Post-ADC Inference: Valid Inference After Active Data Collection

Shuichi Nishino[1,2] **Tomohiro Shiraishi**[1,2] **Teruyuki Katsuoka**[1] **Ichiro Takeuchi**[1,2]
[1]**Nagoya University** [2]**RIKEN**
nishino.shuichi.nagoyaml@gmail.com
takeuchi.ichiro.n6@f.mail.nagoya-u.ac.jp

Abstract

The validity of statistical inference depends critically on how data are collected. When data gathered through active data collection (ADC) are reused for a post-hoc inferential task, conventional inference can fail because the sampling is adaptively biased toward regions favored by the collection strategy. This issue is especially pronounced in black-box optimization, where sequential model-based optimization (SMBO) methods such as the tree-structured Parzen estimator (TPE) and Gaussian process upper confidence bound (GP-UCB) preferentially concentrate evaluations in promising regions. We study statistical inference on actively collected data when the inferential target is constructed in a data-dependent manner after data collection. To enable valid inference in this setting, we propose post-ADC inference, a framework that accounts for the biases arising from both the active data collection process and the subsequent data-driven target construction. Our method builds on selective inference and provides valid p-values and confidence intervals that correct for both sources of bias. The framework covers a class of ADC processes satisfying structural conditions, with the statistical model imposing assumptions only on the observation noise and none on the underlying black-box function. Empirical results also show that post-ADC inference provides valid inference for data collected by GP-UCB and TPE.

1 Introduction

Quantifying the statistical significance of insights obtained from data fundamentally depends on how the data are collected. In principle, valid statistical inference, such as the computation of p-values or confidence intervals, requires that data be sampled through a procedure aligned with the inferential objective. However, in practice, datasets collected for one task are often reused for another. When the data collection process is biased with respect to the new task, the resulting statistical conclusions may be invalid. This issue becomes particularly critical in settings involving active data collection (ADC). Many statistical methods assume that data are independently and identically distributed (i.i.d.), but this assumption is violated when data are collected adaptively based on past observations. We refer to this setting as post-ADC analysis.

As a proof-of-concept (PoC) of post-ADC analysis, we study black-box optimization problems by sequential model-based optimization (SMBO) methods [1–5]. SMBO is a widely used approach for black-box optimization, in which a surrogate model of the objective function is iteratively updated and used to actively select evaluation points. It has been successfully applied to hyperparameter tuning in machine learning, as well as to optimization problems across various scientific domains. Representative methods include density-based methods such as the tree-structured Parzen estimator (TPE) [3] and Gaussian process-based approaches such as Gaussian process upper confidence bound (GP-UCB) [2]. A common characteristic of these methods is that the collected data are highly biased. For example, in maximization problems, samples tend to concentrate in regions with high observed

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

values, while regions with low values are underexplored. Consequently, the dataset is not an i.i.d. sample from the underlying distribution, but rather the outcome of a goal-directed adaptive process.

This bias poses a fundamental challenge when such data are used for a post-hoc inferential task. The difficulty is compounded when the inferential target is itself constructed in a data-driven manner from the same adaptively collected data. For example, after an ADC process, one may wish to formulate a new statistical question about the underlying black-box function, such as a two-sample test comparing "high" and "low" regions. However, because the sampling process depends on past observations and the inferential target is defined after observing the resulting dataset, naive statistical inference can yield misleading conclusions, resulting in overconfident or spurious findings. Figure 1 provides a concrete illustration of this setting.

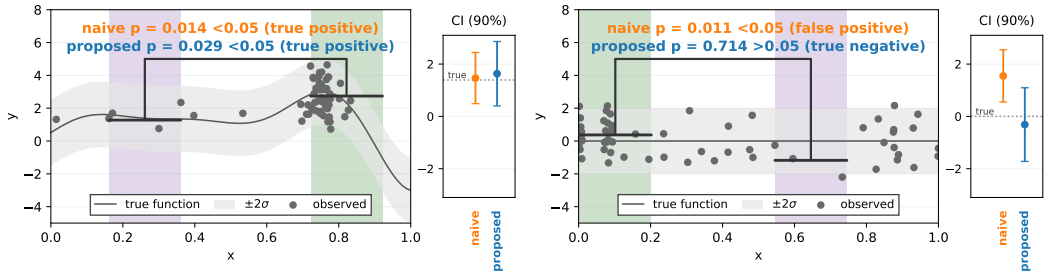

Figure 1: Schematic illustration of a post-ADC analysis. The left panel shows the case where the true black-box function depends on x, indicating underlying differences between regions. In contrast, the right panel shows the case where it does not. Dots indicate data points collected through an SMBO process. Based on observed data, regions with high (green) and low (purple) estimated means are selected, and statistical inference (p-values and confidence intervals) is performed to compare the two regions. Naive inference methods can lead to false-positive findings and miscalibrated confidence intervals, whereas the proposed inference approach provides valid inference.

The goal of this paper is to develop a principled framework for valid statistical inference in post-ADC analysis. Our approach is based on selective inference (SI), which enables valid inference for data-dependent inferential targets by conditioning on the selection event [6–9]. We refer to this framework as *post-ADC inference*. In SI, inference is performed not by treating the target as fixed in advance, but by conditioning on the fact that it was selected after observing the data. The key novelty of our setting is that selection occurs at two levels: not only is the inferential target constructed from the data, but the data themselves are collected through an adaptive process. We therefore treat the entire pipeline—from the ADC process to the data-driven construction of the inferential target—as a single selection event and perform inference conditional on it. This allows us to properly account for the bias induced by ADC. We show that, for a broad class of ADC algorithms, including GP-UCB and TPE, the conditional distribution of the test statistic can be characterized in a closed form. This enables the construction of valid p-values that control type-I error and confidence intervals with valid coverage, even when the data are reused for different tasks after ADC. The proposed post-ADC inference provides exact finite-sample guarantees of statistical significance and coverage in a finite-candidate setting under Gaussian observation noise with known variance. It applies to a range of ADC procedures and data-dependent target construction rules satisfying structural conditions. Importantly, it requires no assumptions on the underlying black-box function or the quality of the surrogate model.

Related work. SMBO is a standard framework for black-box optimization, in which a surrogate model is fitted to past observations and the next evaluation point is selected based on its predictions. Depending on the design of the surrogate model, SMBO methods can be broadly classified into three categories [1–5]. The first category consists of Gaussian process–based methods, which model the objective function with a Gaussian process (GP) surrogate and use acquisition functions, such as expected improvement and GP-UCB, to select evaluation points based on predictive means and uncertainty estimates [1, 2]. The second category consists of density estimation-based methods such as TPE, which use density ratios between high- and low-performing regions and work well with conditional or discrete search spaces [3]. The third category consists of ensemble-based methods such as sequential model-based algorithm configuration (SMAC), which use random forest surrogates and

are robust in nonlinear mixed-type spaces [4]. Despite these differences, all concentrate evaluations in regions that past observations make promising, so the collected data are typically not i.i.d. but biased. Existing SMBO research has therefore emphasized optimization performance, sample efficiency, and regret, while the validity of post hoc inference from actively collected data remains underexplored.

SI is a framework for valid statistical inference after hypotheses or targets are selected from the data [6–9]. Early work focused on variable selection in linear regression, including Lasso, marginal screening, and stepwise selection [7–18]. It has since been extended to clustering, change-point detection, and image or deep learning interpretation, yielding error-rate and coverage guarantees for data-driven targets [19–24]. Most SI studies, however, assume a fixed sampling mechanism and address selection only after the dataset is observed. Related work on adaptive experiments and bandits treats sequentially changing data collection, for example in treatment assignment or arm selection [25–27]. For post-ADC analysis, where the surrogate model and acquisition function make the sampling rule highly data-dependent, a principled SI framework remains largely unexplored. In this study, we propose an SI framework that treats the ADC process and the construction of post hoc inferential targets in a unified manner.

Contributions. Our contributions are threefold. First, in § 4, we formulate post-ADC inference by conditioning on a unified selection event that combines the ADC process and the data-driven construction of the inferential target, thereby yielding valid selective p-values and confidence intervals. Second, in § 5, we show that, for a class of ADC processes satisfying our structural conditions, the conditional distribution of the test statistic admits a closed-form expression, enabling exact finite-sample computation of these quantities. Third, as PoC studies, we consider the post-ADC setting for data collected by GP-UCB and TPE, and empirically demonstrate that the post-ADC inference produces valid inferential results. For reproducibility, we provide the implementation code at `https://github.com/ni-shu/paper_post-adc-inference`.

2 Active Data Collection

We consider ADC on a finite candidate set $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\} \subset \mathbb{R}^d$. At step $n \in [N-1] := \{1, 2, \dots, N-1\}$, the algorithm has access to the dataset $\mathcal{D}_n = ((\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)) \in (\mathcal{X} \times \mathbb{R})^n$. An ADC algorithm is represented by a sequence of update rules $\mathcal{A}^{\mathrm{ADC}} : (\mathcal{X} \times \mathbb{R})^n \to \mathcal{X}$, and the next query point $\boldsymbol{x}_{n+1}$ is given by

$$\boldsymbol{x}_{n+1} = \mathcal{A}^{\mathrm{ADC}}(\mathcal{D}_n).$$

Often, this update rule is expressed as the maximization of an acquisition score $a(\boldsymbol{x}; \mathcal{D}_n)$:

$$\boldsymbol{x}_{n+1} = \arg\max_{\boldsymbol{x} \in \mathcal{X}} a(\boldsymbol{x}; \mathcal{D}_n).$$

The following TPE and GP-UCB are representative examples of ADC algorithms that arise in this form. For notational simplicity, we write $\boldsymbol{x}_{1:n} := (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ and $\boldsymbol{y}_{1:n} := (y_1, \dots, y_n)^\top$.

Example 1: TPE. Fix a quantile level $\gamma \in (0, 1)$ and set $m_n = \lceil \gamma n \rceil$. For $\boldsymbol{y}_{1:n}$, let $\mathcal{H}_n(\boldsymbol{y}_{1:n}) \subset [n]$ denote the indices of the top m_n responses and let $\mathcal{L}_n(\boldsymbol{y}_{1:n}) = [n] \setminus \mathcal{H}_n(\boldsymbol{y}_{1:n})$ denote the remaining indices. Given a positive kernel $\kappa_{\mathrm{TPE}} : \mathcal{X} \times \mathcal{X} \to (0, \infty)$, define

$$g_n(\boldsymbol{x}) = \frac{1}{|\mathcal{H}_n(\boldsymbol{y}_{1:n})|} \sum_{i \in \mathcal{H}_n(\boldsymbol{y}_{1:n})} \kappa_{\mathrm{TPE}}(\boldsymbol{x}, \boldsymbol{x}_i), \qquad \ell_n(\boldsymbol{x}) = \frac{1}{|\mathcal{L}_n(\boldsymbol{y}_{1:n})|} \sum_{i \in \mathcal{L}_n(\boldsymbol{y}_{1:n})} \kappa_{\mathrm{TPE}}(\boldsymbol{x}, \boldsymbol{x}_i).$$

Then, the TPE acquisition score is

$$a^{\mathrm{TPE}}(\boldsymbol{x}; \mathcal{D}_n) = \frac{g_n(\boldsymbol{x})}{\ell_n(\boldsymbol{x})}. \tag{1}$$

The corresponding ADC algorithm $\mathcal{A}^{\mathrm{TPE}}$ is any deterministic maximizer of $a^{\mathrm{TPE}}(\boldsymbol{x}; \mathcal{D}_n)$ over $\mathcal{X}$.

Example 2: GP-UCB. Suppose the optimizer uses a Gaussian-process surrogate with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then the surrogate posterior mean and variance are

$$\mu_n(\boldsymbol{x}; \boldsymbol{y}_{1:n}) = \boldsymbol{k}_n(\boldsymbol{x})^\top (K_n + \sigma^2 I_n)^{-1} \boldsymbol{y}_{1:n}, \qquad s_n^2(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}_n(\boldsymbol{x})^\top (K_n + \sigma^2 I_n)^{-1} \boldsymbol{k}_n(\boldsymbol{x}),$$

where $K_n \in \mathbb{R}^{n\times n}$ has entries $[K_n]_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{k}_n(\boldsymbol{x}) = (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n))^\top \in \mathbb{R}^n$, and σ^2 is the response-noise variance. Under this surrogate, the GP-UCB acquisition score is

$$a^{\text{GP-UCB}}(\boldsymbol{x}; \mathcal{D}_n) = \mu_n(\boldsymbol{x}; \boldsymbol{y}_{1:n}) + \kappa s_n(\boldsymbol{x}), \tag{2}$$

where $\kappa > 0$ controls the exploration-exploitation tradeoff. The corresponding ADC algorithm $\mathcal{A}^{\text{GP-UCB}}$ is any deterministic maximizer of $a^{\text{GP-UCB}}(\boldsymbol{x}; \mathcal{D}_n)$ over $\mathcal{X}$.

Examples 1 and 2 illustrate representative ADC algorithms covered by our framework. More generally, the framework applies to a broad class of ADC algorithms, which is formally characterized in Appendix D.2. Roughly speaking, it covers algorithms whose acquisition rules can be decomposed into a sequence of piecewise-linear operations. For rules such as (1) and (2), this condition is satisfied whenever comparisons between candidate query points can be characterized by linear inequalities in the responses $\boldsymbol{y}_{1:n}$. Consequently, the framework applies to many SMBO algorithms, including the TPE and GP-UCB methods considered here.

We fix a collection budget N and run the ADC algorithm for N queries, producing the collected dataset $\mathcal{D}_N$. Our goal is not to analyze or improve those ADC methods themselves. Rather, our goal is to carry out statistically valid inference that properly accounts for the data collection bias created by those ADC mechanisms.

3 Target of Inference and Difficulty of Naive Inference

To formulate post-ADC inference, we now introduce the data model and the target of inference.

3.1 Data Model for Post-ADC Inference

Let $f_\star : \mathcal{X} \to \mathbb{R}$ be an unknown non-random objective function. At ADC step i, the algorithm queries a point $\boldsymbol{x}_i \in \mathcal{X}$ and observes

$$Y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{3}$$

where the noise variance $\sigma^2 > 0$ is assumed to be either known or estimable from an independent dataset. Importantly, the validity results below rely only on the noise model in (3) and do not require any assumption on $f_\star$ or on the accuracy of the surrogate model used by the ADC algorithm.

We regard the responses $y_1, \ldots, y_N$ observed during ADC in § 2 as realizations of (3). In the following, we use uppercase notation for random variables, $\boldsymbol{Y}_{1:n} = (Y_1, \ldots, Y_n)^\top$, and lowercase notation for realizations, $\boldsymbol{y}_{1:n} = (y_1, \ldots, y_n)^\top$. Similarly, we write the dataset up to step n as $\mathcal{D}_n(\boldsymbol{Y}_{1:n}) = ((\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n))$ when random, and as $\mathcal{D}_n(\boldsymbol{y}_{1:n}) = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ when realized. Furthermore, for notational simplicity, let $\boldsymbol{Y} = \boldsymbol{Y}_{1:N} \in \mathbb{R}^N$ denote the full response vector after ADC, and let $\boldsymbol{y} = \boldsymbol{y}_{1:N} \in \mathbb{R}^N$ denote its realization.

3.2 Target of Inference and Validity Goals

We now define a post-ADC target of inference based on the collected data $\mathcal{D}_N(\boldsymbol{Y})$. First, we introduce representative examples of data-driven targets in Examples 1 and 2.

Example 1: Two-sample inference for high versus low region. This example is illustrated in Figure 1. The high and low regions are defined as the axis-aligned hypercubes of side length ℓ with the largest and smallest average observed responses, respectively:

$$\mathcal{I}_{\boldsymbol{Y}} \in \underset{\mathcal{C}(\boldsymbol{s};\ell)|\mathcal{C}(\boldsymbol{s};\ell)\neq\emptyset, \boldsymbol{s}\in\mathbb{R}^d}{\arg\max} \bar{Y}(\mathcal{C}), \qquad \mathcal{J}_{\boldsymbol{Y}} \in \underset{\mathcal{C}(\boldsymbol{s};\ell)|\mathcal{C}(\boldsymbol{s};\ell)\neq\emptyset, \boldsymbol{s}\in\mathbb{R}^d}{\arg\min} \bar{Y}(\mathcal{C}), \tag{4}$$

where $\ell > 0$ is the side length of the sliding window, $\boldsymbol{s} \in \mathbb{R}^d$ is a sliding-window anchor, $\mathcal{C}(\boldsymbol{s}; \ell) = \{t \in [N] : \boldsymbol{x}_t \in \prod_{j=1}^d [s_j, s_j + \ell]\}$ is the sliding-window index set, $\bar{Y}(\mathcal{C}) = |\mathcal{C}|^{-1} \sum_{t\in\mathcal{C}} Y_t$ is the sample mean over the sliding window $\mathcal{C}$, and $\emptyset$ is the empty set. The target of inference is then the true mean difference between these two regions:

$$\Delta_{\boldsymbol{Y}} = \frac{1}{|\mathcal{I}_{\boldsymbol{Y}}|} \sum_{t\in\mathcal{I}_{\boldsymbol{Y}}} f_\star(\boldsymbol{x}_t) - \frac{1}{|\mathcal{J}_{\boldsymbol{Y}}|} \sum_{t\in\mathcal{J}_{\boldsymbol{Y}}} f_\star(\boldsymbol{x}_t).$$

This problem is similar to a standard two-sample test, but differs in that the comparison target is selected in a data-driven manner after the ADC process. For example, in applications such as scientific parameter optimization, this target corresponds to the true performance gap between high- and low-performance regions in the parameter space.

Example 2: One-sample inference for top-n performance. For a given integer $n \in [N]$, let

$$\mathcal{I}_{\boldsymbol{Y}}^{\mathrm{top},n} \in \underset{\mathcal{S}\subset[N]:\ |\mathcal{S}|=n}{\arg\max} \sum_{t\in\mathcal{S}} Y_t.$$

The target of inference is the average true response over the selected top-n queried points:

$$\Delta_{\boldsymbol{Y}}^{\mathrm{top}} = \frac{1}{n} \sum_{t\in\mathcal{I}_{\boldsymbol{Y}}^{\mathrm{top},n}} f_\star(\boldsymbol{x}_t).$$

This problem can be viewed as a data-driven one-sample inference problem. For example, in applications such as hyperparameter optimization of machine learning models, this target corresponds to the true performance of the top-n parameter configurations found by ADC.

Both Examples 1 and 2 are special cases of a general class of data-dependent targets of the form

$$\Delta_{\boldsymbol{Y}} = \boldsymbol{\eta}_{\boldsymbol{Y}}^\top \boldsymbol{\mu}, \tag{5}$$

where $\boldsymbol{\mu} = (f_\star(\boldsymbol{x}_1), \ldots, f_\star(\boldsymbol{x}_N))^\top$. For Example 1, the weight vector is $\boldsymbol{\eta}_{\boldsymbol{Y}}^{\mathrm{reg}} := |\mathcal{I}_{\boldsymbol{Y}}|^{-1}\mathbf{1}_{\mathcal{I}_{\boldsymbol{Y}}} - |\mathcal{J}_{\boldsymbol{Y}}|^{-1}\mathbf{1}_{\mathcal{J}_{\boldsymbol{Y}}}$, where $\mathbf{1}_{\mathcal{C}} \in \{0,1\}^N$ is the indicator vector of $\mathcal{C} \subset [N]$. For Example 2, the weight vector is $\boldsymbol{\eta}_{\boldsymbol{Y}}^{\mathrm{top}} := n^{-1}\mathbf{1}_{\mathcal{I}_{\boldsymbol{Y}}^{\mathrm{top},n}}$. Note that $\boldsymbol{\eta}_{\boldsymbol{Y}}$ is a random variable that depends on the data $\boldsymbol{Y}$. Naturally, a test statistic for (5) is defined by

$$T(\boldsymbol{Y}) = \boldsymbol{\eta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}.$$

Examples 1 and 2 show typical data-dependent targets covered by our framework. The common feature behind these examples is that the target is constructed after observing the data, through the map $\boldsymbol{Y} \mapsto \boldsymbol{\eta}_{\boldsymbol{Y}}$. Our framework covers a broad class of such target-construction rules, formally defined in Appendix D.2. Informally, the required condition is that the construction of $\boldsymbol{\eta}_{\boldsymbol{Y}}$ can be described by a sequence of piecewise-linear operations. This condition is satisfied, for example, when the weight vector is chosen through comparisons among the observed responses, as long as those comparisons can be characterized as linear inequalities in $\boldsymbol{Y}$. Thus, the framework includes a wide range of data-dependent targets, with data-independent targets appearing as a special case. The two examples above satisfy this requirement. Appendix A provides several additional examples.

One inference goal is to construct a valid confidence interval for $\Delta_{\boldsymbol{Y}}$ using the test statistic $T(\boldsymbol{Y})$. A valid $(1-\alpha)$-confidence interval $C_{1-\alpha}^{\Delta}(\boldsymbol{Y})$ should satisfy

$$\mathbb{P}\big(\Delta_{\boldsymbol{Y}} \in C_{1-\alpha}^{\Delta}(\boldsymbol{Y})\big) = 1-\alpha. \tag{6}$$

In Example 1, this corresponds to estimating the true performance difference between the high-performance and low-performance regions, while controlling the miscoverage probability at α.

Another inference goal is to construct a valid p-value for testing the hypotheses:

$$\mathrm{H}_{0,\boldsymbol{Y}}:\ \Delta_{\boldsymbol{Y}} \le 0 \qquad \text{vs.} \qquad \mathrm{H}_{1,\boldsymbol{Y}}:\ \Delta_{\boldsymbol{Y}} > 0. \tag{7}$$

Formally, a valid p-value should satisfy

$$\mathbb{P}_{\mathrm{H}_0}(p \le \alpha) \le \alpha \qquad \text{for all } \alpha \in (0,1). \tag{8}$$

In Example 1, this tests whether the high-performance region significantly outperforms the low-performance region, rather than reflecting random fluctuation, while controlling the false finding probability at level α.

Although we have focused here on a one-sided test and a two-sided confidence interval, our framework also covers the corresponding inverted procedures, namely the equal-tailed two-sided test and the lower one-sided confidence interval. Appendix B provides their formal definitions and properties.

3.3 Why Naive Inference Fails

The key challenge in this post-ADC analysis is that the same response vector $\boldsymbol{Y}$ is used three times in this workflow. First, $\boldsymbol{Y}$ determines the ADC trajectory $\mathcal{T}_{\boldsymbol{Y}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, i.e., the sequence of queried locations. Thus, the ADC trajectory may tend to select locations with favorable outcomes for $\boldsymbol{Y}$. For example, in an ADC procedure such as SMBO for maximization, regions with high observed values tend to be explored more intensively, whereas regions that appear low due to random noise may be left underexplored. Second, after this data-driven trajectory has been determined, the same $\boldsymbol{Y}$ are used to define the regions $\mathcal{I}_{\boldsymbol{Y}}$ and $\mathcal{J}_{\boldsymbol{Y}}$, and hence the weight vector $\boldsymbol{\eta}_{\boldsymbol{Y}}$, the target $\Delta_{\boldsymbol{Y}}$, and the hypotheses in (7). The important point is that $\boldsymbol{\eta}_{\boldsymbol{Y}}$ is itself a function of $\boldsymbol{Y}$. For example, if the regions are chosen so that the observed difference along the ADC trajectory appears large, then the resulting $\boldsymbol{\eta}_{\boldsymbol{Y}}$ can reflect not only the underlying signal but also favorable random fluctuations in the observed responses. Finally, the same $\boldsymbol{Y}$ are used to compute the test statistic $T(\boldsymbol{Y})$ and to perform inference. Naive post-ADC inference ignores these multiple uses of data and treats the queried locations and the weight vector $\boldsymbol{\eta}_{\boldsymbol{Y}}$ as if they had been fixed in advance. This is a statistically inappropriate procedure, and the validity guarantees in (8) and (6) are no longer guaranteed to hold. As a result, naive post-ADC inference can exhibit inflated type-I error and confidence intervals whose coverage falls below the nominal level. This is the same selection-bias phenomenon studied throughout the selective-inference literature [7, 8]. We will later confirm these failures empirically in § 6.

4 Selective Inference for Post-ADC Analysis

As discussed in § 3.3, the main obstacle in post-ADC inference is that the locations of the queried points, the definition of the inferential target, and the test statistic all depend on the same data $\boldsymbol{Y}$. We now propose a post-ADC inference framework to overcome this obstacle by leveraging SI principles. The key idea is to condition on the selection event that determines the inferential target. Intuitively, conditioning restricts attention to possible response vectors that would lead the ADC algorithm to query the same points in the same order and select the same inferential target. This accounts for the fact that both the acquisition trajectory and the inferential target depend on the observed data.

4.1 Conditioning on the Selection Event

In the setup described in § 3.1 and § 3, the data are used to select both the queried trajectory $\mathcal{T}_{\boldsymbol{Y}}$ and the data-dependent weight vector $\boldsymbol{\eta}_{\boldsymbol{Y}}$ that defines the inferential target. To account for the bias induced by these data dependencies, we condition on the entire selection process $\mathcal{E}_{\boldsymbol{Y}} = (\mathcal{T}_{\boldsymbol{Y}}, \boldsymbol{\eta}_{\boldsymbol{Y}})$ by its realization $\mathcal{E}_{\boldsymbol{y}}$ and perform inference conditional on the selection event:

$$\{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}\} = \{\mathcal{T}_{\boldsymbol{Y}} = \mathcal{T}_{\boldsymbol{y}},\ \boldsymbol{\eta}_{\boldsymbol{Y}} = \boldsymbol{\eta}_{\boldsymbol{y}}\}.$$

We then consider the conditional distribution of the test statistic given the selection event:

$$T(\boldsymbol{Y}) \,\big|\, \{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}\}. \tag{9}$$

This conditioning plays two key roles.

Conditioning on $\mathcal{T}_{\boldsymbol{Y}}$. If the queried trajectory $\mathcal{T}_{\boldsymbol{Y}}$ were fixed in advance, then $\boldsymbol{Y}$ follows the fixed-mean Gaussian model:

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \Sigma), \qquad \boldsymbol{\mu} = (f_\star(\boldsymbol{x}_1), \ldots, f_\star(\boldsymbol{x}_N))^\top, \qquad \Sigma = \sigma^2 I_N. \tag{10}$$

Conditioning on $\mathcal{T}_{\boldsymbol{Y}} = \mathcal{T}_{\boldsymbol{y}}$ therefore restricts the above Gaussian distribution to the subset of $\boldsymbol{Y}$ that reproduce the same queried trajectory $\mathcal{T}_{\boldsymbol{y}}$. Although the resulting conditional distribution is not an ordinary Gaussian distribution, it retains the Gaussian density structure, making it substantially more tractable than the unconditional distribution induced by the ADC process.

Conditioning on $\boldsymbol{\eta}_{\boldsymbol{Y}}$. This fixes the target of inference $\Delta_{\boldsymbol{Y}} = \boldsymbol{\eta}_{\boldsymbol{Y}}^\top \boldsymbol{\mu}$ and the corresponding hypotheses in (7). For the Example 1 in § 3.2, fixing $\boldsymbol{\eta}_{\boldsymbol{Y}}$ is equivalent to fixing the selected high and low regions. Let $\boldsymbol{\eta} = \boldsymbol{\eta}_{\boldsymbol{y}}$ and $T_{\boldsymbol{\eta}}(\boldsymbol{Y}) = \boldsymbol{\eta}^\top \boldsymbol{Y}$. Then the inferential target in (5) is fixed at $\Delta^{\mathrm{sel}} = \boldsymbol{\eta}^\top \boldsymbol{\mu}$ and the hypotheses in (7) become $\mathrm{H}_0 : \Delta^{\mathrm{sel}} \leq 0$ vs. $\mathrm{H}_1 : \Delta^{\mathrm{sel}} > 0$.

Thus, after conditioning on the selection event, the statistic of interest $\boldsymbol{\eta}^\top \boldsymbol{Y}$ becomes a linear function of $\boldsymbol{Y}$, while the conditional distribution retains the Gaussian density structure restricted to the selected data space. This allows us to characterize the conditional distribution in (9) in a relatively tractable form in § 5.

4.2 Selective p-Values and Confidence Intervals

The conditional distribution in (9) still involves nuisance parameters in addition to the effect Δ^{sel}. Under (10), decompose the mean vector as

$$\boldsymbol{\mu} = \underbrace{\Delta^{\mathrm{sel}} \frac{\Sigma\boldsymbol{\eta}}{\boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta}}}_{\text{parameter of interest}} + \underbrace{\left(I_N - \frac{\Sigma\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta}}\right)\boldsymbol{\mu}}_{\text{nuisance parameter (denoted as } \boldsymbol{\mu}^{\mathrm{nui}})} . \tag{11}$$

Thus Δ^{sel} is the parameter of interest, while $\boldsymbol{\mu}^{\mathrm{nui}}$ collects the nuisance component orthogonal to the selected weight vector. To identify the conditional distribution in (9), we need to eliminate the nuisance parameter, which is difficult to marginalize out. We therefore condition on its sufficient statistic, namely the residual after projection onto the weight-vector direction:

$$\mathcal{Q}_{\boldsymbol{Y}} = \left(I_N - \frac{\Sigma\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta}}\right)\boldsymbol{Y}. \tag{12}$$

A formal statement of this sufficient statistic and its proof are provided in Appendix B.1 and C.1 respectively. Therefore, the conditional test statistic used for inference is

$$T_{\boldsymbol{\eta}}(\boldsymbol{Y}) \mid \{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\}. \tag{13}$$

The corresponding cumulative distribution function (CDF) of the statistic is defined as

$$G_{\Delta}^{\mathrm{sel}}(t) = \mathbb{P}_{\Delta^{\mathrm{sel}}=\Delta}(T_{\boldsymbol{\eta}}(\boldsymbol{Y}) \le t \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}). \tag{14}$$

The selective p-value for the directional claim in (7) is evaluated at the boundary value $\Delta^{\mathrm{sel}} = 0$:

$$p_{\mathrm{sel}}(\boldsymbol{y}) = \mathbb{P}_{\Delta^{\mathrm{sel}}=0}(T_{\boldsymbol{\eta}}(\boldsymbol{Y}) \ge T_{\boldsymbol{\eta}}(\boldsymbol{y}) \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}). \tag{15}$$

The same conditional distribution also yields a selective confidence interval:

$$C_{1-\alpha}^{\Delta}(\boldsymbol{y}) = \left\{\Delta \in \mathbb{R} : \alpha/2 \le G_{\Delta}^{\mathrm{sel}}(T_{\boldsymbol{\eta}}(\boldsymbol{y})) \le 1 - \alpha/2\right\}. \tag{16}$$

Appendix B provides the corresponding complementary one-sided and two-sided selective outputs. At this point the inferential outputs are fully specified by selective-inference principles, but their computation still depends on the conditional distribution in (14). The next section shows that the conditional distribution reduces to a truncated normal distribution that can be evaluated exactly.

5 Computing Selective p-Values and Confidence Intervals

In this section, we show that, for a broad class of ADC processes, the conditional distribution of the test statistic in (9) admits a closed-form expression, and thus the selective p-values and confidence intervals can be computed exactly under the finite-sample setup.

5.1 Conditional Distribution of the Test Statistic

As discussed in § 4, after conditioning on the selection event, the distribution for the data $\boldsymbol{Y}$ retains the Gaussian density structure in (10), and the data-dependent selected weight vector $\boldsymbol{\eta}_{\boldsymbol{Y}}$ is fixed at its realized value $\boldsymbol{\eta}$. Consequently, the test statistic follows normal distribution, but truncated to the subset of $\boldsymbol{Y}$ that reproduce the same selection event as the observed data. This observation is formalized in the following lemma.

Lemma 1. *Assume the Gaussian distribution in* (10)*, and let $\boldsymbol{y}$ be the observed realization of $\boldsymbol{Y}$ with $\boldsymbol{\eta}_{\boldsymbol{y}} \neq \boldsymbol{0}$. Fix $\boldsymbol{\eta} = \boldsymbol{\eta}_{\boldsymbol{y}}$. Define*

$$\boldsymbol{a} = \mathcal{Q}_{\boldsymbol{y}}, \qquad \boldsymbol{b} = \frac{\Sigma\boldsymbol{\eta}}{\boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta}}.$$

Then every response vector on the resulting one-dimensional slice can be written as

$$\boldsymbol{Y}(z) = \boldsymbol{a} + \boldsymbol{b}z, \qquad z = T(\boldsymbol{Y}(z)). \tag{17}$$

Let

$$\mathcal{Z} = \left\{ z \in \mathbb{R} : \mathcal{E}_{\boldsymbol{Y}(z)} = \mathcal{E}_{\boldsymbol{y}} \right\}.$$

Then the conditional distribution of $T(\boldsymbol{Y})$ given $\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$ and $\mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}$ is a truncated normal distribution:

$$T(\boldsymbol{Y}) \mid \{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\} \sim \mathrm{TN}(\Delta^{\mathrm{sel}}, v_{\boldsymbol{\eta}}, \mathcal{Z})$$

with mean Δ^{sel} and variance $v_{\boldsymbol{\eta}} = \boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta}$, truncated to $\mathcal{Z}$.

The proof is given in Appendix C.2. Conditioning on the nuisance statistic restricts the response vector to the one-dimensional line $\boldsymbol{Y}(z) = \boldsymbol{a} + \boldsymbol{b}z$. Preserving the observed ADC trajectory and selected inferential target further restricts z to $\mathcal{Z}$, an interval or a finite union of intervals determined by the selection constraints. Appendix D.1 illustrates this construction through a concrete toy example and schematic illustration, while § 5.2 describes the general computation. The selective CDF in (14) is therefore evaluated using the normal distribution truncated to $\mathcal{Z}$. The selective p-value in (15) is computed from the upper tail of this distribution under the boundary null, and the selective confidence interval in (16) is obtained by inverting the same CDF. Furthermore, the selective p-value is valid, in the sense that it controls the type-I error under the null hypothesis H_0 in (7). Similarly, the confidence interval achieves the valid coverage for the selected effect Δ^{sel}. These results are formally stated in the following theorems.

Theorem 1. *Under the setup of Lemma 1, suppose that $\mathbb{P}(\boldsymbol{\eta}_{\boldsymbol{Y}} \neq \mathbf{0}) = 1$. The p-value defined in* (15) *satisfies*

$$\mathbb{P}_{\mathrm{H}_0}(p_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha) \leq \alpha, \quad \forall \alpha \in (0, 1). \tag{18}$$

Theorem 2. *Under the assumptions of Theorem 1, the confidence interval defined in* (16) *satisfies*

$$\mathbb{P}\big(\Delta_{\boldsymbol{Y}} \in C^{\Delta}_{1-\alpha}(\boldsymbol{Y})\big) = 1 - \alpha, \quad \forall \alpha \in (0, 1).$$

Their proofs are given in Appendix C.3. Theorems 1 and 2 show that the proposed post-ADC inference enables us to draw insights from ADC data while controlling the probability of false findings at a pre-specified level α. Proposition 2 in Appendix B.2 further proves the stronger statement that, under the simple null hypothesis $\Delta^{\mathrm{sel}} = 0$, the selective p-value is exactly uniform. Appendix B also provides the formal definition and properties of the two-sided selective test and the one-sided lower confidence interval, which are also valid by the same argument as above.

5.2 Computation of the Truncation Set $\mathcal{Z}$

To obtain an exact computational procedure from the conditional distribution above, we must determine the set $\mathcal{Z}$ of statistic values z that reproduce the same observed selection event as the observed data. Due to space limitations, we omit the details of how to compute $\mathcal{Z}$; see Appendix D and the accompanying code for the full procedure. Roughly speaking, the piecewise-linearity of ADC algorithms and target-construction rules introduced in § 2 and § 3 and formalized in Appendix D.2 imply that the selection event can be represented in the response space by finitely many linear equalities and inequalities in $\boldsymbol{Y}$. Restricting those constraints to the one-dimensional slice $\boldsymbol{Y}(z) = \boldsymbol{a} + \boldsymbol{b}z$ from Lemma 1 turns the problem of computing $\mathcal{Z}$ into solving finitely many linear inequalities in z. These structural conditions are satisfied by GP-UCB with fixed hyperparameters and the TPE procedure defined in Section 2, with explicit derivations provided in Appendix D.3. Applicability to other ADC procedures depends on whether their complete decision processes satisfy the same conditions. Appendix F.8 further discusses the framework's implementability and computational cost, including empirical measurements of inference time and memory usage. Existing SI literature has highlighted the usefulness of randomization for stabilizing inferential outputs [11, 28]. Our framework also accommodates a randomized-response extension, in which independent Gaussian noise is added to the responses used for ADC and target construction, while inference is performed using the original responses. This randomized-response extension retains the validity guarantees and is detailed in Appendix E.

6 Experiments

We evaluate post-ADC inference on synthetic and real data, assessing type I error, coverage, and power. Unless stated otherwise, we use dimension $d = 3$, Gaussian noise variance $\sigma^2 = 1$, $N_{\mathrm{init}} = 10$, and

$N_{\text{steps}} = 50$, giving $N = N_{\text{init}} + N_{\text{steps}}$ samples. We compare TPE (1) and GP-UCB (2), using the high-region-versus-low-region target from § 3. Throughout the experiments, we use the randomized-response extension introduced in § 5.2. Tests use $\alpha = 0.05$, intervals use 0.90 coverage, and rates are estimated from $1{,}000$ independent Monte Carlo replicates per configuration. Implementation details are in Appendix F.1. We compare the proposed method (`post-ADC`) against two baselines: `naive`, a Z-test without selection correction, and `Bonferroni`, which corrects for $|\mathcal{X}|^{N_{\text{steps}}} \cdot 3^N$ possible trajectories and region selections.

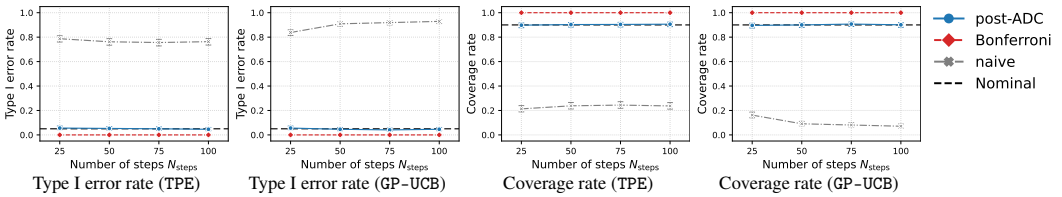

Figure 2: Empirical type I error and coverage across ADC horizons. `post-ADC` remains close to the nominal levels, while `naive` is anti-conservative and `Bonferroni` is conservative. The dashed lines mark the nominal type I error level $\alpha = 0.05$ and coverage level 0.90; both are identified as Nominal in the shared legend.

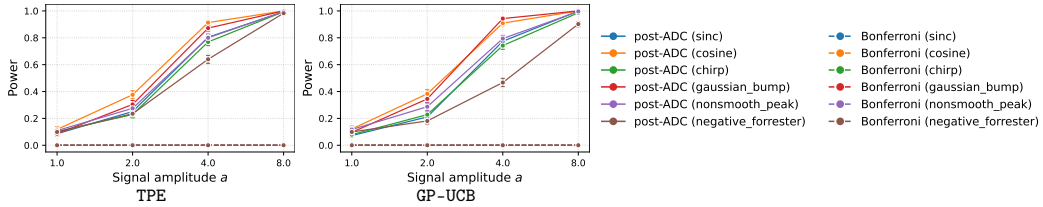

Figure 3: Empirical power. `post-ADC` gains power as the signal increases across six synthetic objective families, while all `Bonferroni` results (dashed lines) remain near zero.

Table 1: Empirical rejection probability on four real datasets.

Acquisition	Method	Concrete	Gas Turbine (CO)	Gas Turbine (NOx)	Power Plant
ucb	`post-ADC`	**0.432**	**0.995**	**0.911**	**0.999**
	`Bonferroni`	0.000	0.014	0.000	0.000
tpe	`post-ADC`	**0.396**	**0.989**	**0.918**	**0.929**
	`Bonferroni`	0.000	0.058	0.000	0.000

6.1 Type I Error and Coverage

Under the global null $f_\star \equiv 0$, every data-dependent inferential target has zero effect, allowing us to assess type I error and coverage. We vary $N_{\text{steps}} \in \{25, 50, 75, 100\}$ and report the rejection probability at $\alpha = 0.05$ and coverage of nominal 90% confidence intervals in Figure 2. `post-ADC` stays near the nominal levels, while `naive` is anti-conservative and `Bonferroni` is overly conservative, matching Theorem 1 and Theorem 2. Additional dimensionality and hyperparameter sensitivity analyses are provided in Appendices F.4 and F.5, respectively, while interval-width results are reported in Appendix F.3. We also investigate type I error rates under non-Gaussian observation noise and with plug-in estimation of the observation-noise variance (Appendices F.6 and F.7, respectively). These experiments assess empirical robustness beyond the assumptions underlying our exact finite-sample guarantees.

6.2 Power Experiments

To evaluate detection performance beyond the global-null setting, we consider six synthetic objective families (`sinc`, `cosine`, `chirp`, `gaussian_bump`, `nonsmooth_peak`, `negative_forrester`),

rescaled to $[-1, 1]$ and multiplied by $a \in \{1, 2, 4, 8\}$; functional forms are in Appendix F.2. Because `naive` is anti-conservative under the null (Figure 2), we compare only `post-ADC` and `Bonferroni` in Figure 3. Despite the potential power loss from conditioning on the ADC trajectory and selected inferential target, `post-ADC` gains power as a increases, whereas `Bonferroni` remains near zero.

6.3 Experiments on Real Data

We use four UCI benchmark datasets as black-box objectives, with dataset, feature-selection, and noise-estimation details in Appendix F.9. For each configuration, we run $1{,}000$ bootstrap replicates. Table 1 gives the $d = 1$ results; `post-ADC` detects signal on all datasets, whereas `Bonferroni` is essentially powerless. Because the true effects are unknown and the observation-noise variance is estimated, these experiments do not directly assess calibration under the exact theoretical assumptions. Instead, they illustrate the empirical behavior of the proposed method in realistic settings with noise-estimation. Analogous results for varying d are in Appendix F.9.

7 Discussion and Limitations

We introduced post-ADC inference for valid inference after active data collection. The key issue is that the same noisy observations determine both where ADC samples and which statistical question is later tested. By treating this pipeline as a single selection event, we derive exact selective p-values and confidence intervals under the stated assumptions. This enables post hoc inference with selective type I error control and confidence interval coverage. Several scope restrictions remain important. First, the exact theory assumes a finite candidate set. A prespecified grid can be used to discretize a continuous domain, but the exact guarantee then applies to the ADC procedure and inferential target defined on that grid, rather than to quantities defined over the original continuous domain. Second, the exact finite-sample guarantee assumes Gaussian observation noise with known variance. Our additional experiments examine non-Gaussian noise and plug-in variance estimation, but extending exact finite-sample guarantees to these settings remains open. Third, our exact computational formulation applies to ADC and target-construction rules expressible as compositions of finitely many piecewise-linear maps, and to data-dependent linear inferential targets satisfying our structural conditions. This includes the fixed-hyperparameter GP-UCB and ordering-based TPE variants considered here. Learning GP hyperparameters from the same ADC responses and constructing nonlinear inferential targets are natural directions for extending the framework. For procedures and targets outside these structural conditions, approximate selective-inference approaches based on numerical sampling or importance sampling may provide a possible extension [13]. Finally, we have not systematically analyzed power and interval width jointly as functions of the acquisition horizon and candidate-set size. Since changing the candidate set can alter both the acquisition trajectory and the selected target, these effects need not be monotone. A systematic empirical and theoretical investigation remains an important direction for future work.

Acknowledgments and Disclosure of Funding

This work was supported by RIKEN Junior Research Associate Program and partially supported by JST CREST (Grant Numbers JPMJCR21D3 and JPMJCR22N2), JSPS KAKENHI (Grant Number JP26H02528), and the RIKEN Center for Advanced Intelligence Project (AIP). The authors declare no competing interests.

References

[1] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. Journal of Global optimization, 13(4):455–492, 1998.

[2] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, page 1015–1022, Madison, WI, USA, 2010. Omnipress.

[3] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In Advances in Neural Information Processing Systems 24, pages 2546–2554, 2011.

[4] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In Carlos A. Coello Coello, editor, Learning and Intelligent Optimization, volume 6683 of Lecture Notes in Computer Science, pages 507–523. Springer, 2011.

[5] Peter I. Frazier. Bayesian optimization. In INFORMS TutORials in Operations Research: Recent Advances in Optimization and Modeling of Contemporary Problems, pages 255–278. INFORMS, 2018.

[6] Yoav Benjamini, Ruth Heller, and Daniel Yekutieli. Selective inference in complex research. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1906):4255–4271, 2009.

[7] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. arXiv preprint arXiv:1410.2597, 2014.

[8] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. The Annals of Statistics, 44(3):907–927, 2016.

[9] Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association, 111(514):600–620, 2016.

[10] Jonathan Taylor and Robert Tibshirani. Post-selection inference for-penalized likelihood models. Canadian Journal of Statistics, 46(1):41–61, 2018.

[11] Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. The Annals of Statistics, 46(2):679–710, 2018.

[12] Xiaoying Tian and Jonathan Taylor. Asymptotics of selective inference. Scandinavian Journal of Statistics, 44(2):480–499, 2017.

[13] Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. In Advances in Neural Information Processing Systems, pages 2469–2477, 2016.

[14] Keli Liu, Jelena Markovic, and Robert Tibshirani. More powerful post-selection inference, with application to the lasso. arXiv preprint arXiv:1801.09037, 2018.

[15] Sangwon Hyun, Max G'sell, and Ryan J Tibshirani. Exact post-selection inference for the generalized lasso path. Electronic Journal of Statistics, 12(1):1053–1097, 2018.

[16] Yoav Benjamini. Selective inference: The silent killer of replicability. 2020.

[17] Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Post-selection adaptive inference for least angle regression and the lasso. arXiv preprint arXiv:1401.3889, 354, 2014.

[18] William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection. arXiv preprint arXiv:1512.02565, 2015.

[19] Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. Journal of the American Statistical Association, pages 1–11, 2022.

[20] Vo Nguyen Le Duy, Shogo Iwazaki, and Ichiro Takeuchi. Quantifying statistical significance of neural network-based image segmentation by selective inference. Advances in Neural Information Processing Systems, 35:31627–31639, 2022.

[21] Daiki Miwa, Duy Vo Nguyen Le, and Ichiro Takeuchi. Valid p-value for deep learning-driven salient region. In Proceedings of the 11th International Conference on Learning Representation, 2023.

[22] Tomohiro Shiraishi, Tatsuya Matsukawa, Shuichi Nishino, and Ichiro Takeuchi. Statistical test for feature selection pipelines by selective inference. In Forty-second International Conference on Machine Learning, 2025.

[23] Tatsuya Matsukawa, Tomohiro Shiraishi, Shuichi Nishino, Teruyuki Katsuoka, and Ichiro Takeuchi. Statistical test for auto feature engineering by selective inference. In The 28th International Conference on Artificial Intelligence and Statistics, 2025.

[24] Jelena Markovic, Lucy Xia, and Jonathan Taylor. Unifying approach to selective inference with applications to cross-validation. arXiv preprint arXiv:1703.06559, 2017.

[25] Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. In Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 1261–1269. PMLR, 2018.

[26] Tobias Freidling, Qingyuan Zhao, and Zijun Gao. Selective randomization inference for adaptive experiments. arXiv preprint arXiv:2405.07026, 2024.

[27] Jiafeng Chen and Isaiah Andrews. Optimal conditional inference in adaptive experiments. arXiv preprint arXiv:2309.12162, 2023.

[28] Danijel Kivaranovic and Hannes Leeb. A (tight) upper bound for the length of confidence intervals with conditional coverage. Electronic Journal of Statistics, 18(1):1677–1701, 2024.

[29] Erich Leo Lehmann, Joseph P Romano, et al. Testing statistical hypotheses, volume 3. Springer, 1986.

A Additional Targets of Inference

The main text already discussed two data-dependent target constructions in § 3.2. Our framework applies more widely to any target construction rule, which is formally defined in Appendix D.2. This subsection collects additional examples of targets that fit our framework, including both data-dependent and data-independent targets.

Data-dependent targets. We summarize two additional data-dependent weight vector families in Table 2. They all fit the same linear form

$$\Delta_{\boldsymbol{y}} = \boldsymbol{\eta}_{\boldsymbol{y}}^\top \boldsymbol{\mu}.$$

Table 2: Additional data-dependent target families beyond the two main-text examples.

#	Name	$\mathcal{I}_{\boldsymbol{y}}$	$\mathcal{J}_{\boldsymbol{y}}$
1	Top-n vs. bottom-m	$\mathcal{I}_{\boldsymbol{y}} = \underset{\lvert\mathcal{S}\rvert=n}{\arg\max} \sum_{t\in\mathcal{S}} y_t$	$\mathcal{J}_{\boldsymbol{y}} = \underset{\lvert\mathcal{S}\rvert=m}{\arg\min} \sum_{t\in\mathcal{S}} y_t$
2	Winner vs. runner-up	$\mathcal{I}_{\boldsymbol{y}} = \{\hat{i}\}$, $\hat{i} = \underset{t\in[N]}{\arg\max}\, y_t$	$\mathcal{J}_{\boldsymbol{y}} = \{\tilde{j}\}$, $\tilde{j} = \underset{t\in[N]\setminus\{\hat{i}\}}{\arg\max}\, y_t$

Data-independent targets. The framework also covers useful data-independent targets, where the weight vector is fixed once the queried design points are fixed. A first example is the observed GP posterior mean at a target location $\boldsymbol{x}^\circ$,

$$\widehat{\mu}_N(\boldsymbol{x}^\circ) = \boldsymbol{k}_N(\boldsymbol{x}^\circ)^\top (K_N + \sigma^2 I_N)^{-1}\boldsymbol{y} = (\boldsymbol{\eta}_{\mathcal{T}}^{\mathrm{GP}})^\top \boldsymbol{y}, \qquad \boldsymbol{\eta}_{\mathcal{T}}^{\mathrm{GP}} := (K_N + \sigma^2 I_N)^{-1}\boldsymbol{k}_N(\boldsymbol{x}^\circ).$$

A second example is the observed average response over a pre-specified region $\mathcal{R} \subset \mathbb{R}^d$,

$$\overline{y}(\mathcal{R}) = \frac{1}{|\mathcal{I}(\mathcal{R})|} \sum_{t\in\mathcal{I}(\mathcal{R})} y_t = (\boldsymbol{\eta}_{\mathcal{T}}^{\mathcal{R}})^\top \boldsymbol{y}, \qquad \mathcal{I}(\mathcal{R}) := \{t : \boldsymbol{x}_t \in \mathcal{R}\},$$

provided $|\mathcal{I}(\mathcal{R})| > 0$.

In these data-independent cases, conditioning remains essential because the ADC trajectory depends on the responses, but the selection event can be simplified to $\{\mathcal{T}_{\boldsymbol{Y}} = \mathcal{T}_{\boldsymbol{y}}\}$. Our framework still yields valid selective tests and confidence intervals for these simpler targets.

B Formal Selective Inference Statements

This appendix provides the formal statements about the elimination of the nuisance parameter by conditioning on the residual statistic in (12). It also provides the formal definitions and properties of the selective p-values and confidence intervals. The one-sided selective p-value in (15) and the equal-tailed two-sided selective confidence interval in (16) have already been formally defined in § 4.2. Therefore, this appendix focuses on the complementary one-sided selective confidence interval and the two-sided selective test, providing their formal definitions and basic properties.

Throughout, let $G_\delta^{\mathrm{sel}}(\cdot)$ denote the conditional CDF in (14) evaluated at effect value δ.

B.1 Elimination of Nuisance Parameter

In § 4.2, we eliminate the nuisance parameter from the conditional distribution by conditioning on an appropriate residual statistic. We formalize that step here.

Proposition 1 (Nuisance parameter-elimination by conditioning)**.** *Let $\boldsymbol{\eta} = \boldsymbol{\eta}_{\boldsymbol{y}}$, and assume the Gaussian model in* (10)*. Define*

$$\mathcal{Q}_{\boldsymbol{Y}} = \left(I_N - \frac{\Sigma\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta}}\right)\boldsymbol{Y}.$$

Then, for any fixed q, the conditional distribution of $T_{\boldsymbol{\eta}}(\boldsymbol{Y}) = \boldsymbol{\eta}^\top \boldsymbol{Y}$ given

$$\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \qquad \mathcal{Q}_{\boldsymbol{Y}} = q$$

depends on the mean parameter $\boldsymbol{\mu}$ only through $\Delta^{\mathrm{sel}} = \boldsymbol{\eta}^\top \boldsymbol{\mu}$. In particular, it does not depend on the nuisance component $\boldsymbol{\mu}^{\mathrm{nui}}$ in (11)*.*

The proof is given in Appendix C.1. Proposition 1 shows that the residual obtained after projecting onto the selected weight-vector direction can be conditioned on to remove the nuisance component from the conditional law of the target statistic. Conditioning on $\mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}$ therefore yields a selective distribution for $T_{\boldsymbol{\eta}}(\boldsymbol{Y})$ that depends on the mean parameter only through Δ^{sel}. In this sense, $\mathcal{Q}_{\boldsymbol{Y}}$ plays the role of a sufficient statistic for the nuisance parameter. The elimination of nuisance parameters by conditioning is standard in conditional inference [29], and the same residual statistic appears in classical selective inference for linear regression [8, 9].

B.2 One-Sided Selective Test

The one-sided selective p-value used in the main text is already defined formally in (15). It tests $\mathrm{H}_0 : \Delta^{\mathrm{sel}} \leq 0$ versus $\mathrm{H}_1 : \Delta^{\mathrm{sel}} > 0$. Its validity is established in Theorem 1; we also provide the stronger exact-uniformity statement.

Proposition 2 (Validity of the one-sided selective test)**.** *Under the assumptions of Theorem 1, the test that rejects for small values of the induced p-value $p_{\mathrm{sel}}(\boldsymbol{Y})$ is exactly uniform under the simple null hypothesis and valid for the composite null. In particular, for every $\alpha \in (0, 1)$,*

$$\mathbb{P}_{\Delta^{\mathrm{sel}}=0}(p_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha \,|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}) = \alpha \qquad \textit{a.s.,} \tag{19}$$

and hence

$$\mathbb{P}_{\Delta^{\mathrm{sel}}=0}(p_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha) = \alpha. \tag{20}$$

Consequently,

$$\mathbb{P}_{\mathrm{H}_0}(p_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha \,|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}) \leq \alpha \qquad \textit{a.s.,}$$

and hence

$$\mathbb{P}_{\mathrm{H}_0}(p_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha) \leq \alpha.$$

B.3 One-Sided Selective Confidence Interval

As noted at the end of § 4.1, the same conditional distribution also supports the inversion of the directional selective test to a lower one-sided selective confidence interval.

Definition 1 (Lower one-sided selective confidence interval)**.** *The lower one-sided selective confidence interval for the effect is obtained by inverting the family of one-sided tests $\mathrm{H}_0 : \Delta^{\mathrm{sel}} \leq \delta$ versus $\mathrm{H}_1 : \Delta^{\mathrm{sel}} > \delta$, namely*

$$C_{1-\alpha}^{\Delta,+}(\boldsymbol{y}) = \left\{\delta \in \mathbb{R} : G_\delta^{\mathrm{sel}}(T(\boldsymbol{y})) \leq 1 - \alpha\right\}. \tag{21}$$

Proposition 3 (Properties of the lower one-sided selective confidence interval)**.** *Under the assumptions of Theorem 1, the set $C_{1-\alpha}^{\Delta,+}(\boldsymbol{y})$ is a lower interval for each observed response vector $\boldsymbol{y}$, and the induced interval $C_{1-\alpha}^{\Delta,+}(\boldsymbol{Y})$ has exact coverage both conditionally and marginally. Writing $\Delta_{\boldsymbol{Y}} = \boldsymbol{\eta}_{\boldsymbol{Y}}^\top \boldsymbol{\mu}$ for the effect, we have*

$$\mathbb{P}\Big(\Delta_{\boldsymbol{Y}} \in C_{1-\alpha}^{\Delta,+}(\boldsymbol{Y}) \,\Big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\Big) = 1 - \alpha \qquad \textit{a.s.,} \tag{22}$$

and therefore

$$\mathbb{P}\Big(\Delta_{\boldsymbol{Y}} \in C_{1-\alpha}^{\Delta,+}(\boldsymbol{Y})\Big) = 1 - \alpha. \tag{23}$$

B.4 Two-Sided Selective Test

The same conditional distribution also supports an equal-tailed two-sided selective test of the point null.

Definition 2 (Equal-tailed two-sided selective test)**.** *The equal-tailed two-sided selective p-value for testing* $\mathrm{H}_0 : \Delta^{\mathrm{sel}} = 0$ *versus* $\mathrm{H}_1 : \Delta^{\mathrm{sel}} \neq 0$ *is*

$$p^{\mathrm{ts}}_{\mathrm{sel}}(\boldsymbol{y}) = 2\min\left\{G^{\mathrm{sel}}_0(T(\boldsymbol{y})), 1 - G^{\mathrm{sel}}_0(T(\boldsymbol{y}))\right\}.$$

Proposition 4 (Validity of the equal-tailed two-sided selective test)**.** *Under the assumptions of Theorem 1, the test that rejects for small values of* $p^{\mathrm{ts}}_{\mathrm{sel}}(\boldsymbol{Y})$ *is exactly uniform under the point null, both conditionally and marginally. In particular, for every* $\alpha \in (0, 1)$*,*

$$\mathbb{P}_{\Delta^{\mathrm{sel}}=0}\big(p^{\mathrm{ts}}_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha \,\big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\big) = \alpha \qquad \textit{a.s.,} \tag{24}$$

and hence

$$\mathbb{P}_{\Delta^{\mathrm{sel}}=0}\big(p^{\mathrm{ts}}_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha\big) = \alpha. \tag{25}$$

B.5 Two-Sided Selective Confidence Interval

The equal-tailed two-sided selective confidence interval used in the main text is already defined formally in (16). Equivalently, it can be written as the inversion of the equal-tailed two-sided selective tests:

$$C^{\Delta}_{1-\alpha}(\boldsymbol{y}) = \left\{\delta \in \mathbb{R} : 2\min\left\{G^{\mathrm{sel}}_\delta(T(\boldsymbol{y})), 1 - G^{\mathrm{sel}}_\delta(T(\boldsymbol{y}))\right\} \geq \alpha\right\}. \tag{26}$$

Proposition 5 (Properties of the two-sided selective confidence interval)**.** *Under the assumptions of Theorem 1, the set* $C^{\Delta}_{1-\alpha}(\boldsymbol{y})$ *is an interval for each observed response vector* $\boldsymbol{y}$*, and the induced interval* $C^{\Delta}_{1-\alpha}(\boldsymbol{Y})$ *has exact coverage both conditionally and marginally. Writing* $\Delta_{\boldsymbol{Y}} = \boldsymbol{\eta}_{\boldsymbol{Y}}^{\top}\boldsymbol{\mu}$ *for the effect, we have*

$$\mathbb{P}\big(\Delta_{\boldsymbol{Y}} \in C^{\Delta}_{1-\alpha}(\boldsymbol{Y}) \,\big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\big) = 1 - \alpha \qquad \textit{a.s.,} \tag{27}$$

and therefore

$$\mathbb{P}\big(\Delta_{\boldsymbol{Y}} \in C^{\Delta}_{1-\alpha}(\boldsymbol{Y})\big) = 1 - \alpha. \tag{28}$$

C Proofs

C.1 Proof of Proposition 1

Proof. Under the Gaussian model in (10), let

$$T_{\boldsymbol{\eta}}(\boldsymbol{Y}) = \boldsymbol{\eta}^{\top}\boldsymbol{Y}, \qquad v_{\boldsymbol{\eta}} = \boldsymbol{\eta}^{\top}\Sigma\boldsymbol{\eta},$$

and define

$$\mathcal{Q}_{\boldsymbol{Y}} = \left(I_N - \frac{\Sigma\boldsymbol{\eta}\boldsymbol{\eta}^{\top}}{v_{\boldsymbol{\eta}}}\right)\boldsymbol{Y}.$$

Then

$$\mathrm{Cov}(\mathcal{Q}_{\boldsymbol{Y}}, T_{\boldsymbol{\eta}}(\boldsymbol{Y})) = \left(I_N - \frac{\Sigma\boldsymbol{\eta}\boldsymbol{\eta}^{\top}}{v_{\boldsymbol{\eta}}}\right)\Sigma\boldsymbol{\eta} = 0.$$

Since $(T_{\boldsymbol{\eta}}(\boldsymbol{Y}), \mathcal{Q}_{\boldsymbol{Y}})$ is jointly Gaussian, it follows that $T_{\boldsymbol{\eta}}(\boldsymbol{Y})$ and $\mathcal{Q}_{\boldsymbol{Y}}$ are independent. Moreover,

$$T_{\boldsymbol{\eta}}(\boldsymbol{Y}) \sim N\big(\Delta^{\mathrm{sel}}, v_{\boldsymbol{\eta}}\big), \qquad \Delta^{\mathrm{sel}} = \boldsymbol{\eta}^{\top}\boldsymbol{\mu}.$$

For any fixed q, every vector $\boldsymbol{z}$ satisfying $\mathcal{Q}_{\boldsymbol{z}} = q$ can be written as

$$\boldsymbol{z} = \frac{\Sigma\boldsymbol{\eta}}{v_{\boldsymbol{\eta}}}t + q, \qquad t = T_{\boldsymbol{\eta}}(\boldsymbol{z}).$$

Therefore, after conditioning on both $\mathcal{Q}_{\boldsymbol{Y}} = q$ and the selection event $\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$, the conditional density of $T_{\boldsymbol{\eta}}(\boldsymbol{Y})$ is proportional to

$$\mathbf{1}\big\{\mathcal{E}_{\Sigma\boldsymbol{\eta}t/v_{\boldsymbol{\eta}}+q} = \mathcal{E}_{\boldsymbol{y}}\big\}\exp\left\{-\frac{(t-\Delta^{\mathrm{sel}})^2}{2v_{\boldsymbol{\eta}}}\right\}.$$

The normalizing constant depends on the truncation region induced by the selection event, but it depends on the mean parameter only through Δ^{sel}. Hence the nuisance parameter is eliminated by conditioning on $\mathcal{Q}_{\boldsymbol{Y}}$. □

C.2 Proof of Lemma 1

Proof. Fix the observed selection event, and let $\boldsymbol{\eta} = \boldsymbol{\eta}_{\boldsymbol{y}}$ so that $\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$ implies $\boldsymbol{\eta}_{\boldsymbol{Y}} = \boldsymbol{\eta}$. Let f^{pre} denote densities under the pre-selection Gaussian distribution from § 4.2.

For notational simplicity, let $v_{\boldsymbol{\eta}} = \boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta}$, and define

$$T_{\boldsymbol{\eta}}(\boldsymbol{Y}) := \boldsymbol{\eta}^\top \boldsymbol{Y}.$$

Under the conditioning on $\mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}$, we have

$$\mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}} \iff \left(I_N - \frac{\Sigma \boldsymbol{\eta}\boldsymbol{\eta}^\top}{v_{\boldsymbol{\eta}}}\right) \boldsymbol{Y} = \mathcal{Q}_{\boldsymbol{y}} \iff \boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}z$$

for $z = \boldsymbol{\eta}^\top \boldsymbol{Y} \in \mathbb{R}$.

Hence,

$$\begin{aligned}
\left\{\boldsymbol{Y} \in \mathbb{R}^N : \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\right\} &= \left\{\boldsymbol{Y} \in \mathbb{R}^N : \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}z, z \in \mathbb{R}\right\} \\
&= \left\{\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}z \in \mathbb{R}^N : \mathcal{E}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{E}_{\boldsymbol{y}},\ z \in \mathbb{R}\right\} \\
&= \left\{\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}z \in \mathbb{R}^N : z \in \mathcal{Z}\right\},
\end{aligned}$$

where

$$\mathcal{Z} := \left\{z \in \mathbb{R} : \mathcal{E}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{E}_{\boldsymbol{y}}\right\}.$$

This proves (17).

Under the pre-selection Gaussian distribution, $T_{\boldsymbol{\eta}}(\boldsymbol{Y})$ and $\mathcal{Q}_{\boldsymbol{Y}}$ are jointly Gaussian. Their covariance is

$$\mathrm{Cov}(T_{\boldsymbol{\eta}}(\boldsymbol{Y}), \mathcal{Q}_{\boldsymbol{Y}}) = \boldsymbol{\eta}^\top \Sigma \left(I_N - \frac{\Sigma \boldsymbol{\eta}\boldsymbol{\eta}^\top}{v_{\boldsymbol{\eta}}}\right)^\top = \boldsymbol{0}^\top.$$

Therefore, $T_{\boldsymbol{\eta}}(\boldsymbol{Y})$ and $\mathcal{Q}_{\boldsymbol{Y}}$ are independent under the pre-selection Gaussian distribution, and hence

$$f^{\mathrm{pre}}_{T_{\boldsymbol{\eta}}(\boldsymbol{Y}),\, \mathcal{Q}_{\boldsymbol{Y}}}(t, q) = f^{\mathrm{pre}}_{T_{\boldsymbol{\eta}}(\boldsymbol{Y})}(t)\, f^{\mathrm{pre}}_{\mathcal{Q}_{\boldsymbol{Y}}}(q),$$

and therefore

$$f^{\mathrm{pre}}_{T_{\boldsymbol{\eta}}(\boldsymbol{Y}) | \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}}(t) = f^{\mathrm{pre}}_{T_{\boldsymbol{\eta}}(\boldsymbol{Y})}(t).$$

Also,

$$T_{\boldsymbol{\eta}}(\boldsymbol{Y}) \sim \mathcal{N}(\Delta^{\mathrm{sel}}, v_{\boldsymbol{\eta}}) \qquad \text{under the pre-selection Gaussian distribution.}$$

Since

$$\left\{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}},\ \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\right\} = \left\{T_{\boldsymbol{\eta}}(\boldsymbol{Y}) \in \mathcal{Z},\ \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\right\},$$

the actual conditional distribution is obtained by restricting the pre-selection Gaussian density to the set $T_{\boldsymbol{\eta}}(\boldsymbol{Y}) \in \mathcal{Z}$. Therefore, for $t \in \mathbb{R}$,

$$\begin{aligned}
&f_{T_{\boldsymbol{\eta}}(\boldsymbol{Y}) | \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}},\, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}}(t) \\
&\quad = \frac{f^{\mathrm{pre}}_{T_{\boldsymbol{\eta}}(\boldsymbol{Y}) | \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}}(t)\, \mathbf{1}\{t \in \mathcal{Z}\}}{\displaystyle\int_{\mathcal{Z}} f^{\mathrm{pre}}_{T_{\boldsymbol{\eta}}(\boldsymbol{Y}) | \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}}(u)\, du} \\
&\quad = \frac{f^{\mathrm{pre}}_{T_{\boldsymbol{\eta}}(\boldsymbol{Y})}(t) \mathbf{1}\{t \in \mathcal{Z}\}}{\displaystyle\int_{\mathcal{Z}} f^{\mathrm{pre}}_{T_{\boldsymbol{\eta}}(\boldsymbol{Y})}(u)\, du} \\
&\quad = f_{\mathrm{TN}(\Delta^{\mathrm{sel}}, v_{\boldsymbol{\eta}}, \mathcal{Z})}(t).
\end{aligned}$$

Finally, on the event $\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$, we have $\boldsymbol{\eta}_{\boldsymbol{Y}} = \boldsymbol{\eta}_{\boldsymbol{y}} = \boldsymbol{\eta}$, and hence

$$T(\boldsymbol{Y}) = T_{\boldsymbol{\eta}}(\boldsymbol{Y}).$$

Therefore,

$$T(\boldsymbol{Y}) \mid \left\{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\right\} \sim \mathrm{TN}(\Delta^{\mathrm{sel}}, v_{\boldsymbol{\eta}}, \mathcal{Z}).$$

This proves the claim. □

C.3 Proof of Theorem 1

Proof. Fix the observed selection event and the sufficient statistic for the observable nuisance parameter. By Lemma 1, the conditional distribution of $T(\boldsymbol{Y})$ is a truncated normal distribution with mean Δ^{sel}, variance $v_{\boldsymbol{\eta}}$, and truncation set $\mathcal{Z}$. Under the null hypothesis (7), we have $\Delta^{\mathrm{sel}} \leq 0$. Let $f_{\Delta,\mathcal{Z}}$ denote the corresponding density.

For $\Delta_1 > \Delta_0$ and $t \in \mathcal{Z}$,

$$\frac{f_{\Delta_1,\mathcal{Z}}(t)}{f_{\Delta_0,\mathcal{Z}}(t)} = C(\Delta_0, \Delta_1, \mathcal{Z}) \exp\left(\frac{(\Delta_1 - \Delta_0)t}{v_{\boldsymbol{\eta}}}\right),$$

where $C(\Delta_0, \Delta_1, \mathcal{Z})$ does not depend on t. Hence this family has a monotone likelihood ratio in the statistic t, and is therefore stochastically increasing in Δ. In particular, if $\bar{F}_{\Delta,\mathcal{Z}}(t) = \mathbb{P}_\Delta\big(T \geq t \mid T \in \mathcal{Z}\big)$ denotes the truncated upper-tail probability, then

$$\bar{F}_{\Delta,\mathcal{Z}}(t) \leq \bar{F}_{0,\mathcal{Z}}(t) \qquad \text{for all } \Delta \leq 0 \text{ and } t \in \mathcal{Z}. \tag{29}$$

The selective p-value in (15) is exactly $p_{\mathrm{sel}}(\boldsymbol{y}) = \bar{F}_{0,\mathcal{Z}}(T(\boldsymbol{y}))$, that is, the upper-tail probability under the least-favorable null value $\Delta^{\mathrm{sel}} = 0$.

Now fix $\alpha \in (0,1)$, and let c_α satisfy

$$\bar{F}_{0,\mathcal{Z}}(c_\alpha) = \alpha.$$

Under the simple null hypothesis $\Delta^{\mathrm{sel}} = 0$, the conditional distribution of $T(\boldsymbol{Y})$ is the truncated normal distribution indexed by $\Delta = 0$, so the probability-integral transform gives exact uniformity:

$$\begin{aligned}\mathbb{P}_{\Delta^{\mathrm{sel}}=0}\big(p_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha \,\big|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\big) &= \mathbb{P}_0\big(T(\boldsymbol{Y}) \geq c_\alpha \mid T(\boldsymbol{Y}) \in \mathcal{Z}\big) \\ &= \bar{F}_{0,\mathcal{Z}}(c_\alpha) = \alpha,\end{aligned}$$

which proves the exact-uniformity claims in (19) and (20).

For the composite null $\Delta^{\mathrm{sel}} \leq 0$, conditionally on the observed selection event and the sufficient statistic for the observable nuisance parameter,

$$\begin{aligned}\mathbb{P}_{\mathrm{H}_0}\big(p_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha \,\big|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\big) &= \mathbb{P}_\Delta\big(T(\boldsymbol{Y}) \geq c_\alpha \mid T(\boldsymbol{Y}) \in \mathcal{Z}\big) \\ &= \bar{F}_{\Delta,\mathcal{Z}}(c_\alpha) \\ &\leq \bar{F}_{0,\mathcal{Z}}(c_\alpha) = \alpha,\end{aligned}$$

where the inequality follows from (29). Taking expectations over the sufficient statistic for the observable nuisance parameter and the selection event preserves the inequality and yields

$$\mathbb{P}_{\mathrm{H}_0}\big(p_{\mathrm{sel}}(\boldsymbol{Y}) \leq \alpha\big) \leq \alpha.$$

This proves (18). □

C.4 Proof of Proposition 3, Proposition 4, Theorem 2, and Proposition 5

Proof. Fix the observed values of $(\mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}})$. By Lemma 1, the conditional distribution of $T(\boldsymbol{Y})$ is a truncated normal family with mean parameter Δ, variance $v_{\boldsymbol{\eta}}$, and truncation set $\mathcal{Z}$. As in the proof of Theorem 1, this family has a monotone likelihood ratio in the statistic T, hence is stochastically increasing in Δ. Therefore, for fixed $T(\boldsymbol{y})$, the conditional CDF $G^{\mathrm{sel}}_\Delta(T(\boldsymbol{y}))$ is continuous and nonincreasing in Δ. Since the conditional distribution shifts to the right as Δ increases, we also have

$$G^{\mathrm{sel}}_\Delta(T(\boldsymbol{y})) \to 1 \quad \text{as } \Delta \to -\infty, \qquad G^{\mathrm{sel}}_\Delta(T(\boldsymbol{y})) \to 0 \quad \text{as } \Delta \to \infty.$$

Hence the acceptance set in (21) is a lower interval, and the acceptance set in (16) is an interval.

For the two-sided interval, the inversion identity in (26) is immediate from the scalar equivalence

$$2\min\{u, 1-u\} \geq \alpha \iff \frac{\alpha}{2} \leq u \leq 1 - \frac{\alpha}{2}.$$

Now evaluate that interval at the true effect $\Delta = \Delta_{\boldsymbol{Y}}$. Because the truncated normal distribution is continuous, the probability integral transform yields

$$G^{\mathrm{sel}}_{\Delta_{\boldsymbol{Y}}}(T(\boldsymbol{Y})) \sim \mathrm{Unif}(0,1) \qquad \text{conditionally on } (\mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}). \tag{30}$$

Hence

$$\begin{aligned}\mathbb{P}\Big(\Delta_{\boldsymbol{Y}} \in C^{\Delta,+}_{1-\alpha}(\boldsymbol{Y}) \,\Big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\Big) &= \mathbb{P}\big(G^{\mathrm{sel}}_{\Delta_{\boldsymbol{Y}}}(T(\boldsymbol{Y})) \le 1-\alpha \,\big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\big) \\ &= 1-\alpha,\end{aligned}$$

which proves (22). Likewise,

$$\begin{aligned}\mathbb{P}\big(\Delta_{\boldsymbol{Y}} \in C^{\Delta}_{1-\alpha}(\boldsymbol{Y}) \,\big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\big) &= \mathbb{P}\Big(\frac{\alpha}{2} \le G^{\mathrm{sel}}_{\Delta_{\boldsymbol{Y}}}(T(\boldsymbol{Y})) \le 1-\frac{\alpha}{2} \,\Big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\Big) \\ &= 1-\alpha,\end{aligned}$$

which proves (27).

Taking expectations of (22) and (27) over $(\mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}})$ yields (23) and (28). Under the point null $\Delta^{\mathrm{sel}} = 0$, (30) becomes $G^{\mathrm{sel}}_0(T(\boldsymbol{Y})) \sim \mathrm{Unif}(0,1)$ conditionally on $(\mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}})$. Therefore,

$$\begin{aligned}\mathbb{P}_{\Delta^{\mathrm{sel}}=0}\big(p^{\mathrm{ts}}_{\mathrm{sel}}(\boldsymbol{Y}) \le \alpha \,\big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\big) &= \mathbb{P}_{\Delta^{\mathrm{sel}}=0}\Big(G^{\mathrm{sel}}_0(T(\boldsymbol{Y})) \le \frac{\alpha}{2} \,\Big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\Big) \\ &\quad + \mathbb{P}_{\Delta^{\mathrm{sel}}=0}\Big(G^{\mathrm{sel}}_0(T(\boldsymbol{Y})) \ge 1-\frac{\alpha}{2} \,\Big|\, \mathcal{E}_{\boldsymbol{Y}}, \mathcal{Q}_{\boldsymbol{Y}}\Big) \\ &= \alpha,\end{aligned}$$

which proves (24). Taking expectations yields (25). The statement in Theorem 2 is exactly the marginal coverage property in (28). □

D Computation of the Interval Set $\mathcal{Z}$

D.1 A Toy Example

To give intuition for why the conditional distribution in our setting becomes truncated normal, we start with the smallest example that already mirrors the setting summarized in § 2 and § 3. Fix $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}\}$ and a collection budget $N = 2$. The first query is always $\boldsymbol{x}_1 = \boldsymbol{x}^{(1)}$, and for $\mathcal{D}_1 = \{(\boldsymbol{x}_1, Y_1)\}$ define

$$\mathcal{A}^{\mathrm{ADC}}(\mathcal{D}_1) = \begin{cases}\boldsymbol{x}^{(2)}, & Y_1 < 0, \\ \boldsymbol{x}^{(3)}, & Y_1 \ge 0.\end{cases}$$

Equivalently, the queried trajectory is $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$ when $Y_1 < 0$ and $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(3)})$ when $Y_1 \ge 0$. This is the simplest example of a piecewise-constant ADC map in § 2: for the fixed first query, the update is constant on the two halfspaces $\{Y_1 < 0\}$ and $\{Y_1 \ge 0\}$. As in § 4, once the realized trajectory is fixed, the realized observed response vector is Gaussian, $\boldsymbol{Y} = (Y_1, Y_2)^\top \sim N(\boldsymbol{\mu}, I_2)$, where Y_2 denotes the second response on the realized branch.

Next define the target-construction map $G : \mathbb{R}^2 \to \mathbb{R}^2$ by selecting the larger observed response:

$$\boldsymbol{\eta}_{\boldsymbol{Y}} = G(\boldsymbol{Y}) = \begin{cases}(0,1)^\top, & Y_2 \ge Y_1, \\ (1,0)^\top, & Y_1 > Y_2.\end{cases}$$

Then the target and test statistic from § 3 are $\Delta_{\boldsymbol{Y}} = \boldsymbol{\eta}_{\boldsymbol{Y}}^\top \boldsymbol{\mu}$ and $T(\boldsymbol{Y}) = \boldsymbol{\eta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}$. This is the simplest example of a piecewise-constant target-construction map in § 3, because G is constant on the two halfspaces $\{Y_2 \ge Y_1\}$ and $\{Y_1 > Y_2\}$.

Now we consider the conditional distribution $T_{\boldsymbol{\eta}}(\boldsymbol{Y}) \,\Big|\, \big\{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}},\ \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\big\}$ in (13).

Because the pair $(\mathcal{T}_{\boldsymbol{y}}, \boldsymbol{\eta}_{\boldsymbol{y}})$ has four possible realizations, the selection event has four branches, each of which corresponds to a different constraint on $\boldsymbol{Y}$. Figure 4 summarizes these four cases. If $\boldsymbol{\eta} = (0,1)^\top$, then $T_{\boldsymbol{\eta}}(\boldsymbol{Y}) = Y_2$ and $\mathcal{Q}_{\boldsymbol{Y}} = (Y_1, 0)^\top$, so conditioning on $\mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}$ fixes $Y_1 = y_1$. The trajectory condition then contributes only the already-fixed sign of y_1, while the remaining event $\boldsymbol{\eta}_{\boldsymbol{Y}} = \boldsymbol{\eta}$ is simply $Y_2 \ge y_1$. Hence the admissible set is $\mathcal{Z} = [y_1, \infty)$. If $\boldsymbol{\eta} = (1,0)^\top$, then $T_{\boldsymbol{\eta}}(\boldsymbol{Y}) = Y_1$ and $\mathcal{Q}_{\boldsymbol{Y}} = (0, Y_2)^\top$, so conditioning fixes $Y_2 = y_2$. The event $\boldsymbol{\eta}_{\boldsymbol{Y}} = \boldsymbol{\eta}$ becomes $Y_1 > y_2$, and the trajectory condition contributes $Y_1 < 0$ when the observed trajectory is $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$ and $Y_1 \ge 0$ when it is $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(3)})$. Therefore $\mathcal{Z} = (y_2, 0)$ in the former case and $\mathcal{Z} = [\max\{0, y_2\}, \infty)$ in the latter.

	$\boldsymbol{\eta_y} = (0,1)^\top$	$\boldsymbol{\eta_y} = (1,0)^\top$
$\mathcal{T}_{\boldsymbol{y}} = (1,2)$	$Y_2 \geq y_1 \iff \mathcal{Z} = [y_1, \infty)$	$y_2 < Y_1 < 0 \iff \mathcal{Z} = (y_2, 0)$
$\mathcal{T}_{\boldsymbol{y}} = (1,3)$	$Y_2 \geq y_1 \iff \mathcal{Z} = [y_1, \infty)$	$Y_1 \geq 0,\ Y_1 > y_2 \iff \mathcal{Z} = [\max\{0, y_2\}, \infty)$

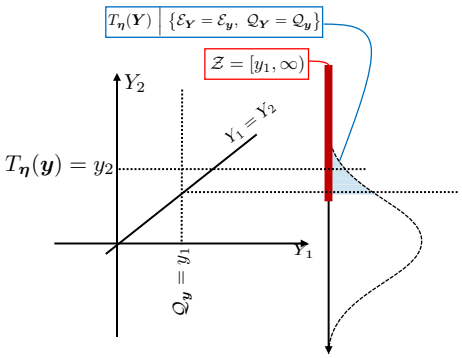

Figure 4: Toy example summary. Left: The four possible realizations of the selection event by trajectory and selected weight vector, with each cell showing the corresponding constraint and the truncation set. Right: A schematic illustration for the branch $\mathcal{T}_{\boldsymbol{y}} = (1,2)$ and $\boldsymbol{\eta_y} = (1,0)^\top$.

Thus

$$T_{\boldsymbol{\eta}}(\boldsymbol{Y}) \mid \{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}},\ \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\} \sim \mathrm{TN}(\Delta^{\mathrm{sel}}, 1, \mathcal{Z}),$$

which is a truncated normal distribution with mean $\Delta^{\mathrm{sel}} = \boldsymbol{\eta}^\top \boldsymbol{\mu}$ and variance $\boldsymbol{\eta}^\top I_2 \boldsymbol{\eta} = 1$, truncated to the set $\mathcal{Z}$ defined above.

This simple example gives intuition for the general case: after conditioning on the selection event and the sufficient statistic for the nuisance parameter, the conditional distribution of the test statistic becomes a one-dimensional Gaussian distribution truncated to a set $\mathcal{Z}$ that can be represented by a finite collection of linear inequalities.

D.2 A Sufficient Condition for Computing $\mathcal{Z}$

Our framework enables the exact computation of $\mathcal{Z}$ for a wide class of problems, including many widely used ADC algorithms and target-construction rules, provided that they satisfy the piecewise-constant properties in Definitions 3 and 4.

Definition 3 (Piecewise-constant ADC maps considered in this paper)**.** *We say that an ADC algorithm $\mathcal{A}^{\mathrm{ADC}}$ belongs to the class considered in this paper if, for every stage $n = 1, \ldots, N-1$ and every fixed queried sequence $\boldsymbol{x}_{1:n} \in \mathcal{X}^n$, there exists a finite polyhedral partition $\{Q_{n,\ell}(\boldsymbol{x}_{1:n})\}_{\ell=1}^{L_n(\boldsymbol{x}_{1:n})}$ of $\mathbb{R}^n$ with associated points $\boldsymbol{x}_{n+1,\ell} \in \mathcal{X}$ such that, for each ℓ,*

$$\mathcal{A}^{\mathrm{ADC}}(\mathcal{D}_n) = \boldsymbol{x}_{n+1,\ell} \qquad \forall \mathcal{D}_n \in (\mathcal{X} \times \mathbb{R})^n \quad \text{such that} \quad \boldsymbol{y}_{1:n} \in Q_{n,\ell}(\boldsymbol{x}_{1:n}).$$

Definition 4 (Piecewise-constant target-construction maps considered in this paper)**.** *We say that a data-dependent target $\Delta_{\boldsymbol{Y}}$ belongs to the target class considered in this paper if its weight vector is defined by a target-construction map*

$$G : \mathbb{R}^N \to \mathbb{R}^N, \qquad \boldsymbol{Y} \mapsto \boldsymbol{\eta}_{\boldsymbol{Y}},$$

that is polyhedral piecewise constant. That is, there exists a finite polyhedral partition $\{P_k\}_{k=1}^K$ of $\mathbb{R}^N$ such that, for each k,

$$\boldsymbol{\eta}_{\boldsymbol{Y}} = \boldsymbol{c}_k \qquad \forall \boldsymbol{Y} \in P_k,$$

for some vector $\boldsymbol{c}_k \in \mathbb{R}^N$. Data-independent maps are a special case of this definition.

With these definitions in place, we introduce the trajectory map associated with an ADC algorithm,

$$H_{\mathcal{A}} : \mathbb{R}^N \to \mathcal{X}^N, \qquad \boldsymbol{Y} \mapsto \mathcal{T}_{\boldsymbol{Y}},$$

which returns the queried trajectory produced by the algorithm from the response vector. When the ADC algorithm in § 2 is polyhedral piecewise constant in the sense of Definition 3, the map $H_{\mathcal{A}}$ is itself polyhedral piecewise constant. This geometric property is the key to the proof below.

Definition 5 (Problems considered in this paper)**.** *A post-ADC analysis problem belongs to the class considered in this paper if its target of inference belongs to the target class in Definition 4, its ADC algorithm satisfies Definition 3, and the weight vector $\boldsymbol{\eta}_{\boldsymbol{Y}}$ is nonzero.*

Proposition 6 (Exact computation of the interval set $\mathcal{Z}$ for problems considered in this paper)**.** *Fix an observed response vector $\boldsymbol{y}$ with $\boldsymbol{\eta}_{\boldsymbol{y}} \neq \mathbf{0}$, and define the one-dimensional slice $\boldsymbol{Y}(z) = \boldsymbol{a} + \boldsymbol{b}z$ as in* (17)*. If the problem belongs to the class considered in this paper with target-construction map G and ADC algorithm $\mathcal{A}$, then*

$$\mathcal{Z}_{G,\mathcal{A}} = \{z \in \mathbb{R} : G(\boldsymbol{Y}(z)) = G(\boldsymbol{y}),\ H_{\mathcal{A}}(\boldsymbol{Y}(z)) = H_{\mathcal{A}}(\boldsymbol{y})\}$$

is a finite union of intervals obtained from finitely many linear equalities and inequalities in z. Consequently, $\mathcal{Z}_{G,\mathcal{A}}$ can be computed exactly by solving finitely many linear constraints.

Proof. For each target-construction cell P_k, the event

$$\{\boldsymbol{Y} \in P_k,\ G(\boldsymbol{Y}) = G(\boldsymbol{y})\}$$

is a polyhedron because P_k is polyhedral and G is constant on P_k. Taking the union over finitely many cells shows that the target-construction event

$$\{G(\boldsymbol{Y}) = G(\boldsymbol{y})\}$$

is a finite union of polyhedra. By the same reasoning, the trajectory event

$$\{H_{\mathcal{A}}(\boldsymbol{Y}) = H_{\mathcal{A}}(\boldsymbol{y})\}$$

is a finite union of polyhedra. Their intersection is therefore a finite union of polyhedra in $\mathbb{R}^N$. Intersecting these polyhedra with the line

$$\{\boldsymbol{a} + \boldsymbol{b}z :\ z \in \mathbb{R}\}$$

produces a finite union of intervals on the z-axis, each described by linear equalities and inequalities in z. □

D.3 Algorithm-Specific Interval Computation Details

This subsection provides more details on how to compute the interval set $\mathcal{Z}$ for the GP-UCB and TPE algorithms.

To compute the interval $[L_z, U_z]$ around a seed point $z \in \mathbb{R}$, it is enough to derive conditions under which applying the same ADC algorithm recursively to $(\mathcal{X}, \boldsymbol{a} + \boldsymbol{b}r)$ yields the same observed selection event as applying it to $(\mathcal{X}, \boldsymbol{a} + \boldsymbol{b}z)$. In the main-text setting, the regions $(\mathcal{I}_{\boldsymbol{a}+\boldsymbol{b}r}, \mathcal{J}_{\boldsymbol{a}+\boldsymbol{b}r})$ are determined by order relations among the observed responses, so once the queried trajectory is fixed they can be represented by linear inequalities in r.

The remaining challenge is therefore to characterize when the queried trajectory and the acquisition-score comparisons from the original ADC run are unchanged. Suppose that the first n ADC steps agree with that run. Then $\boldsymbol{X}_n$ is fixed, and the next queried point $\boldsymbol{x}_{n+1}$ is reproduced whenever

$$\mathcal{A}^{\mathrm{ADC}}(\mathcal{D}_n(r)) = \boldsymbol{x}_{n+1}.$$

For acquisition-maximizing ADC algorithms, this is equivalent to

$$\forall \boldsymbol{x} \in \mathcal{X},\ a(\boldsymbol{x}; \mathcal{D}_n(r)) \leq a(\boldsymbol{x}_{n+1}; \mathcal{D}_n(r)),$$

where $\mathcal{D}_n(r) = \{(\boldsymbol{x}_i, y_i(r))\}_{i=1}^n$ is the restricted dataset formed from the responses $\boldsymbol{y}_{1:n}(r) = \boldsymbol{a}_n + \boldsymbol{b}_n r$.

GP-UCB. For GP-UCB, the posterior mean has the form

$$\mu_n(\boldsymbol{x}; \boldsymbol{y}_{1:n}(r)) = \boldsymbol{k}_n(\boldsymbol{x})^\top (K_n + \sigma^2 I_n)^{-1} \boldsymbol{y}_{1:n}(r),$$

which is affine in the restricted response vector $\boldsymbol{y}_{1:n}(r) = \boldsymbol{a}_n + \boldsymbol{b}_n r$. The posterior variance $s_n^2(\boldsymbol{x})$ depends only on the queried inputs $\boldsymbol{x}_{1:n}$ and is therefore fixed once the history is fixed. Hence, for every candidate $\boldsymbol{x} \in \mathcal{X}$, the GP-UCB comparison

$$\mu_n(\boldsymbol{x}; \boldsymbol{y}_{1:n}(r)) + \kappa s_n(\boldsymbol{x}) \leq \mu_n(\boldsymbol{x}_{n+1}; \boldsymbol{y}_{1:n}(r)) + \kappa s_n(\boldsymbol{x}_{n+1})$$

is a linear inequality in r. Intersecting these inequalities over all $\boldsymbol{x} \in \mathcal{X}$ and over stages $n = 1, \ldots, N-1$ yields a finite family of linear inequalities that characterizes the GP-UCB part of the selection-event-consistent set along the slice.

TPE. Fix $m_n = \lceil \gamma n \rceil$, and let $\mathcal{H}_n(\boldsymbol{y})$ and $\mathcal{L}_n(\boldsymbol{y})$ denote the upper and lower TPE partitions induced by $\boldsymbol{y}$. To keep the same TPE acquisition comparison on the slice in (17), we condition on the quantile-partition event

$$y_i(r) \geq y_j(r) \qquad \forall i \in \mathcal{H}_n(\boldsymbol{y}),\ \forall j \in \mathcal{L}_n(\boldsymbol{y}). \tag{31}$$

Because each $y_i(r)$ has the form $a_i + b_i r$, (31) is again a finite collection of linear inequalities. Up to the probability-zero tie boundary, this conditioning is equivalent to requiring $\mathcal{H}_n(\boldsymbol{y}(r)) = \mathcal{H}_n(\boldsymbol{y})$ and $\mathcal{L}_n(\boldsymbol{y}(r)) = \mathcal{L}_n(\boldsymbol{y})$.

Once the partition is fixed, the TPE score no longer depends on the response magnitudes. Indeed, define

$$\bar{g}_n(\boldsymbol{x}) := \frac{1}{|\mathcal{H}_n(\boldsymbol{y})|} \sum_{i \in \mathcal{H}_n(\boldsymbol{y})} \kappa_{\mathrm{TPE}}(\boldsymbol{x}, \boldsymbol{x}_i), \qquad \bar{\ell}_n(\boldsymbol{x}) := \frac{1}{|\mathcal{L}_n(\boldsymbol{y})|} \sum_{i \in \mathcal{L}_n(\boldsymbol{y})} \kappa_{\mathrm{TPE}}(\boldsymbol{x}, \boldsymbol{x}_i).$$

Then for every r satisfying (31),

$$a^{\mathrm{TPE}}(\boldsymbol{x}; \mathcal{D}_n(r)) = \frac{\bar{g}_n(\boldsymbol{x})}{\bar{\ell}_n(\boldsymbol{x})},$$

which is constant in r. Since $\kappa_{\mathrm{TPE}} > 0$, all denominators are positive, and the next-point condition

$$a^{\mathrm{TPE}}(\boldsymbol{x}; \mathcal{D}_n(r)) \leq a^{\mathrm{TPE}}(\boldsymbol{x}_{n+1}; \mathcal{D}_n(r))$$

is equivalent to the constant inequality

$$\bar{g}_n(\boldsymbol{x})\bar{\ell}_n(\boldsymbol{x}_{n+1}) \leq \bar{g}_n(\boldsymbol{x}_{n+1})\bar{\ell}_n(\boldsymbol{x}).$$

Thus the TPE contribution to the truncation set is determined by linear quantile-partition constraints together with fixed score comparisons on the finite candidate set.

E Randomized-Response Extension

This appendix presents a randomized-response variant of the selective distribution in our setting. The randomization replaces the hard-truncation distribution on $\mathcal{Z}$ by a randomized distribution on $\widetilde{\mathcal{Z}}$, stabilizes the resulting conditional distribution, and avoids the interval-length instability studied in randomized selective inference [11, 28].

E.1 Randomized Selection and Conditional Distribution

Throughout this subsection, assume that both the response covariance and the randomization covariance are isotropic:

$$\Sigma = \sigma^2 I_N, \qquad \Gamma = \tau^2 I_N.$$

Let

$$\boldsymbol{\Omega} \sim \mathcal{N}(\mathbf{0}, \Gamma)$$

be independent of $\boldsymbol{Y}$, where $\tau^2 \geq 0$ is a known randomization level, and define the randomized response vector

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{\Omega}.$$

The optimizer and target-construction map are now applied to $\widetilde{\boldsymbol{Y}}$, whereas inference is still carried out for the original responses $\boldsymbol{Y}$. Let

$$\widetilde{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{\omega}$$

denote the randomized response vector corresponding to the realization $(\boldsymbol{y}, \boldsymbol{\omega})$, where $\boldsymbol{y}$ and $\boldsymbol{\omega}$ are realizations of $\boldsymbol{Y}$ and $\boldsymbol{\Omega}$, respectively, and fix the resulting weight vector $\boldsymbol{\eta} = \boldsymbol{\eta}_{\widetilde{\boldsymbol{y}}}$ and corresponding effect $\Delta^{\mathrm{sel,rnd}} = \boldsymbol{\eta}^\top \boldsymbol{\mu}$. As in the unrandomized case, let

$$v_{\boldsymbol{\eta}} = \boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta}, \qquad \tau^2_{\boldsymbol{\eta}} = \boldsymbol{\eta}^\top \Gamma \boldsymbol{\eta} = \tau^2 \|\boldsymbol{\eta}\|_2^2.$$

Definition 6 (Randomized selection event and the sufficient statistic for the nuisance parameter)**.** *Define*

$$\widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}} = \big(\mathcal{T}_{\widetilde{\boldsymbol{Y}}}, \boldsymbol{\eta}_{\widetilde{\boldsymbol{Y}}}\big),$$

with observed value $\widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{y}}} = \big(\mathcal{T}_{\widetilde{\boldsymbol{y}}}, \boldsymbol{\eta}_{\widetilde{\boldsymbol{y}}}\big)$*, and*

$$\widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = \left(I_N - \frac{(\Sigma + \Gamma)\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\boldsymbol{\eta}^\top(\Sigma + \Gamma)\boldsymbol{\eta}}\right)\widetilde{\boldsymbol{Y}} = \left(I_N - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\boldsymbol{\eta}^\top\boldsymbol{\eta}}\right)\widetilde{\boldsymbol{Y}}. \tag{32}$$

The statistic used for inference remains the weight vector in the unrandomized responses,

$$T(\boldsymbol{Y}) = \boldsymbol{\eta}^\top \boldsymbol{Y},$$

with the weight vector fixed by the randomized selection. Its observed value is $T(\boldsymbol{y}) = \boldsymbol{\eta}^\top \boldsymbol{y}$.

Proposition 7 (Randomized conditional distribution and CDF)**.** *Suppose the selection map belongs to the class considered in Appendix D.2 when viewed as a function of $\widetilde{\boldsymbol{Y}}$. Define*

$$\widetilde{\boldsymbol{a}} = \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{y}}}, \qquad \widetilde{\boldsymbol{b}} = \frac{(\Sigma+\Gamma)\boldsymbol{\eta}}{\boldsymbol{\eta}^\top(\Sigma+\Gamma)\boldsymbol{\eta}} = \frac{\boldsymbol{\eta}}{\boldsymbol{\eta}^\top\boldsymbol{\eta}}, \qquad \widetilde{\boldsymbol{Y}}(z) = \widetilde{\boldsymbol{a}} + \widetilde{\boldsymbol{b}} z, \tag{33}$$

and let

$$\widetilde{\mathcal{Z}} = \left\{ z \in \mathbb{R} : \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}(z)} = \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{y}}} \right\}. \tag{34}$$

Then $\widetilde{\mathcal{Z}}$ is a finite union of intervals. Moreover, if

$$Z \sim \mathcal{N}(\Delta^{\mathrm{sel,rnd}}, v_{\boldsymbol{\eta}}), \qquad R \sim \mathcal{N}(0, \tau_{\boldsymbol{\eta}}^2)$$

are independent, then

$$\mathcal{L}\Big(T(\boldsymbol{Y}) \,\Big|\, \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{y}}}, \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{y}}}\Big) = \mathcal{L}\Big(Z \,\Big|\, Z + R \in \widetilde{\mathcal{Z}}\Big). \tag{35}$$

Equivalently, writing

$$\phi_{\mu,s^2}(x) := \frac{1}{\sqrt{2\pi s^2}} \exp\!\left(-\frac{(x-\mu)^2}{2s^2}\right)$$

for the $\mathcal{N}(\mu, s^2)$ density and

$$w_{\widetilde{\mathcal{Z}}}(t) := \int_{\mathbb{R}} \phi_{0,\tau_{\boldsymbol{\eta}}^2}(r)\, \mathbf{1}\{t + r \in \widetilde{\mathcal{Z}}\}\, dr,$$

then for every $\Delta \in \mathbb{R}$, if $Z_\Delta \sim \mathcal{N}(\Delta, v_{\boldsymbol{\eta}})$ is independent of R, the corresponding randomized selective CDF is

$$G^{\mathrm{rnd}}_{\Delta,\widetilde{\mathcal{Z}}}(t) := \mathbb{P}\Big(Z_\Delta \le t \,\Big|\, Z_\Delta + R \in \widetilde{\mathcal{Z}}\Big) = \frac{\displaystyle\int_{-\infty}^{t} \phi_{\Delta,v_{\boldsymbol{\eta}}}(u)\, w_{\widetilde{\mathcal{Z}}}(u)\, du}{\displaystyle\int_{\mathbb{R}} \phi_{\Delta,v_{\boldsymbol{\eta}}}(u)\, w_{\widetilde{\mathcal{Z}}}(u)\, du}. \tag{36}$$

Proof. Let $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{\Omega}$ and fix the resulting weight vector $\boldsymbol{\eta} = \boldsymbol{\eta}_{\widetilde{\boldsymbol{y}}}$.

$$U := \boldsymbol{\eta}^\top \widetilde{\boldsymbol{Y}} = \boldsymbol{\eta}^\top \boldsymbol{Y} + \boldsymbol{\eta}^\top \boldsymbol{\Omega} = T(\boldsymbol{Y}) + R,$$

where $R = \boldsymbol{\eta}^\top \boldsymbol{\Omega} \sim \mathcal{N}(0, \tau_{\boldsymbol{\eta}}^2)$ is independent of $T(\boldsymbol{Y}) = \boldsymbol{\eta}^\top \boldsymbol{Y} \sim \mathcal{N}(\Delta^{\mathrm{sel,rnd}}, v_{\boldsymbol{\eta}})$. Because $\Sigma = \sigma^2 I_N$ and $\Gamma = \tau^2 I_N$, we have

$$\Sigma + \Gamma = (\sigma^2 + \tau^2) I_N,$$

so the projector in (32) reduces to

$$P_{\boldsymbol{\eta}} := I_N - \frac{(\Sigma+\Gamma)\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\boldsymbol{\eta}^\top(\Sigma+\Gamma)\boldsymbol{\eta}} = I_N - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\boldsymbol{\eta}^\top\boldsymbol{\eta}},$$

and $\widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = P_{\boldsymbol{\eta}} \widetilde{\boldsymbol{Y}}$. The direction vector in (33) likewise reduces to its Euclidean form. Also,

$$\widetilde{\boldsymbol{Y}} - \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = (I_N - P_{\boldsymbol{\eta}})\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{b}}\, U.$$

Hence

$$\widetilde{\boldsymbol{Y}} = \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} + \widetilde{\boldsymbol{b}}\, U.$$

Conditioning on $\widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{y}}}$ therefore yields the one-dimensional representation

$$\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{a}} + \widetilde{\boldsymbol{b}} u, \qquad u = U,$$

with $\widetilde{\boldsymbol{a}}$ and $\widetilde{\boldsymbol{b}}$ as in (33). Since the selection map belongs to the class considered in Appendix D.2 as a function of $\widetilde{\boldsymbol{Y}}$, intersecting its polyhedral cells with this line shows that $\widetilde{\mathcal{Z}}$ in (34) is a finite union of intervals. Therefore,

$$\widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{y}}} \iff U \in \widetilde{\mathcal{Z}} \iff T(\boldsymbol{Y}) + R \in \widetilde{\mathcal{Z}},$$

under the condition $\widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{y}}}$.

$$\begin{aligned}
\mathrm{Cov}\Big(T(\boldsymbol{Y}), \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}}\Big) &= \mathrm{Cov}\big(\boldsymbol{\eta}^\top \boldsymbol{Y}, P_{\boldsymbol{\eta}}(\boldsymbol{Y} + \boldsymbol{\Omega})\big) \\
&= \boldsymbol{\eta}^\top \Sigma P_{\boldsymbol{\eta}}^\top \\
&= \mathbf{0}^\top, \\
\mathrm{Cov}\Big(R, \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}}\Big) &= \mathrm{Cov}\big(\boldsymbol{\eta}^\top \boldsymbol{\Omega}, P_{\boldsymbol{\eta}}(\boldsymbol{Y} + \boldsymbol{\Omega})\big) \\
&= \boldsymbol{\eta}^\top \Gamma P_{\boldsymbol{\eta}}^\top \\
&= \mathbf{0}^\top.
\end{aligned}$$

Thus the jointly Gaussian pair $\big(T(\boldsymbol{Y}), R\big)$ is independent of $\widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}}$. Hence

$$\begin{aligned}
&\mathcal{L}\Big(T(\boldsymbol{Y}) \,\Big|\, \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{y}}}, \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{y}}}\Big) \\
&\quad = \mathcal{L}\Big(T(\boldsymbol{Y}) \,\Big|\, T(\boldsymbol{Y}) + R \in \widetilde{\mathcal{Z}}, \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{y}}}\Big) \\
&\quad = \mathcal{L}\Big(T(\boldsymbol{Y}) \,\Big|\, T(\boldsymbol{Y}) + R \in \widetilde{\mathcal{Z}}\Big).
\end{aligned}$$

Replacing $T(\boldsymbol{Y})$ by an equal-distribution copy $Z \sim \mathcal{N}(\Delta^{\mathrm{sel,rnd}}, v_{\boldsymbol{\eta}})$, independent of R, gives (35). At this stage the conditioning on $\widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}}$ has already been removed, so no independence between Z and $\widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}}$ is needed; Z is only a copy of $T(\boldsymbol{Y})$ used to represent the resulting one-dimensional distribution. Now fix $\Delta \in \mathbb{R}$, let $Z_\Delta \sim \mathcal{N}(\Delta, v_{\boldsymbol{\eta}})$ be independent of R, and let $t \in \mathbb{R}$. By Bayes' rule,

$$\begin{aligned}
G^{\mathrm{rnd}}_{\Delta,\widetilde{\mathcal{Z}}}(t) &= \mathbb{P}\Big(Z_\Delta \le t \,\Big|\, Z_\Delta + R \in \widetilde{\mathcal{Z}}\Big) \\
&= \frac{\mathbb{P}\Big(Z_\Delta \le t,\, Z_\Delta + R \in \widetilde{\mathcal{Z}}\Big)}{\mathbb{P}\Big(Z_\Delta + R \in \widetilde{\mathcal{Z}}\Big)} \\
&= \frac{\displaystyle\int_{-\infty}^{t} \phi_{\Delta, v_{\boldsymbol{\eta}}}(u)\, w_{\widetilde{\mathcal{Z}}}(u)\, du}{\displaystyle\int_{\mathbb{R}} \phi_{\Delta, v_{\boldsymbol{\eta}}}(u)\, w_{\widetilde{\mathcal{Z}}}(u)\, du}.
\end{aligned}$$

This proves (36). Differentiating in t yields the corresponding density

$$f^{\mathrm{rnd}}_{\Delta,\widetilde{\mathcal{Z}}}(t) = \frac{\phi_{\Delta, v_{\boldsymbol{\eta}}}(t)\, w_{\widetilde{\mathcal{Z}}}(t)}{\displaystyle\int_{\mathbb{R}} \phi_{\Delta, v_{\boldsymbol{\eta}}}(u)\, w_{\widetilde{\mathcal{Z}}}(u)\, du}.$$

□

E.2 Randomized Inferential Outputs

For the observed selection event and the sufficient statistic for the nuisance parameter associated with $\widetilde{\boldsymbol{Y}}$, write

$$G^{\mathrm{sel,rnd}}_{\Delta}(t) = \mathbb{P}_{\Delta^{\mathrm{sel,rnd}} = \Delta}\Big(T(\boldsymbol{Y}) \le t \,\Big|\, \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{y}}},\ \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{y}}}\Big),$$

By Proposition 7, this is exactly the CDF in (36) for the observed selection-consistent set $\widetilde{\mathcal{Z}}$.

Definition 7 (Randomized selective p-value and confidence interval)**.** *The randomized one-sided selective p-value is*

$$p_{\mathrm{sel}}^{\mathrm{rnd}}(\boldsymbol{y},\boldsymbol{\omega}) = \mathbb{P}_{\Delta^{\mathrm{sel,rnd}}=0}\Big(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}) \,\Big|\, \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{y}}},\ \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}} = \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{y}}}\Big),$$

and the corresponding equal-tailed randomized selective confidence interval is

$$C_{1-\alpha}^{\Delta,\mathrm{rnd}}(\boldsymbol{y},\boldsymbol{\omega}) = \Big\{\Delta \in \mathbb{R} : \frac{\alpha}{2} \leq G_{\Delta}^{\mathrm{sel,rnd}}(T(\boldsymbol{y})) \leq 1 - \frac{\alpha}{2}\Big\}. \tag{37}$$

When $\tau^2 = 0$, these expressions reduce to (15) *and* (16)*.*

Proposition 8 (Validity of the randomized selective procedures)**.** *Under the assumptions of Proposition 7, the induced randomized selective p-value is exactly uniform under the simple null hypothesis and valid for the composite null. In particular, for every $\alpha \in (0,1)$,*

$$\mathbb{P}_{\Delta^{\mathrm{sel,rnd}}=0}\Big(p_{\mathrm{sel}}^{\mathrm{rnd}}(\boldsymbol{Y},\boldsymbol{\Omega}) \leq \alpha \,\Big|\, \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}}, \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}}\Big) = \alpha \qquad \textit{a.s.}$$

and therefore

$$\mathbb{P}_{\Delta^{\mathrm{sel,rnd}}=0}\big(p_{\mathrm{sel}}^{\mathrm{rnd}}(\boldsymbol{Y},\boldsymbol{\Omega}) \leq \alpha\big) = \alpha.$$

Consequently,

$$\mathbb{P}_{\mathrm{H}_0}\Big(p_{\mathrm{sel}}^{\mathrm{rnd}}(\boldsymbol{Y},\boldsymbol{\Omega}) \leq \alpha \,\Big|\, \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}}, \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}}\Big) \leq \alpha \qquad \textit{a.s.} \tag{38}$$

and therefore

$$\mathbb{P}_{\mathrm{H}_0}\big(p_{\mathrm{sel}}^{\mathrm{rnd}}(\boldsymbol{Y},\boldsymbol{\Omega}) \leq \alpha\big) \leq \alpha. \tag{39}$$

Writing

$$\Delta_{\boldsymbol{Y},\boldsymbol{\Omega}}^{\mathrm{rnd}} = \boldsymbol{\eta}_{\boldsymbol{Y}+\boldsymbol{\Omega}}^{\top}\boldsymbol{\mu}$$

for the effect induced by the randomized responses, we also have

$$\mathbb{P}\Big(\Delta_{\boldsymbol{Y},\boldsymbol{\Omega}}^{\mathrm{rnd}} \in C_{1-\alpha}^{\Delta,\mathrm{rnd}}(\boldsymbol{Y},\boldsymbol{\Omega}) \,\Big|\, \widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}}, \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}}\Big) = 1 - \alpha \qquad \textit{a.s.} \tag{40}$$

and hence

$$\mathbb{P}\Big(\Delta_{\boldsymbol{Y},\boldsymbol{\Omega}}^{\mathrm{rnd}} \in C_{1-\alpha}^{\Delta,\mathrm{rnd}}(\boldsymbol{Y},\boldsymbol{\Omega})\Big) = 1 - \alpha. \tag{41}$$

Proof. For the validity claim, fix the observed selection event and the sufficient statistic for the nuisance parameter associated with $\widetilde{\boldsymbol{Y}}$. By (35), the conditional density of $T(\boldsymbol{Y})$ under parameter value Δ is proportional to

$$\exp\!\left(-\frac{(t-\Delta)^2}{2v_{\boldsymbol{\eta}}}\right)\mathbb{P}\Big(R \in \widetilde{\mathcal{Z}} - t\Big), \qquad \widetilde{\mathcal{Z}} - t := \Big\{r \in \mathbb{R} :\ t + r \in \widetilde{\mathcal{Z}}\Big\}.$$

The factor $\mathbb{P}(R \in \widetilde{\mathcal{Z}} - t)$ does not depend on Δ. Hence, for $\Delta_1 > \Delta_0$,

$$\frac{f_{\Delta_1,\widetilde{\mathcal{Z}}}^{\mathrm{rnd}}(t)}{f_{\Delta_0,\widetilde{\mathcal{Z}}}^{\mathrm{rnd}}(t)} = C(\Delta_0,\Delta_1,\widetilde{\mathcal{Z}})\exp\!\left(\frac{(\Delta_1-\Delta_0)t}{v_{\boldsymbol{\eta}}}\right),$$

where $C(\Delta_0,\Delta_1,\widetilde{\mathcal{Z}})$ is constant in t. Thus the randomized family has a monotone likelihood ratio in T, so it is stochastically increasing in Δ. The same least-favorable-null-value argument used in the proof of Theorem 1 therefore gives (38) and (39).

For the confidence interval, the monotone likelihood ratio property implies that $G_{\Delta}^{\mathrm{sel,rnd}}(T(\boldsymbol{y}))$ is monotone in Δ, so the acceptance set in (37) is an interval. Evaluating that interval at the true randomized effect $\Delta_{\boldsymbol{Y},\boldsymbol{\Omega}}^{\mathrm{rnd}}$, continuity of the conditional distribution gives the probability integral transform

$$G_{\Delta_{\boldsymbol{Y},\boldsymbol{\Omega}}^{\mathrm{rnd}}}^{\mathrm{sel,rnd}}(T(\boldsymbol{Y})) \sim \mathrm{Unif}(0,1) \qquad \text{conditionally on } \Big(\widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}}, \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}}\Big),$$

which yields (40). Averaging over $(\widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{Y}}}, \widetilde{\mathcal{Q}}_{\widetilde{\boldsymbol{Y}}})$ gives (41). □

F Experimental Details

F.1 Detailed Setup

The main experimental settings are described in § 6; this subsection provides the remaining details.

Candidate grid and region family. For each dimension $d \in \{1, 2, 3\}$, we use the axis-aligned grid $\mathcal{X} = \{0, \frac{1}{m_d - 1}, \dots, 1\}^d$ with

$$m_d := \mathrm{round}\Big(m_1^{1/d}\Big), \qquad m_1 = 1024.$$

For the high-region-versus-low-region target, we form the family $\mathfrak{C}_\ell = \{\mathcal{C}(\boldsymbol{s}; \ell) \subset [N] : \ \mathcal{C}(\boldsymbol{s}; \ell) \neq \emptyset, \ \boldsymbol{s} \in \mathbb{R}^d\}$ of axis-aligned hypercubes with side length

$$\ell_d := \ell_1^{1/d}, \qquad \ell_1 = 0.2,$$

and then construct the selected pair $(\mathcal{I}_{\boldsymbol{y}}, \mathcal{J}_{\boldsymbol{y}})$ via (4).

Default ADC hyperparameters. For GP-UCB, the default exploration parameter is $\kappa = 2$, and the surrogate is GP regression with an RBF kernel of variance 1 and dimension-adjusted length scale

$$\lambda_{\mathrm{RBF},d} := \lambda_{\mathrm{RBF},1}\sqrt{d}, \qquad \lambda_{\mathrm{RBF},1} = 0.1,$$

with noise variance matched to the response model. For TPE, the default quantile level is $\gamma = 0.2$ and the Gaussian-kernel bandwidth is 0.1.

Computational resources. All experiments were conducted on a machine equipped with a 48-core AMD EPYC 9474F CPU (3.6 GHz) and 768 GB of memory. However, the proposed inference method can run on a standard laptop or desktop computer. Runtime and memory usage are reported in § F.8.

F.2 Synthetic Objective Families for the Power Experiment

This subsection provides the six synthetic objective families used in Figure 3. Each family is constructed from a separable one-dimensional template aggregated across coordinates, followed by rescaling to the range $[-1, 1]$ over the finite candidate set $\mathcal{X}$.

Six synthetic objective families and the amplitude parameter a. Let $u \in [0, 1]$ and define the six base one-dimensional functions

$$\begin{aligned}
g_{\mathrm{sinc}}(u) &= \mathrm{sinc}\big(10(u - \tfrac{1}{2})\big), && \mathrm{sinc}(t) = \frac{\sin(\pi t)}{\pi t} \ \ (\mathrm{sinc}(0) = 1),\\
g_{\cos}(u) &= -\cos(2\pi u),\\
g_{\mathrm{chirp}}(u) &= \sin(2\pi u^2),\\
g_{\mathrm{bump}}(u) &= \exp\left(-\frac{(u - 0.7)^2}{2 \cdot 0.08^2}\right),\\
g_{\mathrm{peak}}(u) &= 1 - |u - 0.4|,\\
g_{\mathrm{negFor}}(u) &= -(6u - 2)^2 \, \sin(12u - 4).
\end{aligned}$$

For $\boldsymbol{x} = (x_1, \dots, x_d) \in [0, 1]^d$, define the coordinate-aggregated raw objective

$$\tilde{f}_k(\boldsymbol{x}) = \frac{1}{d}\sum_{j=1}^{d} g_k(x_j), \qquad k \in \{\mathrm{sinc}, \cos, \mathrm{chirp}, \mathrm{bump}, \mathrm{peak}, \mathrm{negFor}\}.$$

We then scale each family to the range $[-1, 1]$ over the discrete candidate set $\mathcal{X}$:

$$m_k = \min_{\boldsymbol{x} \in \mathcal{X}} \tilde{f}_k(\boldsymbol{x}), \qquad M_k = \max_{\boldsymbol{x} \in \mathcal{X}} \tilde{f}_k(\boldsymbol{x}), \qquad s_k(\boldsymbol{x}) = \frac{2\tilde{f}_k(\boldsymbol{x}) - (M_k + m_k)}{M_k - m_k}.$$

Finally, the amplitude parameter a enters multiplicatively as

$$f_\star^{(k,a)}(\boldsymbol{x}) = a \, s_k(\boldsymbol{x}).$$

In particular, $a = 0$ yields the global null $f_\star \equiv 0$.

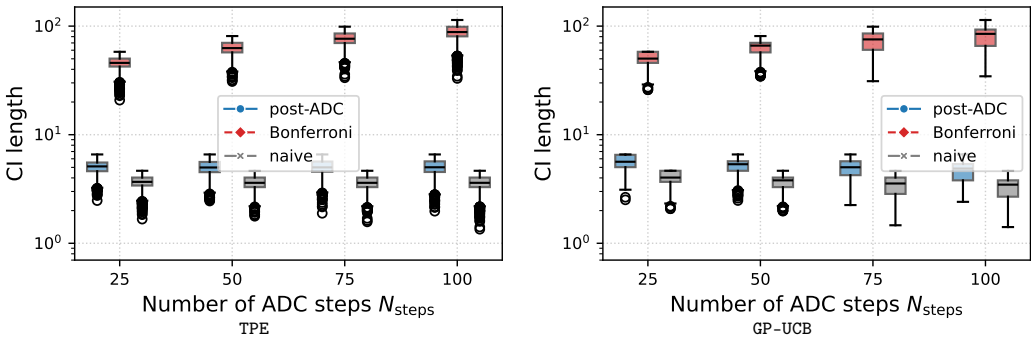

Figure 5: Distribution of selective confidence interval lengths (log scale) for $d = 3$ across horizons and acquisition rules.

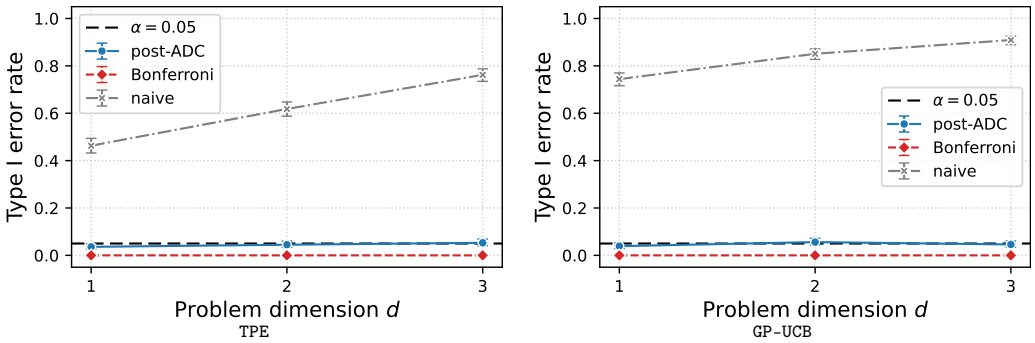

Figure 6: Type I error under the global null $f_\star \equiv 0$ as the problem dimension varies over $d \in \{1, 2, 3\}$, with $N_{\text{steps}} = 50$ fixed. The horizontal dashed line marks the nominal level $\alpha = 0.05$.

F.3 CI Length

This subsection provides the interval-length analysis that complements the coverage results in § 6.1. The setting is the same as in § 6.1: the global null $f_\star \equiv 0$ with $d = 3$ and varying horizon N_{steps}. Figure 5 shows the boxplots of interval lengths. Across the displayed settings, the intervals of `post-ADC` are wider than those of `naive`, which is consistent with the additional conditioning required for validity. Note that this result does not mean that `post-ADC` is worse than `naive`, because `naive` is anti-conservative and hence does not achieve the nominal coverage level (see Figure 2). `Bonferroni` intervals are dramatically wider, which aligns with its extreme conservativeness in the type-I error experiment. Increasing the number of ADC steps changes both the observations used for inference and the selection event on which inference is conditioned. Consequently, the width of the selective confidence interval need not decrease monotonically with N_{steps}.

F.4 Type-I Error vs. Dimension

This subsection complements Figure 2 by fixing $N_{\text{steps}} = 50$ and varying the dimension over $d \in \{1, 2, 3\}$ under the same global-null model $f_\star \equiv 0$. Figure 6 reports the resulting empirical rejection probability for GP-UCB and TPE.

`post-ADC` remains close to the nominal level across the displayed dimensions, while `naive` remains anti-conservative. As in the main-text null experiment, `Bonferroni` stays overly conservative throughout this sweep.

F.5 Type-I Error vs. Acquisition Hyperparameters

We also examine how the null rejection probability changes when the acquisition-rule hyperparameters are varied while holding the remaining experimental conditions fixed at $d = 3$ and $N_{\text{steps}} = 50$. For GP-UCB, we sweep the exploration parameter κ over the grid from 0.5 to 1024. For TPE, we sweep

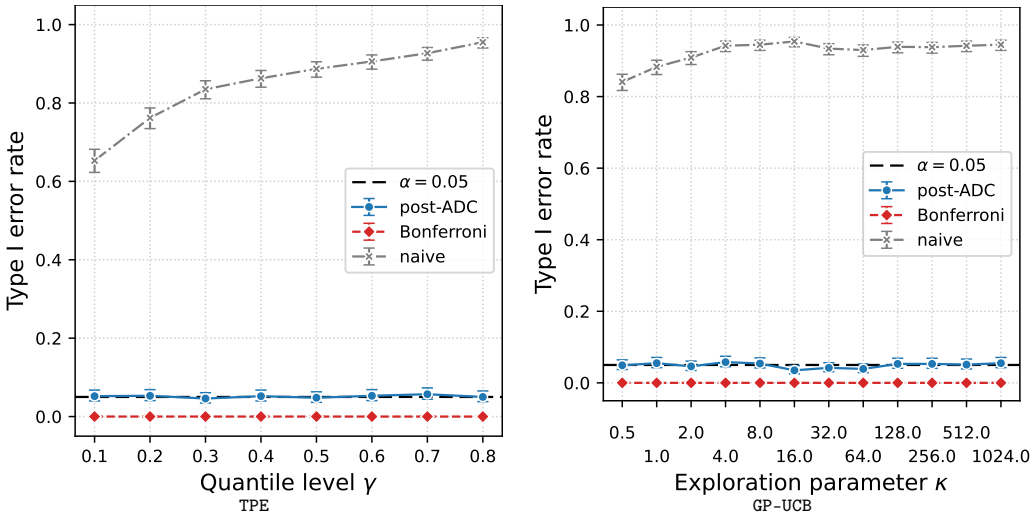

Figure 7: Type I error under the global null $f_\star \equiv 0$ as the acquisition hyperparameters vary, with $d = 3$ and $N_{\text{steps}} = 50$ fixed. The horizontal dashed line marks the nominal level $\alpha = 0.05$.

the quantile level γ over the grid $\{0.1, 0.2, \ldots, 0.8\}$. These denser sensitivity sweeps use 1,000 Monte Carlo replicates per setting. Figure 7 summarizes the resulting type-I error curves. Across these sweeps, `post-ADC` remains close to the nominal level, while `naive` remains anti-conservative over broad portions of the hyperparameter ranges.

F.6 Non-Gaussian Noise

We investigate the type I error rate of `post-ADC` under non-Gaussian observation noise. We consider the global null $f_\star \equiv 0$ with $d = 3$, $N_{\text{init}} = 10$, and $N_{\text{steps}} = 50$. We examine the following five noise distribution families:

- `skewnorm`: Skew-normal distribution.
- `exponorm`: Exponentially modified normal distribution.
- `gennormsteep`: Generalized normal distribution with shape parameter $\beta < 2$, yielding a sharper peak than the Gaussian distribution.
- `gennormflat`: Generalized normal distribution with shape parameter $\beta > 2$, yielding a flatter peak than the Gaussian distribution.
- `t`: Student's t-distribution.

The Gaussian distribution is a special or limiting case of each family, and all noise distributions are standardized to have zero mean and unit variance. For each family, we vary the 1-Wasserstein distance W_1 from the standard Gaussian distribution over $\{0.01, 0.02, 0.03, 0.04\}$. We run 1,000 independent Monte Carlo replications per setting and estimate the type I error rate of `post-ADC` at $\alpha = 0.05$. Figure 8 presents the results for the TPE and GP-UCB acquisition rules. Although the empirical rejection rates tend to deviate more from the nominal level as W_1 increases, they remain generally close to the nominal level over the range examined.

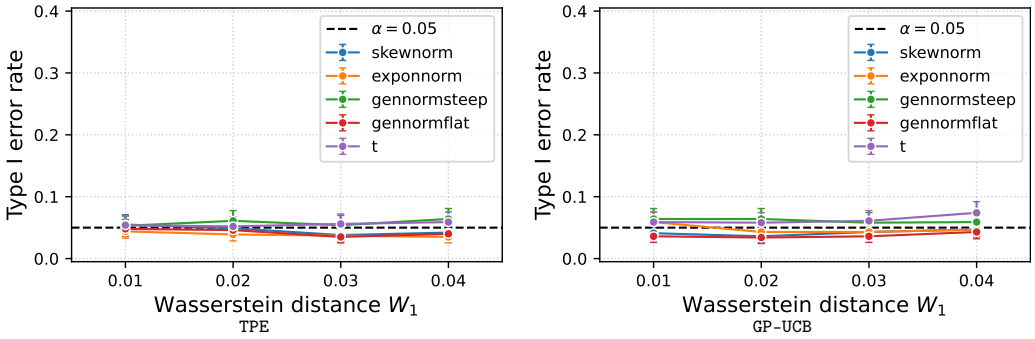

Figure 8: Empirical type I error rates under five non-Gaussian noise distribution families as the 1-Wasserstein distance W_1 from the standard Gaussian distribution varies. The dashed horizontal line indicates $\alpha = 0.05$.

F.7 Plug-In Variance Estimation

We examine how plug-in estimation of the observation-noise variance affects the type I error rate under the global null $f_\star \equiv 0$ with Gaussian observation noise. We set $d = 3$ and $N_{\text{init}} = 10$. We estimate the observation-noise variance by maximizing the GP marginal likelihood with an RBF-plus-white-noise kernel using all observations, including both the initial and subsequently collected observations. The resulting estimate is then plugged into the inference procedure. We vary $N_{\text{steps}} \in \{25, 50, 75, 100\}$ and evaluate rejection rates at $\alpha \in \{0.05, 0.10, 0.20\}$ using 1,000 independent Monte Carlo replications per setting. Figure 9 shows the empirical rejection rates. The empirical rejection rates show slight inflation at $N_{\text{steps}} = 25$, but tend to become more stable and closer to the nominal levels as N_{steps} increases.

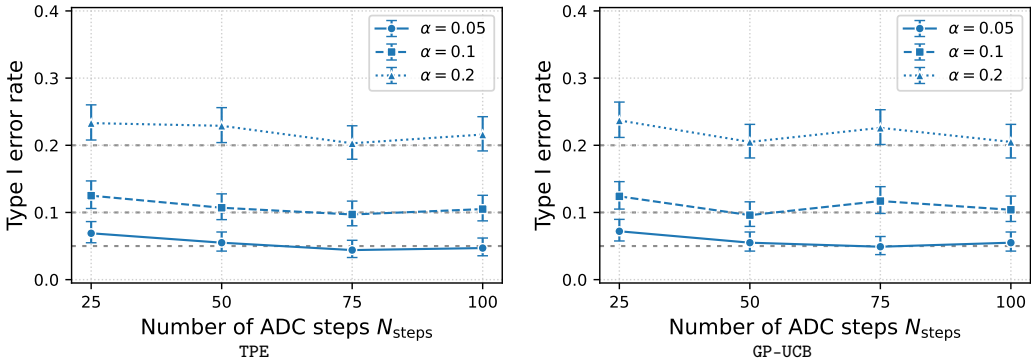

Figure 9: Empirical type I error rates with plug-in estimation of the observation-noise variance under the global null with Gaussian observation noise, for varying N_{steps} and nominal significance levels $\alpha \in \{0.05, 0.10, 0.20\}$. The dashed horizontal lines indicate the nominal levels.

F.8 Implementability and Computational Cost

Implementability Under our structural assumptions, the joint selection event corresponding to the observed ADC trajectory and selected target can be represented as a finite union of polyhedral cells (Proposition 6). After restricting the response vector to the one-dimensional line used for selective inference (Lemma 1), computing the truncation set reduces to solving finitely many linear inequalities in a single scalar variable. Once the procedure-specific inequalities have been generated, they can be processed by a common interval-computation and selective-inference backend. Supporting another ADC procedure satisfying our assumptions therefore requires deriving and implementing its procedure-specific constraint-generation step, as illustrated for GP-UCB and TPE in Appendix D.3, but does not require reimplementing the subsequent truncated-normal inference machinery. The implementation is modular within the supported class, but it is not a black-box wrapper for arbitrary ADC algorithms.

Computational Cost The computational cost is governed primarily by the number of inequalities generated by the observed ADC trajectory and target-construction rule. This number generally increases with the number of ADC steps and candidate comparisons performed at each step. Consequently, longer acquisition horizons and larger candidate sets tend to increase runtime. Because the resulting inequalities are one-dimensional, they can be processed sequentially without solving a high-dimensional optimization problem. The main memory costs arise from retaining the quantities required to reproduce the observed decisions and construct the corresponding constraints.

To evaluate the practical scalability of the proposed framework, we measured the inference time and memory usage for computing a single post-ADC p-value. Inference time was measured as wall-clock time, and memory usage was measured as the peak resident set size (RSS) of the process. We used the Gaussian global null $f_\star \equiv 0$, $d = 3$, $N_{\text{init}} = 10$, and $N_{\text{steps}} \in \{25, 50, 75, 100\}$ for both TPE and GP-UCB, with $1{,}000$ independent Monte Carlo replicates per setting. Each experiment was run using a single worker. The computational environment is described in § F.1. Figure 10 shows that both inference time and peak RSS generally increase with the number of ADC steps. Even at $N_{\text{steps}} = 100$, however, the maximum observed peak RSS remains below 1 GB, while the median inference time remains below 10 s for TPE and below 100 s for GP-UCB.

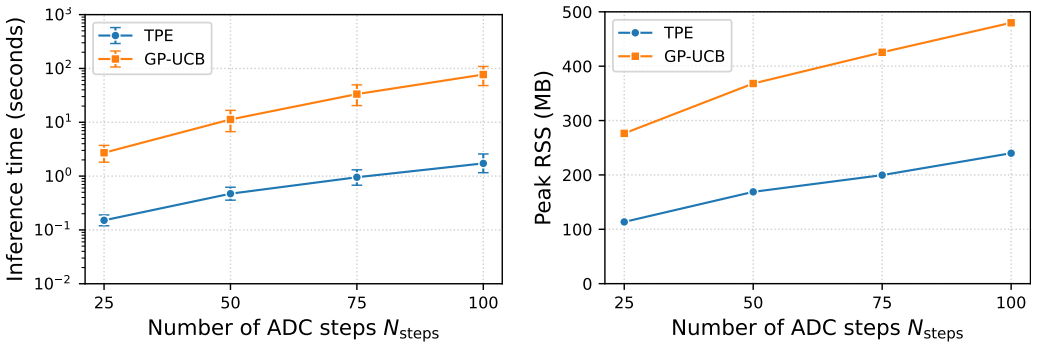

Figure 10: Scalability of post-ADC p-value inference with respect to the number of ADC steps. Left: median inference time, with error bars indicating the interquartile range (logarithmic scale). Right: maximum observed peak resident set size (RSS) across $1{,}000$ replicates for each setting.

F.9 Real-Data Experiment Details

Datasets. We use four UCI regression benchmarks, summarised in Table 3. All datasets are licensed under the CC BY 4.0 license. For each dataset we project the inputs onto a d-dimensional subspace spanned by the top-d features ranked by mutual information with the response; feature sets for each $d \in \{1, 2, 3\}$ are listed in Table 3. The candidate set $\mathcal{X}$ is formed by taking all distinct projected input vectors in the dataset and, if their count exceeds 1024, sub-sampling 1024 uniformly at random. Response noise variance σ^2 is estimated by maximising the GP marginal likelihood (RBF kernel, length scale fitted jointly) on the full dataset; this $\hat{\sigma}^2$ is then used as the fixed noise level in the inference procedure.

Table 3: Real datasets used in § 6.3. Features are listed in order of mutual-information rank; the top d are used for each value of d.

Dataset	N	$\lvert\mathcal{X}\rvert$	Response	Top-3 features (rank order)
Concrete Compressive Strength	1030	412	compressive strength	water, cement, superplasticizer
Gas Turbine (CO)	36733	1024	CO emission	TIT, TEY, CDP
Gas Turbine (NOx)	36733	1024	NOx emission	TIT, AT, TEY
Combined Cycle Power Plant	9568	1024	net hourly output	AT, V, AP

Experimental protocol. For each dataset and each $(d, \text{acquisition})$ combination, we draw $1{,}000$ independent bootstrap replicates. In each replicate, $N = N_{\text{init}} + N_{\text{steps}} = 10 + 50 = 60$ samples are drawn uniformly without replacement from the full dataset and used as the observed responses.

Table 4: Empirical rejection probability on five real datasets with $d = 2$ and $1{,}000$ bootstrap replicates.

Acquisition	Method	Concrete	Gas Turbine (CO)	Gas Turbine (NOx)	Power Plant
ucb	`post-ADC`	**0.145**	**0.994**	**0.836**	**1.000**
	`Bonferroni`	0.000	0.000	0.000	0.000
tpe	`post-ADC`	**0.246**	**0.952**	**0.781**	**0.992**
	`Bonferroni`	0.000	0.000	0.000	0.000

Table 5: Empirical rejection probability on five real datasets with $d = 3$ and $1{,}000$ bootstrap replicates.

Acquisition	Method	Concrete	Gas Turbine (CO)	Gas Turbine (NOx)	Power Plant
ucb	`post-ADC`	**0.166**	**0.999**	**0.741**	**1.000**
	`Bonferroni`	0.000	0.003	0.000	0.000
tpe	`post-ADC`	**0.167**	**0.975**	**0.747**	**0.999**
	`Bonferroni`	0.000	0.006	0.000	0.000

The regions $(\mathcal{I}_{\boldsymbol{y}}, \mathcal{J}_{\boldsymbol{y}})$ are constructed exactly as in the synthetic experiments. We compare only `post-ADC` and `Bonferroni`.

Results for $d = 2$ and $d = 3$. Tables 4 and 5 report the empirical rejection probabilities for $d = 2$ and $d = 3$, respectively. The results are consistent with the $d = 1$ results in Table 1: `post-ADC` detects signal on datasets with informative structure, while `Bonferroni` remains essentially powerless throughout.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and Introduction state that the paper proposes post-ADC inference, gives valid selective p-values and confidence intervals under Gaussian observation noise, and demonstrates the method on GP-UCB and TPE. These claims are matched by the formal setup and validity results in § 3–§ 5 and by the experiments in § 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: § 7 discusses the limitations of the proposed approach.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The assumptions are stated in the model and problem setup in § 3.1–§ 5 and formalized in the appendices. Complete proofs for the main validity and reduction statements are provided in Appendices B, C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: § 6 reports the experimental setup and Appendix F.1 gives the details of the experimental setup to reproduce the results, including the hyperparameters, data preprocessing procedures, and evaluation criteria.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: Implementation code is provided in the supplementary material. It includes instructions for reproducing the main experimental results and generator scripts for synthetic data. In real-data experiments, we used the UCI benchmark datasets described in Appendix F.9.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The paper provides all necessary training and test details

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: This study is a paper on a new statistical testing method and properly discusses statistical significance of the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: We described the information on computer resources in Appendix F.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is methodological research with synthetic experiments and public benchmark datasets, and it does not involve human-subject experiments, private data collection, or release of high-risk models or datasets.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The work is foundational statistical methodology for valid post-hoc inference after active data collection and has the potential to broadly influence the machine learning

community. On the other hand, it does not present significant ethical concerns or foreseeable societal consequences because this work is theoretical and, as of now, has no direct applications that might impact society or raise ethical considerations.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: This study is on the statistical reliability of AI and poses no risk of misuse. The paper does not release pretrained generative models, scraped datasets, or other assets with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: We appropriately credited all the code, data, and models used in this study.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: This paper does not release pretrained generative models, scraped datasets, or other assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval or an equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The core methodology does not use LLMs as an important, original, or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.