- Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.
- Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal Inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.
- Xiaohan Wei and Stanislav Minsker. Estimation of the covariance structure of heavy-tailed distributions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Shirong Xu and Will Wei Sun. Do More Predictions Improve Statistical Inference? Filtered Prediction-Powered Inference, 2026.
- Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47, 2019.
- Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29, 2024.
- Runjia Zou, Daniela Witten, and Brian Williamson. Generalized Prediction-Powered Inference, with Application to Binary Classifier Evaluation, 2026.
- Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15), 2024.